INTERACTIVE CONTROL ENGINEERING

# Model Predictive Control and Economic Optimization

## Interactive Control Engineering: Principles and Applications

Edited by John D. Hedengren

### Control technology


1 Process Dynamics, Modeling, and Identification
2 Practical PID Control
3 Feedforward, Cascade, and Ratio Control
4 Override, Split-Range, and Constraint Control
5 Delays, Slow Measurements, and State Estimation
6 Multivariable and Plantwide Control
7 Nonlinear Dynamics, Oscillations, and Stabilization
8 Batch, Run-to-Run, and Hybrid Control
9 Model Predictive Control and Economic Optimization


### Industry applications


10 Control in Upstream and Midstream Oil and Gas
11 Control in Petroleum Refining
12 Control in Petrochemicals and Polymers
13 Control in Chemical Plants
14 Control in Semiconductor Manufacturing
15 Control in Mining and Mineral Processing
16 Control in Iron, Steel, and Metals Production
17 Control of Energy Conversion and Storage
18 Control of Nuclear and Fusion Systems
19 Control of Biological Systems and Human Performance
20 Motion Control, Robotics, and Autonomous Systems

INTERACTIVE CONTROL ENGINEERING
Principles and Applications

Volume 9

# Model Predictive Control and Economic Optimization

Planning moves ahead, and choosing where to operate

John D. Hedengren
Editor

Advanced Process Solutions, LLC

*Interactive Control Engineering: Principles and Applications*
Volume 9: Model Predictive Control and Economic Optimization
Edited by John D. Hedengren

First edition, 2026

Copyright © 2026 Advanced Process Solutions, LLC. All rights reserved.

Produced by Advanced Process Solutions, LLC.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means without the prior written permission of the publisher, except for brief quotations in a review or scholarly work.

The simulators described in this book are teaching models. Their equations, parameters and results are simplified for instruction and are not validated for the design, operation or safety assessment of any real plant, vehicle or medical device. The editor and the publisher accept no liability for use of this material outside that purpose.

Names of products and companies are used for identification only and may be trademarks of their owners.

Every chapter has a companion simulator that runs in a web browser at https://apmonitor.com/apps/. This volume online, with chapter files and citation records: https://apmonitor.com/apps/books/v09-mpc-optimization/

**How to cite this volume.** John D. Hedengren, Ed., *Model Predictive Control and Economic Optimization*, vol. 9 of *Interactive Control Engineering: Principles and Applications*. Advanced Process Solutions, LLC, 2026.

Typeset in Times and Helvetica-family type with LaTeX and the memoir class.

# Preface

This book is one of 20 volumes of *Interactive Control Engineering: Principles and Applications*. Each chapter is about one plant, machine or system and one control lesson, and each has a simulator that runs in a web browser with nothing to install. The simulators are collected at https://apmonitor.com/apps/. Nine volumes are organized by control technology and eleven by industry or application area; the last pages of this volume list them all.

### How to read this book with the apps

Chapter 1 teaches the theory this volume relies on and explains, once, the simulator framework that the apps share. Each chapter after it opens with the control question it answers, its learning objectives, what to read first, and the address of its app, also given as a code to scan. The chapter then develops the model, the controllers and the measured results. Every number reported was measured on the app, so a reader can open the app, load the named scenario and reproduce it.

Every chapter ends with exercises. Each is labeled with its kind and the time it should take, and none takes more than ten minutes. Concept checks and calculations can be done on paper. App challenges name a scenario, the settings and the run length. One exercise in each chapter, labeled *Review the agent*, presents a tuning, a model or a claim proposed by an AI assistant and asks what is wrong with it. Short answers to the odd-numbered exercises are at the back of the book, and worked solutions are available to instructors.

### How AI agents were used

AI agents did much of the development work behind this series. Under the editor's direction they wrote and tested the simulators, drafted the papers from which the chapters are drawn, measured the reported results on the running apps, converted the papers to book form, and drafted the introductory and closing chapters and the exercises. Other agents then checked that work: they re-measured reported numbers, audited what each app displays against the state of its model, reviewed the papers as referees, and solved the exercises without seeing the answers. The editor set the scope, chose the plants and the lessons, reviewed the results and is responsible for the content.

That way of working is also a subject of the book. The last chapter of each volume describes what the agents got wrong while this volume was made and how the errors were caught, and it sets out what the editor believes this means for engineers: more of the routine work can be delegated, and more of the engineer's value lies in checking, supervising, choosing the problem and knowing the process.

### About this volume

This volume is about controllers that plan ahead and about choosing where to operate. Chapters 2 to 7 use a model to predict and to coordinate future moves toward a target: across a dead time, through a change of regime, along a canal, through a heat network whose delay follows its flow, and against a forecast that may be wrong. Chapters 8 to 12 ask which target or trajectory is worth reaching, with an optimum that moves with the weather, a comfort band in place of a setpoint, a discharge permit, pumps that are either on or off, and a timetable.

Each chapter also reports where simpler control was enough. Volume 4 gives the selector and override schemes that the predictive controllers are compared with.

John D. Hedengren
2026

# Contents

**Preface** v

**Notation** x

**1 Introduction: Planning Ahead With a Model** 1
- 1.1 The Question This Volume Answers . . . 1
- 1.2 A Short History . . . 2
- 1.3 Notation and Conventions . . . 3
- 1.4 Prediction and Optimization on One Heated Tank . . . 3
- 1.5 Instruments and Final Elements in This Volume . . . 12
- 1.6 The Simulator Framework . . . 13
- 1.7 Working With AI Agents While Studying . . . 17
- 1.8 Chapter Roadmap . . . 17
- 1.9 Exercises . . . 20

**2 Hot-Air Balloon Simulator for Landing an Integrating Plant Behind a Dead Time** 22
- 2.1 Introduction . . . 23
- 2.2 Balloon Flight and the Control of Altitude . . . 24
- 2.3 Flight Model . . . 25
- 2.4 The Chain From Blast Valve to Altitude . . . 30
- 2.5 Controllers . . . 31
- 2.6 Results . . . 35
- 2.7 Technology in Practice . . . 41
- 2.8 Limitations . . . 42
- 2.9 Discussion . . . 43
- 2.10 Conclusion . . . 44
- 2.11 Exercises . . . 45

**3 Friction Stir Welding Simulator for Regime-Dependent Tuning and Predictive Control** 49
- 3.1 Introduction . . . 50
- 3.2 Friction Stir Welding and the Arrival of Temperature Control . . . 51
- 3.3 Process Model . . . 52
- 3.4 Three Plants in One Weld . . . 59
- 3.5 Controllers . . . 60
- 3.6 Results . . . 63
- 3.7 Technology in Practice . . . 68
- 3.8 Limitations . . . 70
- 3.9 Discussion . . . 70
- 3.10 Conclusion . . . 71
- 3.11 Exercises . . . 73

**4 Boiler-Turbine Simulator Where Fast Megawatts Are Borrowed From Drum Pressure** 76
- 4.1 Introduction . . . 77
- 4.2 From a Peak-Load Unit to a Control Benchmark . . . 78
- 4.3 The Unit as Modeled . . . 79
- 4.4 What the Unit Does Before Any Control . . . 83
- 4.5 The Unit Master as Implemented . . . 84
- 4.6 Eight Cases and a Challenge, Measured . . . 87

4.7 Technology in Practice . . . 90
4.8 Limitations . . . 93
4.9 Conclusion . . . 93
4.10 Exercises . . . 95

**5 Irrigation Canal Simulator for Distant Downstream Control, Resonance and Releasing Water Ahead of Demand 98**
5.1 Introduction . . . 99
5.2 Automation of Irrigation Canals . . . 100
5.3 Process Model . . . 101
5.4 Implemented Control . . . 103
5.5 Results . . . 108
5.6 Technology in Practice . . . 117
5.7 Limitations . . . 118
5.8 Conclusion . . . 119
5.9 Exercises . . . 121

**6 Pricing Heat in a District Heating Network Whose Delay Follows the Flow 125**
6.1 Introduction . . . 126
6.2 Hot Water Networks and the Price of Heat . . . 127
6.3 A Plant, Three Trunks and Six Substations . . . 128
6.4 Why the Flow Sets the Delay . . . 132
6.5 Loops, Planners and a Price . . . 133
6.6 What the Cases Measure . . . 136
6.7 Technology in Practice . . . 142
6.8 Limitations . . . 143
6.9 Conclusion . . . 144
6.10 Exercises . . . 146

**7 Reservoir Flood Control on an Ensemble Forecast With Scenario Trees 149**
7.1 Introduction . . . 150
7.2 Flood Pools, Value Functions and Ensemble Forecasts . . . 151
7.3 A Declared Basin From Rain to the Town . . . 152
7.4 Why the Pool Must Be Emptied Before the Storm . . . 156
7.5 Deciding the Morning's Release . . . 157
7.6 Measurements on Ten Cases and a Water Year . . . 159
7.7 Technology in Practice . . . 165
7.8 Limitations . . . 166
7.9 Conclusion . . . 167
7.10 Exercises . . . 169

**8 A Data Center Cooling Simulator Whose Optimum Moves With the Wet Bulb 172**
8.1 Introduction . . . 173
8.2 Background . . . 175
8.3 Process Model . . . 176
8.4 Selection of the Supervisory Decision Variables . . . 181
8.5 Controller Design . . . 182
8.6 Results . . . 185
8.7 Technology in Practice . . . 194
8.8 Limitations . . . 196
8.9 Conclusion . . . 197
8.10 Exercises . . . 199

**9 A Building HVAC Simulator With a Comfort Band Instead of a Setpoint 202**
9.1 Introduction . . . 203
9.2 How a Building Came to Have a Band to Spend . . . 204
9.3 Process Model . . . 206
9.4 Dynamic Character . . . 213
9.5 Control Structure as Implemented . . . 214
9.6 Results . . . 218
9.7 Technology in Practice . . . 225
9.8 Limitations . . . 227
9.9 Discussion . . . 228
9.10 Conclusion . . . 228
9.11 Exercises . . . 230

**10 Wastewater Aeration Simulator for Supervisory Setpoint Selection Under a Discharge Permit 234**
10.1 Introduction . . . 235
10.2 Where Aeration Control Came From . . . 236
10.3 Process Model . . . 237
10.4 Controller Design . . . 243
10.5 Results . . . 247
10.6 Technology in Practice . . . 255
10.7 Limitations . . . 257
10.8 Discussion . . . 258
10.9 Conclusion . . . 259
10.10 Exercises . . . 261

**11 Water Distribution Simulator That Schedules Integer Pumps Under a Pressure Floor 265**
11.1 Introduction . . . 266
11.2 From Pumping Costs to Pressure Management . . . 267
11.3 The Town as Modeled . . . 268
11.4 Why the Hour Matters . . . 271
11.5 The Control Layers as Programmed . . . 272
11.6 Ten Cases, Measured . . . 274
11.7 Technology in Practice . . . 279
11.8 Limitations . . . 281
11.9 Conclusion . . . 281
11.10 Exercises . . . 283

**12 Automatic Train Operation Simulator Where the Timetable Decides Where to Coast 287**
12.1 Introduction . . . 288
12.2 From a Driver's Judgment to Planned Profiles . . . 289
12.3 The Train, the Line and the Rail as Modeled . . . 290
12.4 The Structure of an Energy-Optimal Run . . . 293
12.5 Planners and Tracking as Implemented . . . 295
12.6 Nine Cases and a Challenge, Measured . . . 297
12.7 Technology in Practice . . . 302
12.8 Limitations . . . 304
12.9 Conclusion . . . 305
12.10 Exercises . . . 307

**13 Summary and Outlook 311**
13.1 What the Chapters Showed . . . 311

13.2 Choosing a Method . . . 314
13.3 Where Practice Is Heading . . . 315
13.4 Engineering With AI Agents . . . 318
13.5 Gaps in Predictive Control and Economic Optimization . . . 320
13.6 Exercises . . . 324

**Answers to Selected Exercises** **326**

**Glossary** **331**

**Index of Control Topics** **334**

**Index of Applications** **336**

**About the Editor** **338**

**Volumes in This Series** **339**

## Notation

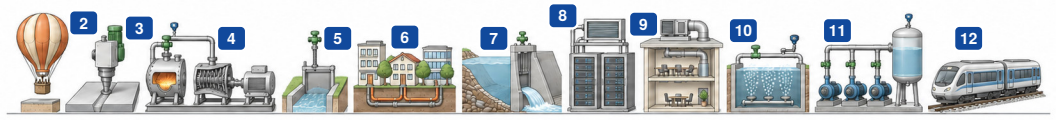
The chapters share the symbols below. A symbol used by one chapter only is defined where it first appears, and each chapter states its own units.

| Symbol | Meaning |
|---|---|
| SP | setpoint, the value a controller is asked to hold |
| PV | process variable, the measurement a controller acts on |
| CV | controlled variable |
| MV | manipulated variable; also OP, the controller output |
| DV | disturbance variable |
| $e$ | control error, SP $-$ PV |
| $K_c$ | controller gain |
| $\tau_I$, $\tau_D$ | integral (reset) time and derivative time |
| $K_p$ | process gain, the steady-state change in PV per unit change in MV |
| $\tau_p$ | process time constant |
| $\theta_p$ | process dead time |
| $\tau_c$ | closed-loop time constant chosen in a tuning rule |
| $\Delta t$ | controller sample time |
| $s$ | Laplace variable |
| $x$, $u$, $y$, $d$ | state, input, output and disturbance of a state-space model |
| $A$, $B$, $C$ | matrices of the linear model $\dot{x} = Ax + Bu$, $y = Cx$ |
| $Q$, $R$ | weights on outputs (or states) and on input moves in an optimal controller |
| $\hat{x}$ | an estimate of $x$ |

Abbreviations used throughout: PID, proportional-integral-derivative; PI, proportional-integral; FOPDT, first-order plus dead time; MPC, model predictive control; MHE, moving horizon estimation; IAE, integral of the absolute error; RGA, relative gain array. Times are in seconds unless a chapter says otherwise. Controller output and valve position are in percent of span. A scenario or case number refers to the numbered list in the chapter's app.

CHAPTER 1

# Introduction: Planning Ahead With a Model




**The chapters** 2 hot-air balloon 3 friction stir weld 4 boiler-turbine unit 5 irrigation canal 6 district heating 7 flood reservoir 8 data center cooling 9 office building 10 aeration basins 11 water network pumps 12 electric train

*Notice that the first four units are regulated by a predictive controller, the next five decide what the loops should hold, where a price, a permit or a forecast sets the target, and the last two are planned in whole decisions. Illustration, not to scale.*

**IN PLAIN TERMS** **Control That Looks Ahead**

**The main idea.** An ordinary controller reacts to the error it sees now. A predictive controller uses a model of the plant to guess what comes next, and it plans its moves several steps ahead.

**The challenge.** The plan is only as good as the model and the forecast behind it. Limits, prices and machines that are either on or off make the best plan harder to find, and it has to be found again every few moments.

**Why it matters.** Such plans run heat networks, dams, data halls, offices, water works and trains, where they save energy and money.

**THREE WAYS TO READ THIS CHAPTER**

| | | |
|---|---|---|
| **Explore** | 7 min | This page, Section 1.2 with its timeline and two History boxes, the Did you know? boxes in Sections 1.4.2, 1.4.7 and 1.4.13, Figure 1.2, and Key points in Section 1.8. No equations. |
| **Study** | 45 min | The whole chapter in order, with each number of the tank example worked, then Exercises 1 to 10. |
| **Practice** | 5 min | This page, Section 1.4.13 with its two tables, Section 1.5 with Figure 1.6, Figure 1.2, and Section 1.8. |

## 1.1 The Question This Volume Answers

A feedback loop acts on the error it sees now. A predictive controller acts on the errors a model says are coming, and an optimizer chooses what the loops should aim for. Both cost something: a model to build and keep right, a program to solve every period, and a layer of logic that operators must trust. This volume answers one question: when does planning ahead with a model pay for itself against simpler control, and how is such a plan built so that it survives a wrong model, a limit and an uncertain forecast?

Eleven chapters answer it on eleven plants, each with an app that the reader runs in a browser. Chapters 2 to 5 use prediction to regulate: a balloon behind a dead time, a weld whose gain changes along the joint, a boiler that trades pressure for megawatts and a canal whose gates are kilometers apart. Chapters 6 to 10 use it to decide what the regulators should hold, on a heating network, a flood reservoir, a data hall, an office building and an aeration basin, where a price, a permit or a forecast sets the target. Chapters 11 and 12 plan in whole decisions: pumps that are on or off, and a train that powers, holds, coasts or brakes.

A reader who has worked through the volume can do the following on a plant that is new to the reader.

- Write the prediction of a step-response or state-space model and the quadratic program a predictive controller solves each period.

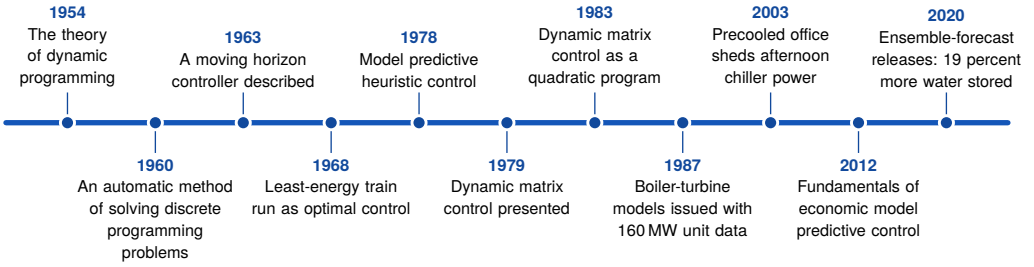



**Figure 1.1** Milestones of prediction, optimization and planning in the systems of this volume. The dates are from [5] (1954), [6] (1960), [2] (1963, 1979 and 1983), [8] (1968), [7] (1978), [9] (1987), [10] (2003), [11] (2012) and [12] (2020). Notice that the optimization came first and the plants that store energy or water came last. Events are evenly spaced, not to a time scale.

- Choose a horizon, a move suppression and a disturbance model, and say what each does to speed, robustness and offset.
- Decide which limits are hard and which are soft, and price a soft limit so that it is broken only when it must be.
- Find the active constraint of a steady-state optimum and the back-off that variability forces.
- Set up an economic plan, a scenario tree, a price-coordinated plan and an integer schedule, and read their answers through the structure that optimal control predicts.
- Measure what each of these buys against a well-tuned simpler controller.

This chapter develops each of these on one small example and names the chapter that uses it. Surveys of the theory and of industrial practice [1–3] and a textbook treatment [4] give fuller derivations. PID forms and tuning are taken as known from Volume 2.

## 1.2 A Short History

The milestones of Figure 1.1 follow the methods of this volume through very different plants. Two came from mathematics before any controller ran them. A paper of 1954 set out the theory of dynamic programming [5], the backward recursion with which Chapter 12 plans a run and Chapter 7 values the water it holds back, and a paper of 1960 gave an automatic method of solving discrete programming problems [6], the branch and bound that schedules the pumps of Chapter 11.

> ⧖ HISTORY **1963 to 1979: Planning Ahead Moves Into Plants**
>
> Solving an open-loop optimization again at every step is older than its industrial use. A moving horizon controller was described in 1963, and a 1967 optimal control text anticipated the practice. The linear quadratic Gaussian controller, grown from Kalman's work of the early 1960s, had little impact on the process industries, whose plants brought constraints, nonlinearity, model error and performance criteria of their own. Industry built its own methods. A conference paper of 1976, summarized in 1978, described model predictive heuristic control on an impulse-response model, applied to the main fractionator of a cracking unit, a power plant steam generator and a polyvinyl chloride plant. Engineers at an oil company first applied a method of their own in 1973 and presented it as dynamic matrix control at a national meeting in 1979 [2, 7].

Optimal control reached one of this volume's plants before predictive control reached any. A paper published in 1968 asked for the run between two stations that uses the least energy and still stops at the next one at a specified instant. Because the speed limit must be observed strictly, it treated the run as a problem with a bounded state variable and obtained the optimal control law [8], the question of Chapter 12.

> ⧖ HISTORY **1983 to 1986: Limits Become Rows of a Quadratic Program**
>
> The first algorithms controlled plants without limits well, but they handled limits in an ad hoc way. A 1980 application to a cracking unit's reactor and regenerator added an equation to the model whenever a predicted input came close to its limit, to drive it back. A 1983 conference paper instead posed dynamic matrix control as a quadratic program in which input and output limits appear explicitly, and a fuller description followed in 1986 [13]. Its Hessian is positive definite for a linear plant, so the program is convex and standard codes solve it. On a pyrolysis furnace the controller set the fuel gas pressure of three burners to hold three temperatures, and it proved most profitable inside on-line optimization, moving a plant smoothly from one constrained operating point to another [2].

The later milestones are the plants of this volume. A report of December 1987 gave nonlinear models of boiler-turbine-alternator units with an extensive set of data from a 160 MW oil-fired unit, and offered both on a disk [9]. Chapter 4 runs its model. In October 2003 an occupied office held its zones at the low end of the comfort range until 2 pm and then let them float, and its chiller power fell by 80 to 100 percent from 2 to 5 pm without a comfort complaint [10], the band that Chapter 9 spends. A 2012 paper set out the fundamentals of economic model predictive control [11], the form of Section 1.4.8. At one flood-control reservoir, releases recommended from a 15-day ensemble streamflow forecast were tried in operation, and in the 2020 water year, the third driest in a 127-year record, the lake ended the winter with 19 percent more water stored [12], the forecast that Chapter 7 plans on.

## 1.3 Notation and Conventions

The decision index is $k$, one control period apart, and the controller plans $N$ periods ahead, the prediction horizon. The plan moves the input $M$ times, the move horizon, often in blocks of several periods. The setpoint is $r$, the predicted output is $\hat{y}$ and the input change is $\Delta u$. The weight on tracking is $Q$ and the move suppression is $R$, as in Equation (1.19) of the framework section. A slack $s \geq 0$ relaxes a soft limit and carries a quadratic price $\rho$ and a linear price $\nu$. An energy or fuel price is $\pi$, a Lagrange multiplier or coordinating price is $\lambda$, a costate is $p$ and a value function is $V$. Money is in US dollars.

The chapters keep their own symbols, and several collide. Chapter 2 and Chapter 3 write the step response as $a_i$, the prediction horizon as $P$ and the move suppression as $\lambda$, and Chapter 5 and Chapter 10 also price moves with $\lambda$. This chapter writes the step response $g_i$, because the framework's $a$ is the controller action. Chapter 6 uses $\lambda_j$ for the price of heat, Chapter 12 uses $\lambda$ for the price of time and $p$ for the costate, and Chapter 7 calls the costate of stored water its water value $w$. Linear slack prices appear as $\nu$, $\sigma_i$ and $\sigma_1$ in Chapters 5, 8 and 10. The framework's $h$ is its integration step, so no symbol here uses it.

## 1.4 Prediction and Optimization on One Heated Tank

### 1.4.1 The Running Example

The running example is a tank of liquid that a process draws from, kept warm by an electric heater. The process needs the liquid at 50 °C or warmer, and the tank must stay below 80 °C. Each hour the tank keeps 80 % of its excess over the 20 °C room, and each kilowatt-hour of heat raises it by one kelvin, as shown in Equation (1.1).

$$T_{k+1} = 20 + \phi\,(T_k - 20) + u_k - d_k, \qquad \phi = 0.8, \quad 0 \leq u_k \leq 20\,\mathrm{kW}, \quad 50 \leq T_k \leq 80\,^{\circ}\mathrm{C} \tag{1.1}$$

where $T_k$ is the temperature in °C at hour $k$, $u_k$ the heater power in kW held over the hour and $d_k$ the heat leaving with the drawn liquid, 2 kW on an ordinary hour. The equation is the exact hourly sample of a first-order store with a steady gain of $K = 1/(1-\phi) = 5\,\mathrm{K/kW}$ and a time constant of 4.48 h, a fast loss chosen for clean arithmetic rather than for a particular product. Holding 50 °C takes

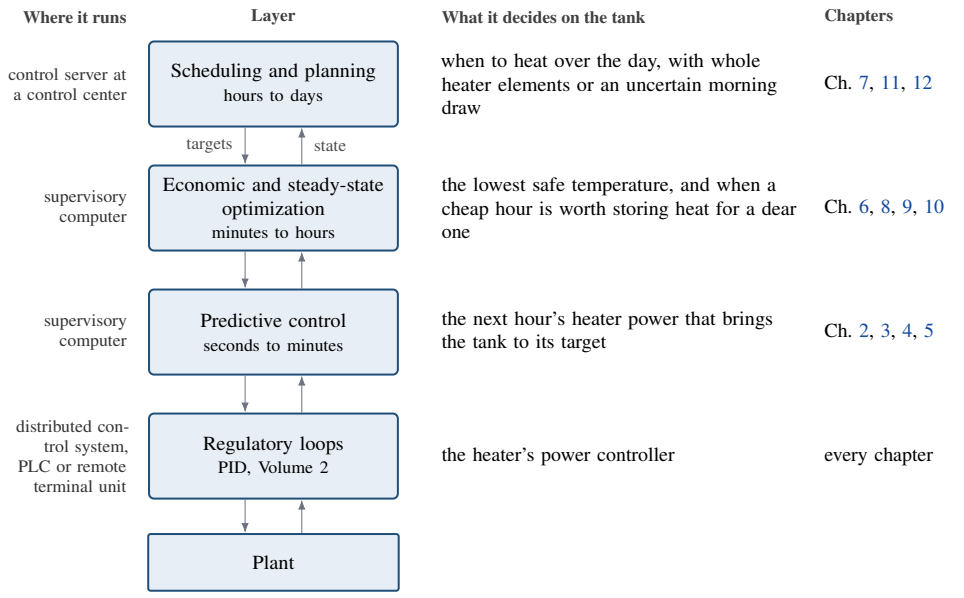



**Figure 1.2** Layers of a Control System, Where Each Runs, What Each Decides on the Running Example and the Chapters That Work in It. Down Arrows Carry Targets, Up Arrows Carry Measurements and Limits. Platforms from [2, 14] and the chapters' technology sections.

$u = 8\,\text{kW}$. Electricity costs \$0.10 per kWh from 00:00 to 06:00, \$0.40 from 17:00 to 21:00 and \$0.20 otherwise, and the draw rises to 6 kW from 07:00 to 09:00 and from 18:00 to 21:00.

Figure 1.2 places the volume's chapters in the layers of a plant's control system and states what each layer decides on the tank. The regulatory layer is the subject of Volume 2. The layers above it are this volume's subject, and the sections below climb them in order.

### 1.4.2 Prediction Models

A predictive controller needs the effect of a move on the future output. The oldest form is the step response, the output change $i$ periods after a unit step of the input. For the tank it is Equation (1.2).

$$g_i = \frac{1-\phi^i}{1-\phi} = 5\left(1-0.8^i\right), \qquad g_1 = 1,\ g_2 = 1.8,\ g_3 = 2.44,\ g_4 = 2.952\ \text{K/kW} \tag{1.2}$$

The prediction splits into a free response $f$, what the output would do if the input stayed where it is, and a forced response from the planned moves, as shown in Equation (1.3).

$$\hat{y} = f + G\Delta u, \qquad G_{ij} = g_{i-n_j} \text{ for } i > n_j, \text{ else } 0 \tag{1.3}$$

where $\hat{y}$ stacks the $N$ predicted outputs, $\Delta u$ stacks the $M$ moves and move $j$ is made at period $n_j$. The matrix $G$ is the dynamic matrix of dynamic matrix control (DMC). The same prediction follows from a state-space model such as Equation (1.18) by repeated substitution, and Equation (1.1) is that model with one state. A step response is easy to obtain from a plant test and needs no physics. A state-space model carries a state that an estimator can correct and handles an unstable or integrating plant without a truncated response.

The chapters use every form. Chapter 2 identifies a step response once from a reduced copy of its model, and Chapter 3 writes one in closed form from a fitted first-order model, with two candidates from two step tests on the same weld. Chapter 5 fits an integrator with a dead time to each reach. Chapter 4 linearizes its balances at the estimate at every decision, Chapter 10 linearizes along the

current plan, and Chapter 6 adds a predictor built only from a building's recorded data. In each case the controller's model is kept apart from the plant's, so a parameter the reader changes is a model error, as it would be on a real plant.

> ✦ DID YOU KNOW? **A Controller Whose Clock Followed the Load**
>
> In one of the first predictive control applications, on a power plant steam generator, the response time of the process varied inversely with the load. The controller met that nonlinearity by executing with a variable sample time [2]. Chapter 4 meets a gain that changes with load by linearizing its model at every decision instead.

### 1.4.3 The Quadratic Program and the Receding Horizon

The plan minimizes the tracking error over the horizon plus a price on the moves, the objective of Equation (1.19). Substituting the prediction gives a quadratic function of the moves, the quadratic program (QP) of Equation (1.4).

$$\min_{\Delta u} (r - f - G\Delta u)^\top Q (r - f - G\Delta u) + R\Delta u^\top \Delta u \quad \text{s.t.} \quad u_{\min} \le u \le u_{\max}, \;\; y_{\min} \le \hat{y} \le y_{\max} \tag{1.4}$$

Its Hessian $G^\top QG + RI$ is positive definite whenever $R > 0$, so the program has one minimum. With no limit active the minimum is linear in the predicted error, as shown in Equation (1.5).

$$\Delta u = \left(G^\top QG + RI\right)^{-1} G^\top Q (r - f) = K_{\text{mpc}} (r - f) \tag{1.5}$$

The gain matrix $K_{\text{mpc}}$ can be computed once, which is how the original DMC ran [1]. Only the first move is applied. At the next period the output is measured, the prediction is corrected and the plan is made again, the receding horizon of every predictive controller in this volume [15].

With one move ($M = 1$), unit weights and no move suppression, Equation (1.5) reduces to $\Delta u = \sum g_i e_i / \sum g_i^2$, where $e_i = r - f_i$. For a setpoint step from 50 to 55 °C on the tank, every $e_i$ is 5 K. A horizon of one hour gives $\Delta u = 5/g_1 = 5\,\text{kW}$, the move that reaches the new setpoint in one hour. A horizon of three hours gives $5 \times 5.24/10.19 = 2.57\,\text{kW}$. As $N$ grows the move tends to $5/K = 1\,\text{kW}$, the move that reaches the setpoint only at steady state.

When a limit is active the program has no closed form. It is solved by an active-set or interior-point method, a dual active-set method in the framework's apps (Section 1.6.3), and the solution is not the unconstrained plan clipped to the limits. The difference matters most for output limits, which clipping cannot honor at all, and for plans with several moves, where a limit on a later move changes the best first move [13]. A day's plan for the tank with $N = 24$ has 24 moves, 24 input bounds and 48 temperature bounds, a program the browser solves in milliseconds.

Chapter 2 and Chapter 3 compute the gain matrix once and handle the input box with a few projected-gradient steps or a clamp. Chapter 4, Chapter 5, Chapter 9 and Chapter 8 solve the full program with limits on outputs at every decision.

### 1.4.4 Horizons and Tuning

Table 1.1 runs the one-move controller on the tank for the 5 K step, with the output-bias correction of Section 1.4.5, and Figure 1.3 shows three of the runs. A short horizon is fast and fragile. With $N = 1$ the tank settles in one hour, but a plant whose gain is twice the model's sits on the stability limit. A long horizon is slow and robust: with $N = 48$ the move is barely more than the steady move, the loop takes 13 hours to stay within 0.25 K, and the plant's gain may be 9.5 times the model's before the loop goes unstable. The horizon acts as a tuning knob, much as the closed-loop time constant does in a PI tuning rule.

The last column of Table 1.1 is the honest one. A gain error of 1.7, the ratio between the two step tests of Chapter 3, overshoots by 5 to 9 K at every horizon. A longer horizon keeps the loop stable

**Table 1.1** One-Move DMC on the Tank for a 5 K Setpoint Step, by Prediction Horizon ($R = 0$)

| Horizon $N$ (h) | First move (kW) | Within 0.25 K after (h) | Plant gain tolerated | Overshoot at gain ×1.7 (K) |
|---|---|---|---|---|
| 1 | 5.00 | 1 | ×2.00 | 9.1 |
| 2 | 3.30 | 3 | ×3.03 | 6.2 |
| 3 | 2.57 | 5 | ×3.89 | 5.1 |
| 6 | 1.74 | 8 | ×5.75 | 6.0 |
| 12 | 1.31 | 10 | ×7.64 | 6.9 |
| 48 | 1.05 | 13 | ×9.50 | 7.6 |

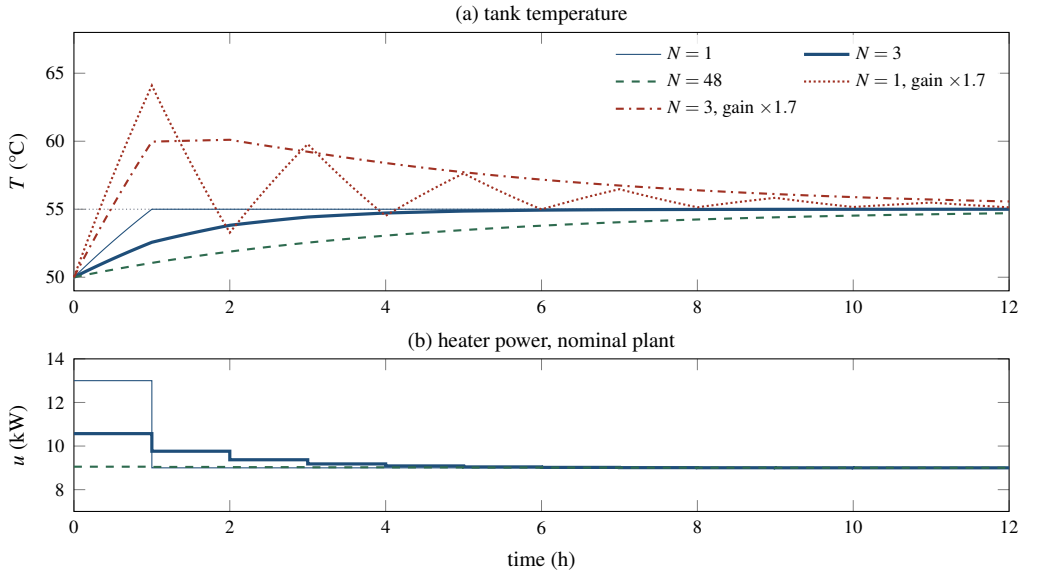



**Figure 1.3** A 5 K Setpoint Step Under One-Move DMC. Thin Solid $N = 1$, Thick Solid $N = 3$, Dashed $N = 48$ on the Nominal Plant. Dotted ($N = 1$) and Dash-Dotted ($N = 3$) on a Plant Whose Heater Gain Is 1.7 Times the Model's. The Dotted Run Drives the Heater to Zero in Its Second Hour.

but does not make the first move right, because the first move is sized by the model's gain. Move suppression has the same effect as a longer horizon: with $N = 3$, $R = 10$ halves the first move to 1.30 kW and raises the tolerated gain to 10.8. Several moves ($M > 1$) with $R = 0$ return the deadbeat plan, which is why practice pairs a long horizon with a few blocked moves and a small $R$ [2].

Chapter 2 finds a horizon too long for its model, whose linear step response fails over long horizons, and fixes it by shortening the horizon and weighting the end. Chapter 3 scales $R$ to the model so that one setting means the same aggressiveness for both of its models, and Chapter 5 and Chapter 10 fix the horizon by its physics first and tune the prices afterward.

### 1.4.5 Offset-Free Tracking

A predictive controller has no integral term, so it removes a steady error only if its prediction learns about the error. Let the draw rise from 2 to 4 kW at hour 2 and not be measured. Figure 1.4(a) shows three answers with $N = 3$. A controller that predicts from the measured temperature but keeps the nominal draw settles 3.89 K low, and its offset grows with the horizon, to 9.50 K at $N = 48$. Two disturbance models remove it [16, 17]. The output bias of DMC adds the present prediction error $b_k = y_k - \hat{y}_{k|k-1}$ of an open-loop model to every point of the horizon. An input-disturbance estimate corrects the draw itself, $\hat{d}_k = \hat{d}_{k-1} - L(y_k - \hat{y}_{k|k-1})$ with $L = 0.5$ here, which is the disturbance state

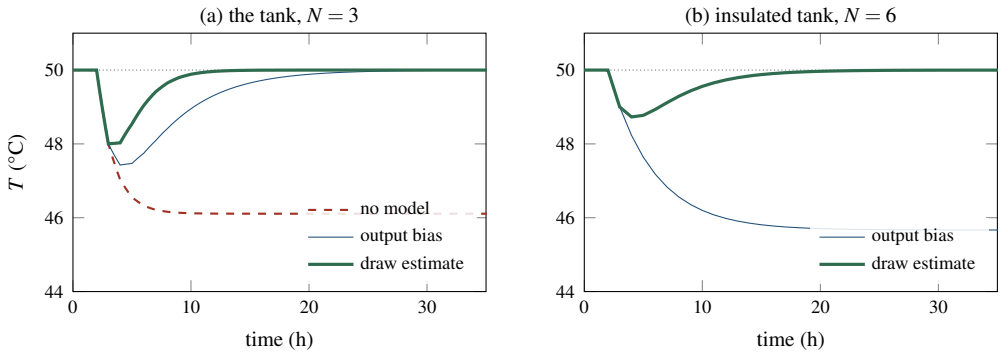



**Figure 1.4** An Unmeasured Rise in the Draw at Hour 2. (a) The Tank of Equation (1.1), Draw 2 to 4 kW: No Disturbance Model (Dashed), Output Bias (Thin Solid), Draw Estimate (Thick Solid). (b) The Tank With $\phi = 1$, Draw 2 to 3 kW: the Output Bias (Thin Solid) Leaves a Steady Offset.

of the Kalman filter in Section 1.6.3. Both settle at 50 °C. The bias needs 15 hours to come within 0.25 K, because it treats a load at the input as a step at the output that the model's 4.48 h lag then releases slowly. The estimator needs 7 hours.

The output bias fails on an integrating plant. With perfect insulation ($\phi = 1$) the tank integrates, $g_i = i$, and a draw 1 kW above the model makes the open-loop model drift away from the plant by $\delta = 1$ K an hour. A constant bias cannot represent a drift, and setting the one-move law to zero at steady state gives the offset of Equation (1.6).

$$T_\infty - r = -\delta \frac{\sum_{i=1}^{N} i\, g_i}{\sum_{i=1}^{N} g_i} = -\delta \frac{2N+1}{3} \tag{1.6}$$

At $N = 6$ it is 4.33 K, as Figure 1.4(b) shows. The draw estimate settles at 50 °C, and a filtered ramp of the bias, the drift term of Chapter 2, gives the same trajectory here. That chapter measures the same failure on its integrating altitude. Chapter 4 adds a fuel-quality disturbance to its Kalman filter and checks that the augmented model stays observable. Chapter 5, Chapter 9 and Chapter 10 estimate an unexplained inflow or load as a random walk, and Chapter 3 keeps the output bias on a self-regulating temperature.

### 1.4.6 Hard and Soft Constraints

An input limit is physical and always holds, so it is written as a hard constraint. An output limit is a wish that a disturbance can make impossible. Suppose a cleaning cycle draws 16 kW from 03:00 to 06:00. From 50 °C, full heat gives 48.0, 46.4 and 45.1 °C at the next three hours, so a plan that insists on $T \geq 50$ has no feasible point and returns nothing [18]. The usual remedy relaxes the limit with a slack, as shown in Equation (1.7).

$$\hat{y}_k \geq y_{\min} - s_k, \qquad s_k \geq 0, \qquad J \leftarrow J + \sum_k \left(\rho\, s_k^2 + \nu\, s_k\right) \tag{1.7}$$

With the slack the plan heats at full power and the tank falls to 45.1 °C, the least violation the heater allows. A plan that sees the cleaning cycle from midnight does better. It needs 59.5 °C at 03:00, reaches it by heating at 17.5 kW from 02:00, and holds the floor throughout at full power, for 5.6 kWh more heat than the plan that breaks it. Preview turns an infeasible limit into a feasible one.

The price decides when the limit gives way. At steady state one more kelvin costs $\pi(1-\phi) = \$0.04$ an hour at \$0.20 per kWh, the Lagrange multiplier of the floor. A quadratic price alone lets the plan sit below the limit by $\lambda/2\rho$, 2 K at $\rho = \$0.01$ per kelvin squared and hour, whenever saving energy

is worth it. A linear price larger than every multiplier makes the soft limit exact: it is broken only when it cannot be held. The largest multiplier sits at the last hour of the horizon, where nothing later pays for the heat, and is the full price, \$0.20 per kelvin. On a 24-hour plan $\nu = 0.19$ lets the tank sag to 42 °C at the end, and $\nu = 0.21$ holds 50 °C throughout.

Chapter 5, Chapter 8, Chapter 10 and Chapter 6 price their slacks with both terms. Chapter 4 keeps its pressure band soft and adds tube MPC, which tightens the nominal limits by the largest gap a bounded disturbance can open. Chapter 11 adds slacks only after a hard program fails, and flags the plan.

### 1.4.7 Steady-State Optimization and Active Constraints

Above the predictive layer, an optimizer chooses the setpoints. At steady state the tank's cost is linear in its temperature, as shown in Equation (1.8).

$$c(T) = \pi u = \pi\left[(1-\phi)(T-20)+d\right], \qquad \lambda = \frac{\mathrm{d}c}{\mathrm{d}T} = \pi(1-\phi) \tag{1.8}$$

At \$0.20 per kWh it is \$1.60 an hour at 50 °C and \$2.00 at 60 °C, so the optimum lies on the floor, an active constraint, and $\lambda =$ \$0.04 per kelvin and hour is what the constraint costs. This is the usual outcome of a steady-state economic optimization, and the optimum often moves from one constraint to another as prices and weather change [19, 20].

A setpoint on an active constraint cannot be held exactly. Let the draw vary by 1 kW (one standard deviation) from hour to hour. Under the draw estimator with $N = 3$ the temperature varies by 1.09 K. Placing the setpoint two standard deviations above the floor keeps the tank below 50 °C for 2.3 % of the hours, at a back-off of 2.17 K that costs \$2.09 a day. An aggressive tuning ($N = 1$, $L = 1$) varies by 1.41 K and costs \$2.71. No hourly feedback goes below 1 K, because a draw that arrives within the hour is past before the heater can answer it. Reducing variance and moving the setpoint closer to the limit is the most common way advanced control earns its keep [21].

Chapter 8 searches a grid of two setpoints every three minutes, reports which constraint binds and watches it move with the wet bulb. Chapter 10 holds a permit ceiling of 2.2 g N/m$^3$ inside a limit of 2.5, a back-off, and Chapter 4 chooses a sliding pressure as a steady-state decision.

> **DID YOU KNOW? The Money Is Made One Level Up**
>
> The first predictive control papers argued that the benefit comes less from steadier control than from the level above it, where steadier control lets a setpoint move closer to a limit. A main fractionator gained \$150,000 a year that way, by raising the flow of its light product [2].

### 1.4.8 Economic MPC and the End of the Horizon

A tracking controller holds a setpoint chosen above it. An economic controller puts the cost itself in the objective and lets the output go wherever the cost says, inside its limits [11, 22]. For the tank the objective is $\sum_k \pi_k u_k$ subject to Equation (1.1) and its limits, a linear program.

Heat bought at hour $j$ and needed at hour $n$ loses a fifth of its excess each hour in between, so its effective price is that of Equation (1.9).

$$\pi_{j\to n} = \frac{\pi_j}{\phi^{\,n-j}} \tag{1.9}$$

For the evening peak at 17:00 the effective prices of the hours 14, 15 and 16 are \$0.39, \$0.31 and \$0.25, all below \$0.40, while 13:00 at \$0.49 is not. The plan follows: it holds the floor through the day, heats at full power for two hours before each price rise, lets the tank coast to 71.6 °C and back to the floor and resumes heating as the stored heat runs out (Figure 1.5). It buys 229.2 kWh against 212 kWh for holding the floor and pays \$44.35 against \$46.40, 4.4 % less. Economic control spends more energy to spend less money.

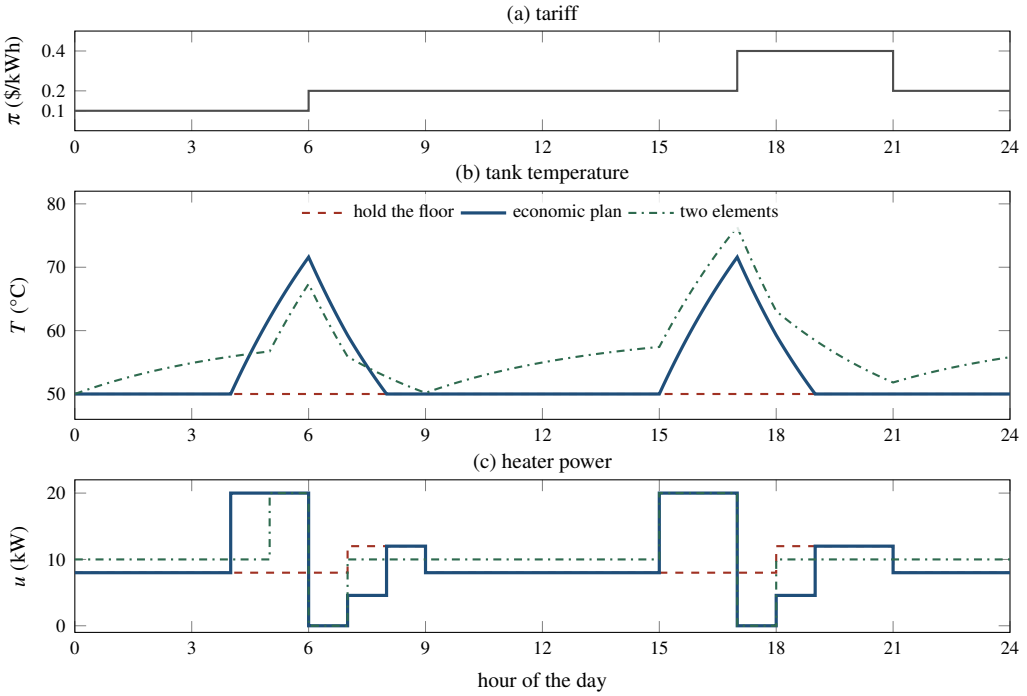



**Figure 1.5** One Day of the Tank. Dashed: Holding the 50 °C Floor With a Continuous Heater. Thick Solid: the Economic Plan With a Continuous Heater. Dash-Dotted: the Best Schedule of Two 10 kW Elements. The Draw Rises From 07:00 to 09:00 and From 18:00 to 21:00.

The horizon must reach the price rise, but no further than the stored heat lasts. Run with a receding horizon, the plan pays \$46.40 a day at $N = 1$, the same as holding the floor, \$44.60 at $N = 2$, \$44.43 at $N = 3$ and the optimum from $N = 4$ on. A longer horizon buys nothing here, because the best plan never stores heat for more than four hours before using it. A terminal value on the heat left at the horizon's end can stand in for a longer horizon. Valued at its marginal worth at the floor, $\phi\,\pi$ per kelvin, it brings $N = 2$ to the optimum but costs \$45.04 at $N = 3$, because the value of stored heat falls as the tank fills. A terminal value is only as good as the price it puts on what is left.

Chapter 9 plans a day of cooling against the tariff and a demand charge and holds the slab's stored heat near its present value at the end. Chapter 6 values heat left in the tank and the pipes at the horizon's end. Chapter 8 and Chapter 10 plan the same cost their supervisors use, and Chapter 11 ends each day's plan at the level the first plan had there.

### 1.4.9 Optimal Control and Dynamic Programming

The factor $\phi^{n-j}$ of Equation (1.9) is a discrete costate. In continuous time, attaching a costate $p$ to the tank's equation $\tau\dot{T} = 20 - T + K(u-d)$ gives the Hamiltonian of Equation (1.10).

$$H = -\pi u + \frac{p}{\tau}\left[20 - T + K\left(u - d\right)\right], \qquad \dot{p} = \frac{p}{\tau} \tag{1.10}$$

$H$ is linear in $u$, so the best input sits at a bound, full power or none, or on an arc that holds a state constraint [23]. The costate grows toward the deadline, so heat is best bought late. To heat the tank from 50 to 60 °C in six hours with the least energy, the plan holds the floor at 8 kW for five hours and heats at 18 kW in the last, 58.0 kWh. A constant 10.71 kW uses 64.3 kWh. In continuous time the switch to full power comes 0.82 h before the deadline and the energy is 57.8 kWh.

Dynamic programming reaches the same plans without the calculus. It computes the least cost from every state to the end, backward in time, as shown in Equation (1.11) [5].

$$V_k(T) = \min_{0 \le u \le u_{\max}} \left[ \pi_k u + V_{k+1}(T') \right], \qquad T' = 20 + \phi\,(T - 20) + u - d_k \tag{1.11}$$

On a 0.1 K grid of temperatures it prices the tank's day at \$44.348 against the linear program's \$44.346, and the slope of $V_k$ at the floor is the costate of each hour, \$0.08, \$0.16 or \$0.32 per kelvin. A grid handles any nonlinearity and any integer input but grows with the number of states, which is why it suits the one or two states of a train or a lake.

Chapter 12 finds this structure on a train. Its Hamiltonian is linear in the force, so the best drive is full power, holding speed, coasting and full braking, and a dynamic program on a distance and speed grid places the switches. Chapter 7 computes a value function on storage by stochastic dynamic programming and uses its slope, the water value, as the terminal price of a shorter plan.

### 1.4.10 Integer Scheduling

Many actuators come in whole units: pumps, compressors, burners, heater elements. Suppose the tank's heater is two 10 kW elements, so $u_k \in \{0, 10, 20\}$, a mixed-integer program. Its relaxation, with $0 \le u_k \le 20$ continuous, is the economic plan of Section 1.4.8 and bounds every integer schedule from below at \$44.35. Rounding the relaxed plan to the nearest element count costs \$44.00, less than the bound, because it breaks the floor in seven hours and drops the tank to 45.1 °C. A rounded plan is not a feasible plan.

Branch and bound finds the best feasible one [6]. It solves the relaxation, picks an input with a fractional value, and splits the problem into two, one with the input rounded down and one rounded up, discarding any branch whose relaxation already costs more than the best schedule found so far. On the window from 12:00 to 20:00 it proves the best of the $3^8 = 6561$ schedules, \$22.00 against a relaxation of \$19.43, after 401 relaxations. The best day with two elements costs \$49.00 (Figure 1.5), \$4.65 above the continuous heater, because 10 kW cannot hold 8 kW and the tank runs warm. The simplest controller of whole elements, the fewest elements that keep the next hour above the floor, costs \$54.00. Scheduling the integers is worth \$5.00 a day here, more than the continuous plan saves over holding the floor.

Chapter 11 schedules two fixed pumps and a variable-speed pump hourly by branch and bound, with minimum run and rest times written as integer rows. Chapter 12 chooses among eleven discrete drive moves at each stage of its dynamic program.

### 1.4.11 Distributed Plans and Prices

One plan for a whole plant can be too large, or the units can belong to different owners. A distributed controller gives each unit its own plan and coordinates them [24]. Let two tanks share a feeder of 14 kW in one hour. Tank 1 wants 10 kW and serves a forgiving process, tank 2 wants 8 kW and serves a sensitive one. Each minimizes its own discomfort $\frac{1}{2} w_i (u_i^* - u_i)^2$ plus a price $\lambda$ on the power it draws, with $w_1 = 0.25$ and $w_2 = 0.75$ in dollars per square kilowatt. Each answers $u_i = u_i^* - \lambda / w_i$, and the feeder raises the price while it is overloaded, as shown in Equation (1.12).

$$\lambda \leftarrow \max\Big[0,\ \lambda + \alpha\Big(\sum_i u_i - C\Big)\Big], \qquad \text{error factor } 1 - \alpha \sum_i \frac{1}{w_i} \tag{1.12}$$

The price settles at \$0.75 per kWh, and both tanks get 7 kW: the forgiving tank gives up three kilowatts, the sensitive one only one. First come, first served would give tank 2 only 4 kW and four times the total discomfort. The step size $\alpha$ decides the iteration. Each round multiplies the error by $1 - 5.33\,\alpha$, so $\alpha = 0.0625$ needs ten rounds to bring the overload under 0.1 kW, $\alpha = 0.1875$ needs one, and from $\alpha = 0.375$ on the price swings between two values forever. With linear costs and no quadratic term each tank takes all or nothing and no price settles.

**Table 1.2** What Each Strategy Cost on One Day of the Tank (Draws Known)

| Strategy | Energy (kWh) | Cost ($) | Lowest $T$ (°C) |
|---|---|---|---|
| Hold the floor, continuous heater | 212.0 | 46.40 | 50.0 |
| Economic plan, $N \geq 4$ | 229.2 | 44.35 | 50.0 |
| Economic plan, $N = 2$ | 216.0 | 44.60 | 50.0 |
| Fewest elements to stay above the floor | 250.0 | 54.00 | 50.0 |
| Best schedule of two elements | 250.0 | 49.00 | 50.0 |
| Relaxed plan rounded to elements | 240.0 | 44.00 | 45.1 |

Chapter 6 coordinates six substations and a plant by a price of heat in this way, with a quadratic term to make the price settle, and reports the gap between its bound and its plan. Chapter 5 divides one plan among three gates that exchange their plans until they agree, and measures how much of the central plan's result the division keeps.

### 1.4.12 Plans Under an Uncertain Forecast

A forecast is a set of possible futures, not one future [25]. On half of the days a batch draws 12 kW from 07:00 to 09:00, and the draw is seen only as it happens. The heat for 07:00 to 08:00 must be decided before then. A plan for the light morning lets the tank fall to 40 °C on a heavy one, and a plan for the mean draw to 45 °C. A scenario tree plans both futures at once, as shown in Equation (1.13).

$$\min \sum_{s} p_s \sum_{k} \pi_k \, u_{s,k} \quad \text{s.t.} \quad u_{s,k} = u_k \;\; \text{for } k < k_r, \qquad T_{s,k} \geq 50 \;\; \text{for every } s,k \tag{1.13}$$

where $s$ indexes the scenarios with probabilities $p_s$ and $k_r$ is the hour from which the scenario is known. The shared decisions before $k_r$ are the non-anticipativity constraint: the plan cannot act on what it does not yet know. The tree heats 10.6 kW in the shared hour, holds the floor on both mornings and costs $15.92 on average from midnight to noon. A controller told the morning in advance would pay $15.72, so the guarantee costs $0.20 a day. Holding a limit on only some of the scenarios, a chance constraint, trades that cost against the share of days the limit may break.

Chapter 7 plans releases on a tree of seven ensemble members with two shared days, and its chance constraint holds the channel on a declared number of them. Chapter 9 and Chapter 10 compare a plan with a forecast against one without, and the aeration plan saves energy only when it is given the typical load curve.

### 1.4.13 Where Prediction and Optimization Pay

Table 1.2 collects the tank's results. Prediction paid where the plan could see something a feedback loop could not: a price rise, a cleaning cycle, a morning that might be heavy. It also paid where the decision was combinatorial. It paid nothing when the horizon was longer than the stored heat lasts, and against an unmeasured draw within the hour it could do no better than feedback. A model error of 1.7 cost more than any tuning recovered.

Table 1.3 lists, for each chapter, the simpler control the predictive or optimizing layer is measured against and what that layer can exploit. Chapter 13 adds the measured results and asks where prediction or optimization paid for itself.

> **DID YOU KNOW? More Than 4600 Applications**
>
> A survey published in 2003 counted more than 4600 predictive control applications reported by vendors, over twice the number of five years before. Refining held 67 percent of the classified ones, and their size ran from a single variable to several hundred [2].

**Table 1.3** The Simpler Control Each Chapter Measures Against, Its Predictive or Optimizing Layer and What That Layer Can Exploit

| Ch. | Simpler control | Predictive or optimizing layer | What it can exploit |
|---|---|---|---|
| 2 | Altitude PID, three-deep cascade | DMC with a terminal weight | Dead time, integrating altitude, a touchdown limit |
| 3 | Two ITAE-tuned PI loops | DMC from a plunge or a traverse step test | A known response, if the test matches the regime |
| 4 | Boiler-following, turbine-following, coordinated PI | Offset-free MPC, tube MPC | Coupling of pressure and output, a pressure limit, fuel quality |
| 5 | Distant downstream PI with feedforward | Central and distributed MPC | Published orders, three coupled gates |
| 6 | Outdoor-compensated supply curve, scheduled boost | Central MPC, dual decomposition | Storage in pipes and tank, a delay set by the flow, fuel prices |
| 7 | Rule curve with forecast pre-release | Stochastic dynamic program, scenario tree | Ensemble forecast, flood pool, energy price |
| 8 | Fixed setpoints, grid-search supervisor | Economic MPC on the same cost | A moving active constraint, continuous decisions |
| 9 | Priced supervisory rule | Day-ahead economic MPC | Slab storage, tariff, demand charge, curtailment notice |
| 10 | Fixed oxygen, ammonia-based PI supervisor | Day-ahead MPC on a linearized model | An eight-hour permit mean, a typical load curve |
| 11 | Level-band pump rule | Hourly mixed-integer schedule | Tariff, tank storage, pressure floor, whole pumps |
| 12 | Flat out, one coast point | Dynamic program with a price of time | Time budget, grades, speed limits |

## 1.5 Instruments and Final Elements in This Volume

Figure 1.6 lists what the controllers of each chapter read and what they move. Temperature, pressure and flow are each read in five chapters, and a modulating valve or a flow loop is a final element in six. The open marks set this volume apart: eight of the eleven planners read something that no instrument on the plant measures, a forecast, a price or a timetable. The quality behind a decision is seldom read at all. Only Chapter 10 reads analyzers, oxygen probes and an ammonium analyzer whose result comes two minutes late and once every five minutes, and Chapter 4 estimates its fuel quality as a disturbance on the fuel from the drum pressure.

Several of the levels are read as a pressure. A column of liquid exerts a pressure at its bottom that is set by its height and its density, whatever the width or shape of the vessel, so a pressure measured there gives the level [26]. The tank transmitter of Chapter 11 works this way, and a density that changes, as in the boiling drum of Chapter 4, changes the reading. Air and water are divided by throttling elements. In a butterfly valve a disk turns in the stream: parallel to the flow it barely obstructs it, across the flow it blocks it, and a tight shutoff needs a soft seat [26]. Butterfly valves share the air of Chapter 10 among its basins, and gates play the same part in the canal and at the dam, as Figure 1.7 shows.

The controllers sit in the layers of Figure 1.2. A survey of industrial practice draws a plant-wide economic optimization every day, a local one every hour, predictive control every minute and the basic loops of the distributed control system every second [2]. Supervisory and lower loops run with cycle times from milliseconds to minutes, and on a network spread over a wide area a control server polls remote terminal units or programmable controllers at intervals such as 5 or 60 s and sends them new setpoints [14]. The chapters span that range: the balloon plans every 2 s, the boiler every 5 s, the data hall every minute, the canal every 300 s, the heat network every 5 min, the building and the aeration basins every 15 min, the water network every hour and the reservoir once a day. Figure 1.2 names the platform of each layer.

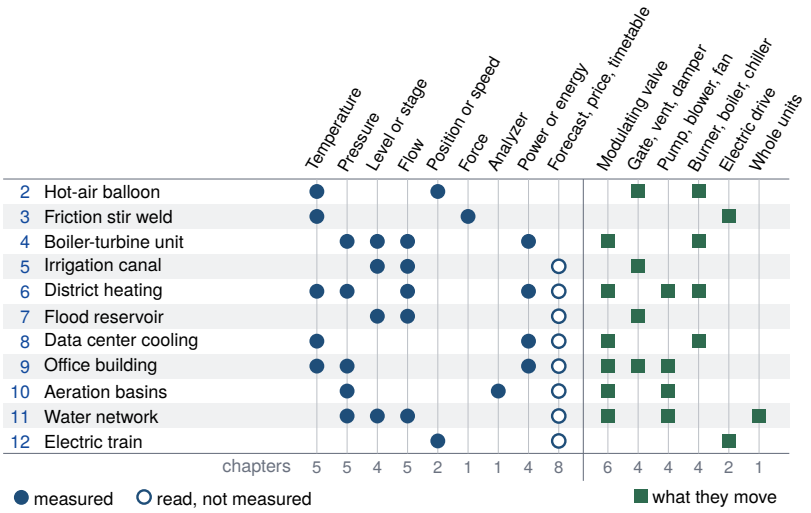



**Figure 1.6** What the controllers of each chapter read (round marks) and move (square marks), from each chapter's variables table and technology section. Pressure includes differential pressure, a flow set through a flow loop or a gate counts as the valve or the gate, the analyzer column holds the oxygen probes and the ammonium analyzer, and a computed value is not marked. Notice the open marks, which are read from a forecast, a tariff or a timetable, not from the plant.

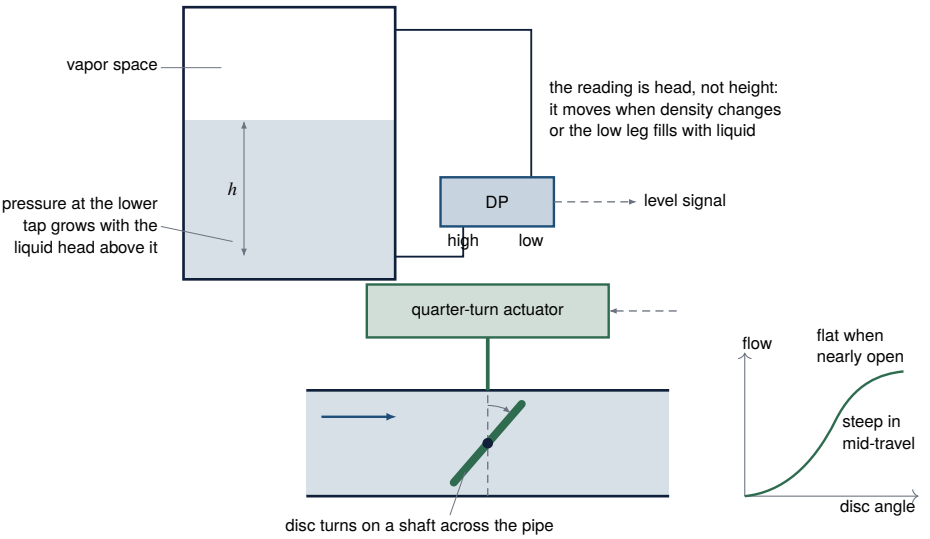



**Figure 1.7** Principle of level by differential pressure (left) and of a butterfly valve with a quarter-turn actuator (right). Notice that the pressure reading is a head, so the density of the liquid sets the level it reports.

## 1.6 The Simulator Framework

Every chapter of this book has an app that runs in a web browser, collected with the others at https://apmonitor.com/apps/. About 240 apps exist across the series. All but a few are built on one shared framework, so the engine, the PID algorithm, the trends and the data export behave the same way in every chapter. The exceptions are standalone studios, such as the cascade studio and the two Temperature Control Lab apps, which follow the same ideas in their own code. This section describes

**Table 1.4** What Every App Provides. Standalone Apps Are Described in Their Own Chapters.

| Element | What it does |
|---|---|
| Run, Pause, Reset | Start, hold and restart the run. Reset keeps the reader's settings. |
| Speed | Simulated time per second of real time. It changes the pace and never the result. |
| Manual and Auto | Hand each actuator to the reader or to the controller, without a bump. |
| Scenario list | Numbered cases, each with a briefing, a completion rule and a reload button. |
| Control panel | Setpoints, manual outputs, tuning, options and disturbances, at a Beginner or an Advanced level. |
| Trends | SP and PV above the MV for each loop, with limits, event marks and a cursor. |
| Scorecard | IAE, overshoot, settling time, MV travel and time at a limit for each loop. |
| Learn card | Predict, Try It, two hints and Explain, opened by the reader in that order. |
| Show Model | Equations, live variable values, editable parameters and assumptions. |
| Export CSV | Every logged sample of the run as a text file with a header row. |
| Instructions | A separate page with background, an operating guide and the model's limitations. |

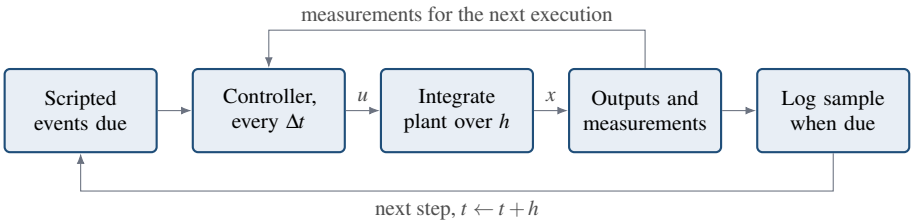



**Figure 1.8** Work of the Engine in One Integration Step

the framework once and states where the standalone apps differ. Table 1.4 lists what a reader meets.

### 1.6.1 The Simulation Engine

Each app carries a plant model written as ordinary differential equations $\frac{dx}{dt} = f(x,u,d,p)$ in the states $x$, the inputs $u$, the disturbances $d$ and the parameters $p$. The engine integrates the model with the classical fourth-order Runge–Kutta method on a fixed step $h$ set for the fastest dynamics of the app. Some apps supply their own step, such as explicit Euler for a discontinuous right-hand side or an exact update for a model driven by events.

Three clocks run at separate rates. The plant advances every step $h$. The controller executes every control period $\Delta t$, equal to $h$ unless the app declares a slower controller, and holds its outputs between executions. The log takes one sample at a fixed interval of simulated time for the trends, the scorecard and the exported file. The order of work in one step is shown in Figure 1.8. The controller also executes once when a scenario is loaded, so every run starts from an initialized controller.

Drawing the screen is kept apart from these clocks. A speed of 5 means five units of the app's time base per second of real time. Speed, frame rate and pauses change how fast a run is shown and never what it computes, so a phone and a workstation produce the same trajectory and the same file.

Noise comes from a deterministic pseudorandom generator. Its seed is declared by the app, or by a scenario that needs its own noise record, and it is restored whenever a scenario is loaded or Reset is pressed. The framework gives the reader no seed setting, so *default seed* in this book means the seed the app ships with. A run repeats bit for bit when the same scenario is loaded and the same settings are made before Run. A change made by hand during a run lands on a different step each time, which is one reason numeric answers carry a tolerance. The standalone simulators draw noise from the browser's own generator, so a noisy run there repeats in its statistics and not sample for sample.

Run and Pause start and hold the clock. Reset restarts the run at $t = 0$, paused, and keeps the reader's settings: every setpoint, tuning value, option, model parameter, disturbance slider and mode the reader changed is put back before the plant and the controller are initialized. Scripted events

replay, and a one-shot disturbance button pressed during the run is cleared. The reload button beside the scenario list restores the scenario's own settings instead.

### 1.6.2 The Velocity-Form PID Algorithm

One PID implementation serves the PID loops of every framework app. It is the velocity (incremental) form: each execution adds the change in the proportional and derivative terms to a running sum, and the integral is a separate term, as shown in Equations (1.14) to (1.17) for execution $k$ and control period $\Delta t$.

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k), \qquad P_k = K_c\,e_k, \qquad D_k = -a\,K_c\,\tau_D\,\frac{\mathrm{PV}^f_k - \mathrm{PV}^f_{k-1}}{\Delta t} \tag{1.14}$$

$$S_k = S_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \tag{1.15}$$

$$v_k = u_b + S_k + I_{k-1}, \qquad u_k = \mathrm{sat}(v_k) \tag{1.16}$$

$$I_k = I_{k-1} + \left[\frac{K_c}{\tau_I}\,e_k + \frac{u_k - v_k}{T_t}\right]\Delta t \tag{1.17}$$

where $a = \pm 1$ is the controller action, $\mathrm{PV}^f$ is the measurement after an optional first-order filter, $u_b$ is the output bias, $S$ is the running sum of proportional and derivative increments, $v$ is the output before limiting, $u$ is the output sent to the actuator, $I$ is the integral term and $T_t$ is the tracking time constant, equal to $\tau_I$ unless a loop sets another value. The function sat clamps $v_k$ to the actuator range and then applies a rate limit if the loop declares one.

The limit acts on the emitted output $u_k$ only. The sum $S_k$ is never clamped, so it equals $P_k + D_k$ and the output before limiting is the position form $u_b + K_c e_k + D_k + I_{k-1}$. Anti-windup touches the integral alone. While the output is limited, the last term of Equation (1.17) drives $I$ toward the value that reproduces the output applied, which is back-calculation [27]. With no integral action, Equation (1.17) is skipped. Clamping the running sum would instead discard each increment that crosses a limit and keep each one that returns, so symmetric noise would push the output off the limit even in a proportional-only loop.

The derivative acts on the measurement, so a setpoint step gives a proportional kick and no derivative kick. A change of $K_c$ acts on the whole error at the next execution, as in the position form. The action $a = +1$ suits a process whose PV rises with the MV, a reverse-acting controller in the ISA convention, and $a = -1$ gives the direct-acting one.

In Manual the controller keeps executing and sets $I_k = u_{\mathrm{man}} - u_b - S_k$, so the output it would send equals the manual value $u_{\mathrm{man}}$. The first automatic output differs from it by one period's change in $P$ and $D$. In Auto the engine copies each controller output to the manual station, so a switch to Manual is also bumpless. The standalone studios with PID loops use velocity-form code of their own, with the limit on the output and anti-windup on the integral alone, in which a gain change applies only to later increments.

### 1.6.3 The Predictive Control and Estimator Options

More than 100 apps offer model predictive control (MPC) as an option beside the structure they open with, which stays the default. Volume 9 develops the theory. In most of them the controller plans with a small linear model in discrete time as shown in Equation (1.18).

$$x_{k+1} = A\,x_k + B\,u_k + E\,w_k + c, \qquad y_k = C\,x_k \tag{1.18}$$

where $w$ is a measured or scheduled disturbance and $c$ is a constant left by linearization. The model comes from a step test at the design point, from the app's balances linearized there or again at every decision, or from the model the reader fits. It never reads the true states or parameters of the run, so a parameter the reader changes in Show Model is model mismatch, as on a real plant.

At each decision the controller minimizes the objective of Equation (1.19) over a horizon of $N$ steps.

$$J = \sum_{k=1}^{N} (y_k - r_k)^\top Q\,(y_k - r_k) + \sum_j \Delta u_j^\top R\,\Delta u_j + \sum_i \rho_i\, s_i^2 \tag{1.19}$$

where $r$ is the setpoint, $Q$ weights the tracking error, $R$ is the move suppression on each input change $\Delta u_j$ and $\rho_i$ is the price of a slack $s_i \geq 0$ that lets an output band be broken when it cannot be held. Actuator limits are hard constraints. The quadratic program is solved in the browser by a dual active-set method [28]. The first move is applied, and the plan is made again at the next decision.

The plan starts from an estimate of the model state. About half of these apps use a steady-state Kalman filter on the model with an added disturbance state, which absorbs what the model did not predict and removes steady offset without an integral term [17]. Others correct the model output with a filtered bias to the same end. About a dozen apps also offer moving horizon estimation (MHE), which solves the estimation problem over a window of recent samples and can hold bounds that a Kalman filter cannot, such as a concentration that is never negative [29].

With the MPC selected, the tuning fields of the loops it replaces leave the Control panel. In most of these apps the Advanced level shows up to three settings, each a multiple of the shipped value: the prediction horizon, the move suppression and the band price. The trend draws the accepted plan ahead of the present time, and a list of constraint prices shows which limits bind. A plan that is not optimal, not finite or outside its limits is rejected, and a fallback, in most apps the shipped controller, holds the actuator until a plan succeeds. The Model Fitting and Tuning Studio builds the same kind of controller from the reader's own data in its tab, MPC from the fit.

### 1.6.4 Scenarios, Progress and Data

An app teaches through scenarios, the numbered cases at the top of the Control panel, nine to thirteen in most apps. Loading one sets the model parameters, the controller settings, the disturbances, the mode, the initial state and any scripted events. The information button beside the list opens a briefing in four parts: what the case is, why it matters, what to do and one detail for the practicing engineer. The Learn card holds the lesson back, with Predict and Try It before the run, hints on request and Explain afterward.

Each scenario has a completion rule that the app judges from the run on the simulated clock: a run length, the scripted events passed, a setting made by the reader, a state held for a stated time or a target met. Choosing a case earns nothing. A completed case is marked in the list, and a badge row shows the count and a level, Explorer after one case, Practitioner at half and Completed at all. Progress is kept in the reader's browser with no account. A certificate of progress or of completion can be issued from it, as a record of self-paced work and not an assessed credential.

One case, usually the last, is the graded challenge. It states a target, such as an error band to hold through a sequence of upsets. Changing the model parameters or the disturbances voids the attempt in most apps, and a record made with Reveal suggested tuning says so.

Show Model opens the equations, live values of the variables, the editable parameters and the assumptions. The trends plot SP and PV above the MV for each loop, with actuator limits and marks at setpoint moves, disturbances, mode changes and scripted events. The scorecard reports for each loop the integral of the absolute error (IAE), overshoot, settling time, MV travel and share of time at a limit, with the app's own measures, and it does not combine them into one score. IAE and travel accumulate from Reset, so two runs are compared at the same clock reading.

Export CSV saves the log as a text file with a header row of column names that carry their units and one row per logged sample. In most apps the time column is in seconds, minutes or hours according to the run length. The log holds the most recent 60 000 samples unless the app sets another size. Apps with the MPC option begin the file with comment lines that start with #, which a script must skip.

An app exercise in this book names the scenario by its number and name as the app shows them, every setting that differs from the scenario's defaults, the run length in simulated time, the seed

(the default seed unless stated), what to read and where (a trend, a scorecard tile or a column of the exported file) and a tolerance on a numeric answer. Settings are made before Run unless the exercise gives a time.

## 1.7 Working With AI Agents While Studying

A reader of this book is likely to have an AI assistant within reach, and the apps make its work easy to test. This section gives the editor's view of how to use one while working through a chapter. In short, the assistant may do the algebra and the scripting, and the reader keeps the checking.

**What to delegate.** Tasks with a clear input and a result that can be tested are good candidates. An assistant can derive the linearization of a chapter's balance at an operating point and the transfer function that follows from it. It can write a short script that reads the exported CSV file and fits a first-order-plus-dead-time model to a step test. It can propose a first tuning from that fit with a named rule, convert units, explain a term in a model or outline a report. The assistant knows nothing about the app until it is told. Give it the equations and parameter values from Show Model, the column names of the exported file and the scenario that produced the data, as described in Section 1.6.4. Ask it to state its assumptions and the units of every result.

**What to check, and how.** An assistant returns a fluent answer whether or not the answer is right, so each result needs a test that it could fail. Four tests cover most of the errors a chapter will produce.

- *Units.* Every term of an equation carries the same units, and a process gain is in PV units per MV unit. A gain computed with the valve as a fraction is 100 times the gain of a loop whose valve is in percent.
- *Signs.* Decide from the physics which way the PV moves when the MV rises. Compare that with the sign of the fitted gain and with the controller action of Section 1.6.2.
- *Steady state.* Set the derivatives to zero. The model must return the operating point that Show Model reports, and a linearization must give zero deviation there.
- *A run on the app.* Load the named scenario at the default seed, enter the proposed numbers and run. The trend and the scorecard decide. A tuning that reads well on paper and holds the valve at a limit for half the run has failed.

A fitted model gets one more test: compare it with data it was not fitted to, such as a second step in the opposite direction. When a check fails, report the failure to the assistant with the evidence and ask for the cause, then check the correction the same way.

**How the exercises train the habit.** One exercise in each chapter is marked *Review the agent*. It presents a tuning, a model or a claim as an AI assistant might propose it, with a flaw, and asks the reader to find the flaw in about five minutes. The flaws are ordinary ones: a unit, a sign, a steady state that does not close, a result measured on the wrong scenario or at a different run length, or a limit that was ignored. Each is found with one of the tests above. The purpose is not distrust. An engineer who can say why an answer is right can hand an assistant much larger tasks and still stand behind the result. That larger picture, in which agents do most of the routine work and the engineer supervises and checks, is the subject of Chapter 13.

## 1.8 Chapter Roadmap

Table 1.5 lists the chapters with the lesson each adds to this one. The level says how much of Section 1.4 a chapter assumes. The time is an estimate for reading the chapter once and running each case of its app once, at about eight minutes a case. Figure 1.2 places the same chapters in the layers of a control system.

**Table 1.5** Roadmap of the Volume

| Ch. | Plant | The lesson | Level | Time |
|---|---|---|---|---|
| 2 | Hot-air balloon | Step-response control of an integrating plant behind a dead time, and a landing under a speed limit. | introductory | 2.5 h |
| 3 | Friction stir welder | One weld holds three plants. Which step test fills the dynamic matrix decides the result. | introductory | 3 h |
| 4 | Drum boiler and turbine | Fast megawatts borrowed from pressure. Offset-free and tube MPC on a two-input unit. | advanced | 2 h |
| 5 | Irrigation main canal | Delay, storage and resonance in each reach. One plan, three plans and plans that talk. | advanced | 2.5 h |
| 6 | District heating network | A delay set by the flow. Heat priced among substations by dual decomposition. | advanced | 2.5 h |
| 7 | Flood control reservoir | Releases decided on an ensemble forecast by a value function and a scenario tree. | advanced | 2.5 h |
| 8 | Data center cooling plant | A cost optimum on a constraint that moves with the weather. | intermediate | 3 h |
| 9 | Office building | A comfort band to spend, thermal mass, tariffs and a curtailment. | intermediate | 3.5 h |
| 10 | Activated sludge aeration | Where a supervisory setpoint belongs under a permit that is a mean. | intermediate | 3.5 h |
| 11 | Town water network | Integer pumps scheduled under a pressure floor and a tariff. | intermediate | 2 h |
| 12 | Electric train | Where to coast, from optimal control and a dynamic program. | introductory | 2 h |

Chapters 2, 3 and 12 form a first course: two single-input predictive controllers and one optimal plan whose structure can be checked by hand. They rest on Sections 1.4.2 to 1.4.5 and 1.4.9. Chapters 8 to 11 are best read after Sections 1.4.6 to 1.4.10, in any order. Chapters 4, 5, 6 and 7 draw on Sections 1.4.11 and 1.4.12 as well and are best left until last. A reader who knows MPC from industrial practice can start at Section 1.4.7 and take Chapters 8, 9, 11 and 7.

**KEY POINTS**

- A predictive controller plans several moves on a model, applies the first and plans again at the next period.
- A longer horizon or a higher price on moves buys robustness at the price of speed, and neither corrects a wrong gain.
- An offset disappears only when the prediction learns about the disturbance, through a bias or an estimated input.
- Most of the money is made in the layer above: where the plant sits against its active limit, and when stored energy is bought.

**Where else this applies.** Steady-state optimization above the loops in Volume 13, Chapter 11, and a buoy planned on a fading wave forecast in Volume 17, Chapter 5.

## References

[1] C. E. García, D. M. Prett, and M. Morari, "Model predictive control: Theory and practice—a survey," *Automatica*, vol. 25, no. 3, pp. 335–348, 1989.

[2] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[3] M. Morari and J. H. Lee, "Model predictive control: Past, present and future," *Computers & Chemical Engineering*, vol. 23, no. 4–5, pp. 667–682, 1999.

[4] E. F. Camacho and C. Bordons, *Model Predictive Control*, ser. Advanced Textbooks in Control and Signal Processing. London: Springer, 2007.

[5] R. Bellman, "The theory of dynamic programming," *Bulletin of the American Mathematical Society*, vol. 60, no. 6, pp. 503–515, 1954.

[6] A. H. Land and A. G. Doig, "An automatic method of solving discrete programming problems," *Econometrica*, vol. 28, no. 3, pp. 497–520, 1960.

[7] J. Richalet, A. Rault, J. L. Testud, and J. Papon, "Model predictive heuristic control: Applications to industrial processes," *Automatica*, vol. 14, no. 5, pp. 413–428, 1978.

[8] K. Ichikawa, "Application of optimization theory for bounded state variable problems to the operation of train," *Bulletin of JSME*, vol. 11, no. 47, pp. 857–865, 1968.

[9] R. D. Bell and K. J. Åström, "Dynamic models for boiler-turbine-alternator units: Data logs and parameter estimation for a 160 MW unit," Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, Research Report TFRT-3192, 1987. [Online]. Available: https://lup.lub.lu.se/record/8602296

[10] P. Xu, P. Haves, M. A. Piette, and J. Braun, "Peak demand reduction from pre-cooling with zone temperature reset in an office building," Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-55800, 2004, in Proceedings of the ACEEE 2004 Summer Study on Energy Efficiency in Buildings. OSTI 838609.

[11] J. B. Rawlings, D. Angeli, and C. N. Bates, "Fundamentals of economic model predictive control," in *Proc. 51st IEEE Conference on Decision and Control*, 2012, pp. 3851–3861.

[12] Lake Mendocino FIRO Steering Committee, "Lake Mendocino forecast informed reservoir operations final viability assessment," UC San Diego, Tech. Rep., Dec. 2020. [Online]. Available: https://cw3e.ucsd.edu/FIRO_docs/LakeMendocino_FIRO_FVA.pdf

[13] C. E. García and A. M. Morshedi, "Quadratic programming solution of dynamic matrix control (QDMC)," *Chemical Engineering Communications*, vol. 46, no. 1–3, pp. 73–87, 1986.

[14] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[15] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[16] K. R. Muske and T. A. Badgwell, "Disturbance modeling for offset-free linear model predictive control," *Journal of Process Control*, vol. 12, no. 5, pp. 617–632, 2002.

[17] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[18] P. O. M. Scokaert and J. B. Rawlings, "Feasibility issues in linear model predictive control," *AIChE Journal*, vol. 45, no. 8, pp. 1649–1659, 1999.

[19] M. L. Darby, M. Nikolaou, J. Jones, and D. Nicholson, "RTO: An overview and assessment of current practice," *Journal of Process Control*, vol. 21, no. 6, pp. 874–884, 2011.

[20] S. Skogestad, "Plantwide control: The search for the self-optimizing control structure," *Journal of Process Control*, vol. 10, no. 5, pp. 487–507, 2000.

[21] M. Bauer and I. K. Craig, "Economic assessment of advanced process control—a survey and framework," *Journal of Process Control*, vol. 18, no. 1, pp. 2–18, 2008.

[22] M. Ellis, H. Durand, and P. D. Christofides, "A tutorial review of economic model predictive control methods," *Journal of Process Control*, vol. 24, no. 8, pp. 1156–1178, 2014.

[23] A. E. Bryson and Y.-C. Ho, *Applied Optimal Control: Optimization, Estimation, and Control*. Routledge, 2018.

[24] R. Scattolini, "Architectures for distributed and hierarchical model predictive control—a review," *Journal of Process Control*, vol. 19, no. 5, pp. 723–731, 2009.

[25] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Systems Magazine*, vol. 36, no. 6, pp. 30–44, 2016.

[26] T. R. Kuphaldt, *Lessons in Industrial Instrumentation*, version 2.33 ed. Open textbook, Creative Commons Attribution 4.0 license, ibiblio.org/kuphaldt/socratic/sinst, 2024.

[27] K. J. Åström and L. Rundqwist, "Integrator windup and how to avoid it," in *Proc. 1989 American Control Conference*, 1989, pp. 1693–1698.

[28] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[29] C. V. Rao, J. B. Rawlings, and D. Q. Mayne, "Constrained state estimation for nonlinear discrete-time systems: Stability and moving horizon approximations," *IEEE Transactions on Automatic Control*, vol. 48, no. 2, pp. 246–258, 2003.

## 1.9 Exercises

**1.1** Theory and calculation · 4 min

The tank of Equation (1.1) keeps $\phi = 0.8$ of its excess over the room each hour and gains one kelvin per kilowatt-hour.

(a) Find the step-response coefficient $g_4$. *Checkpoint:* 2.952 K/kW.

(b) A one-move DMC with $N = 2$ and $R = 0$ sees a setpoint step of 5 K. Find its first move. *Checkpoint:* 3.30 kW.

(c) Find the limit of the first move as $N$ grows without bound. *Checkpoint:* 1 kW.

**1.2** Concept check · 2 min

A one-move DMC was built from a step test on the tank, but the heater on the plant has 1.7 times the gain of the model. Which statement matches Table 1.1?

(a) a horizon of one hour removes the overshoot, because the plan reaches the setpoint in one period

(b) move suppression removes the overshoot, because it shrinks every move in proportion to the gain error

(c) a longer horizon keeps the loop stable but leaves an overshoot, because the first move is sized by the model's gain

(d) the output bias removes the overshoot at once, because it adds the measured error to the prediction

**1.3** Theory and calculation · 4 min

The tank's draw rises from 2 to 4 kW and is not measured. A controller with $N = 1$ predicts from the measured temperature with the nominal draw of 2 kW and no disturbance model, and holds $r = 50\,^{\circ}\mathrm{C}$.

(a) Write the heater power it chooses as a function of the measured temperature $T$. *Checkpoint:* $u = 48 - 0.8\,T$.

(b) Find the temperature at which the tank settles. *Checkpoint:* 48 °C.

**1.4** Theory and calculation · 3 min

The insulated tank ($\phi = 1$) has $g_i = i$. Its draw is 1 kW above the model, so the open-loop model drifts from the plant by 1 K an hour, and a one-move DMC with a constant output bias holds 50 °C. Use Equation (1.6) to find the steady offset for $N = 3$, and say how it changes with the horizon.

**1.5** Concept check · 2 min

A 24-hour economic plan for the tank at \$0.20 per kWh softens the floor of 50 °C with a linear slack price $\nu$ per kelvin and hour. Which value of $\nu$ is the smallest of the four that keeps the floor wherever it can be held?

(a) \$0.21, a little above the price of one kilowatt-hour of heat

(b) \$0.04, the steady-state multiplier $\pi(1-\phi)$ of the floor

(c) \$0.05, a quarter above the steady-state multiplier of the floor

(d) \$0.01, since any positive linear price makes a soft limit exact

**1.6** Theory and calculation · 4 min

The tank's setpoint is backed off from the 50 °C floor by two standard deviations of its temperature, at \$0.20 per kWh.

(a) Find the cost of one kelvin of back-off per day from Equation (1.8). *Checkpoint:* \$0.96 per kelvin and day.

(b) The draw estimator with $N = 3$ holds the temperature to a standard deviation of 1.086 K. Find the back-off and its cost per day. *Checkpoint:* 2.17 K, \$2.09 a day.

**1.7** Theory and calculation · 4 min

The tank needs heat at 07:00, when power costs \$0.20 per kWh. Earlier hours cost \$0.10 up to 06:00 and \$0.20 from 06:00. Use Equation (1.9) to find the effective price of heat bought at 03:00, 04:00 and 05:00 for use at 07:00, and the earliest of these hours worth buying in.

**1.8** Theory and calculation · 3 min

The tank is heated from 50 to 60 °C in six hours with the least energy, with the draw at 2 kW and the floor at 50 °C. The best plan holds the floor and heats once, at the end. Find the power in the last hour and the energy saved against the constant power of 10.71 kW.

**1.9** Concept check · 2 min

The relaxed economic plan for the tank costs \$44.35 a day. Rounded to whole 10 kW elements it costs \$44.00, and the best schedule of whole elements costs \$49.00. What explains this?

(a) the rounded plan found a cheaper schedule that branch and bound missed, because it stopped early

(b) the relaxation is not a lower bound when the integer schedule must end at the same temperature

(c) whole elements are cheaper than a continuous heater, because they run at full efficiency

(d) the rounded plan breaks the floor in seven hours, so it is not a feasible schedule at all

**1.10** Theory and calculation · 4 min

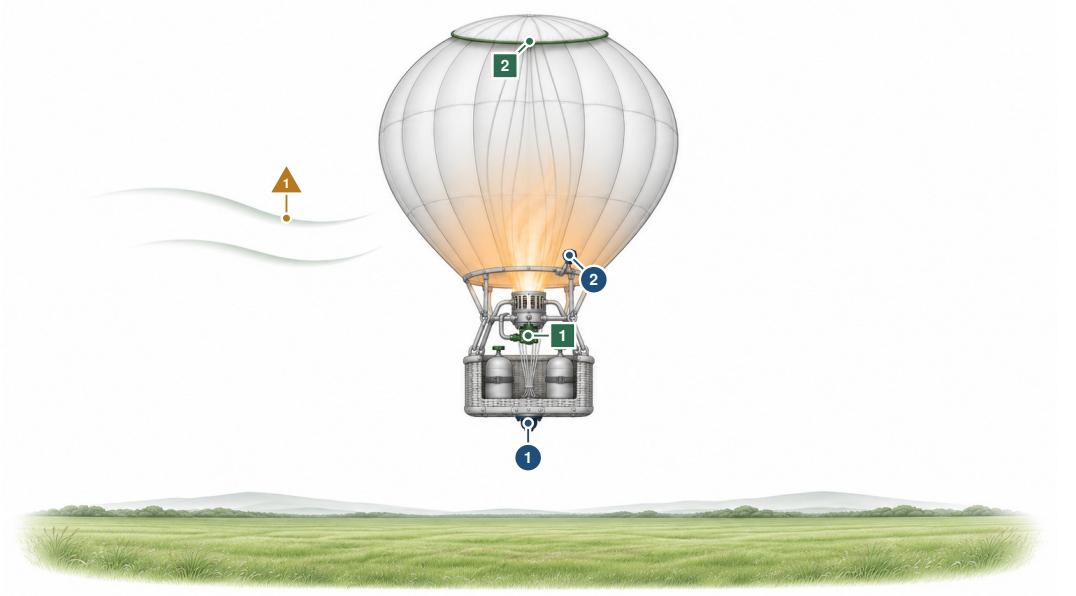
Two tanks share a feeder of 14 kW. They want 10 and 8 kW and weigh their shortfall with $w_1 = 0.25$ and $w_2 = 0.75$, so each answers a price $\lambda$ with $u_i = u_i^{*} - \lambda / w_i$.

(a) Find the price at which the feeder is exactly loaded, and each tank's power. *Checkpoint:* \$0.75 per kWh, 7 and 7 kW.

(b) Find the step size $\alpha$ of Equation (1.12) that settles the price in one round, and the step size at which it swings forever. *Checkpoint:* 0.1875 and 0.375.

CHAPTER 2

# Hot-Air Balloon Simulator for Landing an Integrating Plant Behind a Dead Time




**The controller moves** 1 burner blast valve 2 parachute vent at the crown

**The controller reads** 1 altimeter and variometer 2 envelope temperature sensor

**What disturbs it** 1 wind

*Notice that heat enters at the open mouth and is let out at the crown, and that the balloon answers its burner about half a minute late. Illustration, not to scale.*

### IN PLAIN TERMS Landing a Balloon That Answers Late

**The main idea.** A hot-air balloon rises when its burner heats the air inside it and sinks when that air cools or is let out through a vent at the top. A controller that looks ahead can plan the burns before the balloon needs them.

**The challenge.** The balloon starts to move only seconds after the flame lights, and then it keeps speeding up long after the burner stops. A gentle landing means heating while still high and falling fast, before the instruments show any need.

**Why it matters.** A pilot or controller that waits for the error to appear reaches the ground far too fast.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the fact in Section 2.1, the history in Section 2.2, the first try-it in Section 2.4, the second try-it and the second fact in Section 2.6, and the key points |
| **Study** | 55 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 20 min | the first page, the results (Section 2.6), Technology in Practice (Section 2.7), Limitations (Section 2.8) and the open challenge |

**The control question** A hot-air balloon answers its burner four seconds late and then integrates twice on the way to the altitude, so there is no steady state from which to read a gain. What can one loop, a cascade and a step-response predictive controller do with such a plant, and what does it take to land it under a touchdown-speed limit?

**Learning objectives**

- Trace a blast-valve step through the dead time, the thermal integrator and the drag lag, and explain why no process gain can be read.
- Compute the float temperature, the lift per kelvin and the steady climb rate from the buoyancy and envelope energy balances.
- Explain from the phase of the plant why one PID loop on altitude hunts, and how a three-deep cascade gives each loop one slow element.
- Describe how the dynamic matrix controller identifies its step response, builds its gain matrix once and corrects a drift with a bias and a ramp.
- Choose a prediction horizon by where the identified model stops being true, and tell a model error from an aggressive optimizer.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.2 and 1.4.3 (step response and gain matrix), 1.4.4 (horizons and move suppression), 1.4.5 (the bias on an integrating plant) and 1.6.4 (how an exercise names a run).

**The app** https://apmonitor.com/apps/hot-air-balloon/

**See also** Volume 5, Chapter 2 for a burner whose fuel arrives behind a transport delay, and Volume 5, Chapter 3 for a predictive controller whose model omits part of the delay. Volume 3, Chapter 2 for matching the loop structure to where a disturbance enters, and Volume 3, Chapter 10 for another descent that must arrive slowly.

## 2.1 Introduction

The pilot of a hot-air balloon has one continuous way to change altitude, which is to change the temperature of the air held in the envelope. Opening the blast valve does not move the balloon and does not move the variometer. It lights a flame three meters tall inside an envelope twenty meters across, and the plume has to rise and mix before the mean temperature of the envelope changes at all. On the simulator of this chapter that takes 4 s. What follows the delay is worse than the delay. The envelope temperature is an integrator against a 646 s loss time constant, one kelvin of envelope temperature is worth 75.96 N of lift against 5356 kg of effective mass, quadratic drag turns that imbalance into a climb rate over about 33 s, and the altitude is the integral of the climb rate. A step on the blast valve moves the delivered heat at 4.0 s and the variometer past 0.5 m/s at 24.2 s, and two minutes later the balloon is 264 m higher and still accelerating.

> ✦ **DID YOU KNOW? Six to Fifteen Seconds Before the Balloon Answers**
>
> Flight training material warns that a pilot's input can take from 6 to 15 seconds to show as a reaction of the balloon, because the air inside the envelope is always moving [1]. In the simulator, the variometer passes 0.1 m/s at 13.0 s after a blast-valve step (Table 2.3), inside that range.

The shape of that plant is general. An output that integrates, an actuator that is both delayed and lagged, and a flight or a batch that ends on a hard limit rather than on a setpoint, appear together in ship steering, where rudder angle sets a yaw rate and heading is the integral [2], in building thermal control, where a heating plant with a long time constant drives a room whose temperature has no equilibrium of its own within the comfort band [3], and in every powered descent that must arrive at a fixed touchdown speed, such as the rocket of Volume 3, Chapter 10. An output that integrates behind a long delay is awkward enough that the standard dead-time compensator had to be reworked before it would hold such a plant at all [4], and each reach of the canal in Chapter 5 is such a plant, an integrator behind a wave travel time. The correct answers are structural rather than numerical.

Either the loop is broken into a cascade whose inner loop closes around the fast dynamics [5], or a predictive controller is given the delay and the integration explicitly [6, 7].

A block diagram states all of that and demonstrates none of it. Reading $Ke^{-\theta s}/s^2$ on a page does not produce the specific and counterintuitive experience of holding a burner open for five seconds from trim, seeing nothing happen, setting the valve back where it was, and then finding the balloon 586 m higher seven minutes after the burn began, 485 m higher than the same balloon left alone, because heat that has been bought cannot be given back. The misconception the plant corrects is that a manipulated variable set correctly holds the process where it is. Setting the burner to the heat input that exactly matches the fabric loss freezes the envelope temperature, and freezing the envelope temperature freezes whatever vertical speed the balloon already had. Trimming the actuator is not trimming the plant. Assessments of control education published over the past decade place experimental work with an unknown plant near the center of what a course should provide [8, 9], and a plant with no steady state to read a gain from is where that argument is strongest.

**Pause.** Try Exercise 2.1 before reading on.

Simulators for predictive control usually hide the controller behind a solver call, so the student sees a horizon and a tuning weight and never sees the free response being shifted, the bias being added or the constraint being enforced. This chapter works with a hot-air balloon simulator in which a step-response predictive controller is written out in full, from the step test that identifies it to the projected-gradient iteration that keeps its moves inside the actuator box, and it reports two results that the implementation forced. Section 2.2 traces balloon flight and its altitude control. Section 2.3 gives the nine-state flight model with its split-range actuator, its real transport delay and its terminal constraint on the arrival speed, with every equation and constant needed to rebuild it. Section 2.4 follows a blast-valve step through the chain of lags to the altitude, and Section 2.5 sets out the single loop, the cascade and the dynamic matrix controller that act on the same demand. Section 2.6 compares them on a level-off and on a graded landing that the single loop cannot pass at any gain, quantifies what a linear step-response model costs on a plant whose climb rate saturates as $\sqrt{\Delta T}$, and shows that a constant output bias is not offset-free when the output integrates. Section 2.7 describes the instruments, the burner and the platforms of real balloons and predictive controllers. Sections 2.8 to 2.10 state the limitations, the lessons that transfer and a summary. The general form of dynamic matrix control, its gain matrix and the failure of the bias on an integrating plant are set out in Sections 1.4.2 to 1.4.5, and this chapter keeps what is specific to the balloon.

## 2.2 Balloon Flight and the Control of Altitude

Manned flight began with a hot-air balloon [10], as the history segment below tells.

> **HISTORY 1783: Two Kinds of Balloon in One Year**
>
> The first recorded manned balloon flight was made on November 21, 1783, in a hot-air balloon built by the Montgolfier brothers of France. Pilâtre de Rozier and the Marquis d'Arlandes flew it for 23 minutes and 5.5 miles. Ten days later the physicist Jacques Charles flew the first gas balloon, a varnished silk envelope filled with hydrogen, for two hours and 27 miles, so the two kinds of balloon still flown today appeared in the same year. The modern hot-air balloon grew from a contract of the United States Navy in the 1950s for a small hot-air balloon for military purposes, a balloon the Navy never used [1]. Its propane burner, its fabric envelope and the vent in its crown are the actuators and the plant of this chapter.

What returned with it was an aircraft with two controls and no steering. The burner adds heat at a rate the pilot chooses, and the parachute vent, a disc in the crown of the envelope on a line to the basket, dumps hot air and lets cold air in. The two are never used together. Fuel is the binding resource, and the fuel a flight profile costs was already being computed from the envelope heat balance in the early lighter-than-air literature [11]. Gas balloons solved the same problem with a

different pair of actuators, a valve at the top that releases lifting gas and sand ballast over the side, and both of those are one-way and finite.

Altitude control on a manned balloon is still manual and still open loop in the strict sense. The instrument panel is an altimeter, a variometer, an envelope thermocouple at the crown with a fusible tell-tale, and a fuel gauge, and the technique taught in the certification literature is to anticipate, to burn in short bursts, and to fly the envelope temperature rather than the altitude [12]. The same source sets the standards the landing in this application is graded against, because an approach is flown to a target sink rate and a target ground speed rather than to a target position. The persistence of hand flying is not conservatism. It is a rational response to a plant on which a correction placed thirty seconds late is worse than no correction at all, made by an operator who can see the ground.

Automatic altitude control on balloons is real, current and unmanned. A superpressure balloon holds a fixed volume and trims its altitude with ballast, a pump moving air into and out of an internal ballonet, or a valve, and the loop that decides those actions runs for weeks without a pilot [13]. Stratospheric station-keeping goes further and uses altitude as the only steering authority available, choosing a height at which the wind blows toward the target, so a slow altitude loop becomes the inner loop of a navigation problem solved over days [14]. Both cases have the plant of this chapter, an integrating altitude behind a slow and asymmetric actuator, and both run the loop that the simulator asks the reader to design.

Physical modeling of balloon flight is a settled subject. Ascent and float codes integrate a vertical momentum balance with buoyancy, weight and quadratic drag against a standard atmosphere, and add an envelope energy balance carrying solar input, infrared exchange and the thermal mass of the film [15]. The added mass of the displaced air, which for a sphere in potential flow is half the displaced volume, is part of that balance and dominates the time it takes a climb rate to establish [16]. This application keeps the momentum and energy balances, replaces the radiation terms with a single lumped loss coefficient because a sport flight lasts an hour rather than a day, and adds the burner transport delay that the ascent codes have no reason to carry.

## 2.3 Flight Model

The vehicle is an AX-8 sport balloon of 3000 $\mathrm{m}^3$ (106000 cubic feet) carrying 640 kg of envelope, basket, burner, tanks and four occupants, 77 kg of propane and 40 kg of sand ballast, for an all-up mass of 757 kg. Figure 2.1 shows the arrangement and labels every symbol used below. Nine states are integrated. Three of them carry the lesson, which are the altitude $z$, the climb rate $v_z$ and the envelope temperature $T_{env}$, and the remaining six are the burner delivery position, the vent position, the propane and ballast masses and the horizontal position and speed.

The controlled, manipulated and disturbance variables are set out in Table 2.1. The clock is in seconds, altitudes in meters, temperatures in degrees Celsius except where the gas law needs kelvin, and both actuators in percent of full travel.

### 2.3.1 Vertical Momentum and the Added Mass

Altitude and climb rate follow Newton's second law with buoyancy, weight and quadratic drag, as shown in Equations (2.1) and (2.2).

$$\frac{\mathrm{d}z}{\mathrm{d}t} = v_z \tag{2.1}$$

$$m_{eff}\,\frac{\mathrm{d}v_z}{\mathrm{d}t} = (\rho_a V - m_{tot})\,g - \tfrac{1}{2}\rho_a C_D A\,(v_z - w)\,|v_z - w| \tag{2.2}$$

where $m_{tot} = m_{dry} + m_f + m_b + \rho_{env} V$ is everything the balloon carries including the air inside it, $\rho_a$ is the ambient air density, $V$ is the envelope volume, $C_D$ is the drag coefficient on frontal area $A$, and $w$ is the vertical motion of the air itself. Both sides carry newtons. The mass on the left is not $m_{tot}$, as shown in Equation (2.3).

$$m_{eff} = m_{tot} + k_a\,\rho_a V \tag{2.3}$$

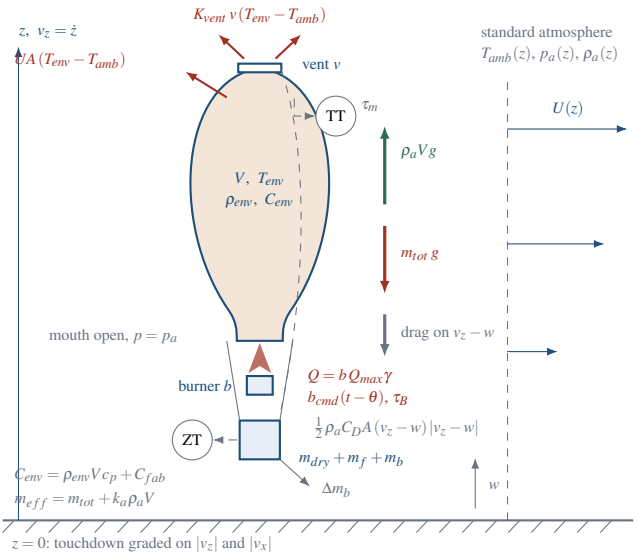



**Figure 2.1** The AX-8 balloon of the simulator with the states, the two actuators, the forces and the heat flows. The envelope of volume $V$ holds air at $T_{env}$ and density $\rho_{env}$ with total thermal capacity $C_{env}$, is open at the mouth and therefore at ambient pressure $p_a$, and loses heat through the fabric as $UA(T_{env}-T_{amb})$ and through the crown vent of opening $v$ as $K_{vent}v(T_{env}-T_{amb})$. Burner fraction $b$ follows the command delayed by $\theta$ with lag $\tau_B$ and delivers $bQ_{max}\gamma$. Net force is buoyancy $\rho_a V g$ less weight $m_{tot}g$ less quadratic drag on the velocity $v_z-w$ relative to the air, accelerating the effective mass $m_{eff}$ that includes the added mass $k_a\rho_a V$. TT is the crown thermocouple with lag $\tau_m$ and ZT is the altimeter and variometer.

**Table 2.1** Classification of Variables. Nominal Values Are Those of the Design Point at 300 m Over a 15 °C Morning.

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $z$ | CV, measured | 300 m | 0–1500 m SP |
| $v_z$ | inner CV, measured | 0 m/s | ±2.5 m/s envelope |
| $T_{env}$ | inner CV, measured | 90.06 °C | $T_{amb}$–120 °C |
| $u$ (demand) | MV | 12.65 % | −100–100 % |
| $b$ (burner) | MV, split range | 12.65 % | 0–100 % |
| $v$ (vent) | MV, split range | 0 % | 0–100 % |
| $T_{sl}$ | DV, measured | 15 °C | −10–38 °C |
| $U_{10}$ | DV, measured | 0 m/s | 0–12 m/s |
| $w$ | DV, measured | 0 m/s | ±3 m/s |
| $\Delta T_{inv}$ | DV, measured | 0 K | 0–12 K |
| $\Delta m_b$ | DV, discrete | — | 20 kg twice |
| $m_f$ | state, measured | 77 kg | 0–77 kg |
| $b_{pos}$, $v_{pos}$ | states | 0.1265, 0 | 0–1 |
| $x$, $v_x$ | states, measured | 0, $U(z)$ | — |

where $k_a$ is the added-mass coefficient, taken as the potential-flow value of 0.5 for a sphere. At the design point $k_a\rho_a V = 1785$ kg against 3570 kg of balloon and envelope air, so a third of what has to be accelerated is the air being pushed out of the way [16]. Linearizing the drag term about a climb rate $|v_z|$ gives a first-order approach with time constant $m_{eff}/(2k_d|v_z|)$ where $k_d = \frac{1}{2}\rho_a C_D A$, which is 32.7 s at 1 m/s. The added mass supplies 10.9 s of that.

### 2.3.2 The Envelope Energy Balance

The envelope is open at the mouth and is therefore at ambient pressure, so its internal energy $pVc_v/R$ does not change and every joule delivered leaves again as the enthalpy of air spilling out of the bottom. Writing that balance out and adding the thermal mass of the fabric gives Equation (2.4), which has the form of a closed vessel holding the air currently inside.

$$\begin{aligned} C_{env}\frac{\mathrm{d}T_{env}}{\mathrm{d}t} &= b\,Q_{max}\,\gamma-(UA+K_{vent}\,v)\,(T_{env}-T_{amb}) \\ C_{env} &= \rho_{env}Vc_p+C_{fab} \end{aligned} \tag{2.4}$$

where $b$ and $v$ are the delivered burner and vent fractions, $Q_{max}$ is the burner rating, $UA$ is the fabric loss coefficient, $K_{vent}$ is the additional loss at a fully open vent, $c_p$ is the specific heat of air and $\gamma$ is the fuel-available factor of Equation (2.5). Both sides carry watts. At the design point the capacity is 2.972 MJ/K, of which the nylon supplies 145 kJ/K, or 4.9 %, and it is the part that keeps heating after the valve shuts. The ratio $C_{env}/UA$ is 646 s, so over any maneuver Equation (2.4) behaves as an integrator. The two halves of the actuator are not gain matched. Full burner against the design-point loss gives $+0.82\,\mathrm{K/s}$ and a fully open vent gives $-1.41\,\mathrm{K/s}$, a ratio of 1.72, and that asymmetry is retained because it is what a parachute vent does and it is what makes one linear controller behave differently going up and coming down.

Propane is consumed in proportion to the delivered heat and the burner dies smoothly as the last half kilogram goes, as shown in Equation (2.5).

$$\frac{\mathrm{d}m_f}{\mathrm{d}t}=-\frac{b\,Q_{max}\,\gamma}{\eta_b\,\mathrm{LHV}},\qquad \gamma=\min\left(\max\left(\frac{m_f}{0.5\,\mathrm{kg}},0\right),1\right) \tag{2.5}$$

where $\eta_b$ is the burner efficiency and LHV is the lower heating value of propane. The smooth gate keeps the four stages of the fixed-step integrator from straddling a discontinuity. At the design point Equation (2.5) runs at 29.9 kg/h, which is 8.3 g/s of load leaving the aircraft continuously and is the reason the float temperature drifts all flight.

### 2.3.3 Actuator Dynamics and the Transport Delay

Both actuators are first order and the burner carries a pure delay in front of its lag, as shown in Equation (2.6).

$$\tau_B\frac{\mathrm{d}b}{\mathrm{d}t}=b_{cmd}(t-\theta)-b,\qquad \tau_V\frac{\mathrm{d}v}{\mathrm{d}t}=v_{cmd}-v \tag{2.6}$$

where $\theta=4\,\mathrm{s}$ is the plume rise and mixing time, $\tau_B=0.8\,\mathrm{s}$ is the delivery lag and $\tau_V=1.0\,\mathrm{s}$ is the vent travel time. The delay is applied to the command through a ring buffer with linear interpolation, written at the controller period and read $\theta$ seconds back, so the right-hand side that the integrator evaluates four times per step stays a pure function of the state. The delay is set in steps of 0.5 s, a whole number of controller periods, so the delivered command is exactly the command issued $\theta$ earlier, held for one controller period. Delaying the command is equivalent to delaying the delivered heat and keeps the delay line out of the derivative evaluation.

### 2.3.4 Split Range and the Descent Envelope

Every architecture manipulates one number, the heat demand $u\in[-1,+1]$, which is split into the two actuators as shown in Equation (2.7).

$$b_{cmd}=\max(u,0),\qquad v_{cmd}=\max(-u,0) \tag{2.7}$$

so the burner and the vent can never be open at once. Using one manipulated variable is what makes the four architectures comparable. The vent also accepts a floor from the override described in Section 2.5. The climb rate the guidance is permitted to ask for is bounded by Equation (2.8).

$$|v_z|\le\min\left(v_{max},\,k_z z+v_{flare}\right) \tag{2.8}$$

where $v_{max}$ is the cruise limit, $k_z$ is the slope of the descent envelope and $v_{flare}$ is what the envelope allows at the ground. Above $z = (v_{max} - v_{flare})/k_z$, which is 85 m at the default settings, Equation (2.8) is a flat cruise limit. Lower down the permitted speed falls with height, which converges the sink rate onto $v_{flare}$ over a time scale of $1/k_z = 50\,\mathrm{s}$. The profile sets the requested terminal rate. Thermal delay and tracking error still affect the physical touchdown speed, so the terminal safety requirement must be checked against the measured arrival and not inferred from the guidance limit alone [17].

### 2.3.5 Atmosphere, Wind and Buoyancy Algebra

The column is a standard lapse rate applied to the day's sea-level temperature, as shown in Equation (2.9).

$$\begin{aligned} T_{amb} &= T_{sl} - Lz + \Delta T_{inv}\, \mathrm{e}^{-((z-z_{inv})/w_{inv})^2} \\ p_a &= p_0 \left(1 - \frac{Lz}{T_{sl}}\right)^{g/RL} \end{aligned} \tag{2.9}$$

with $\rho_a = p_a/(R\,T_{amb})$, where $L$ is the lapse rate, $R$ is the gas constant of air and $\Delta T_{inv}$ is an optional inversion of thickness $w_{inv}$ centered at $z_{inv}$. The inversion is added to the temperature profile only and is not made hydrostatically consistent with the column above it. At 300 m over a 15 °C morning Equation (2.9) gives 13.05 °C, 97773 Pa and 1.19012 kg/m$^3$. Horizontal wind follows a logarithmic profile and vertical air motion is tapered to zero at the surface, as shown in Equation (2.10).

$$U(z) = U_{10}\,\frac{\ln((z+z_0)/z_0)}{\ln(z_{ref}/z_0)}, \qquad w = w_0\left(1 - \mathrm{e}^{-z/z_{th}}\right) \tag{2.10}$$

where $z_0$ is the surface roughness, $z_{ref}$ is the reference height and $z_{th}$ is the height over which a thermal dies out at the ground. The amplitude $w_0$ is the steady vertical air motion plus a thermal or a patch of sink of strength $w_g$ that decays with time constant $\tau_g$ once the balloon has entered it. The horizontal channel is one quadratic drag equation on the same effective mass, $m_{eff}\dot{v}_x = \frac{1}{2}\rho_a C_D A(U(z) - v_x)|U(z) - v_x|$ and $\dot{x} = v_x$, and it feeds nothing back into the vertical channel. It exists because the touchdown verdict depends on ground speed as well as sink rate.

The envelope density is $\rho_{env} = p_a/(R\,T_{env})$ in kelvin, and the temperature that exactly floats a given load follows from neutral buoyancy without iteration, as shown in Equation (2.11).

$$T_{float} = \frac{p_a}{R\,(\rho_a - m_{load}/V)}, \qquad m_{load} = m_{dry} + m_f + m_b \tag{2.11}$$

Equation (2.11) is the reference every architecture in the application is really tracking, and the application draws it as a trace beside the envelope temperature. At the design point it returns 90.06 °C against a 120 °C fabric limit. Differentiating the buoyancy with respect to envelope temperature and equating the resulting force to quadratic drag gives the climb characteristic of Equation (2.12).

$$\frac{\partial F}{\partial T_{env}} = \frac{\rho_{env} V g}{T_{env}}, \qquad v_{z,ss} = \mathrm{sgn}(\Delta T)\sqrt{\frac{\left|\frac{\partial F}{\partial T_{env}}\Delta T\right|}{\frac{1}{2}\rho_a C_D A}} \tag{2.12}$$

where $\Delta T = T_{env} - T_{float}$. At the design point $\partial F/\partial T_{env} = 75.96\,\mathrm{N/K}$ against $k_d = 81.82\,\mathrm{N}$ per $(\mathrm{m/s})^2$, so one kelvin off float buys 0.96 m/s of steady climb, four kelvin buys 1.93 m/s and sixteen buys 3.85 m/s. The square root in Equation (2.12) is the source of most of the difficulty later, because it makes the balloon much easier to start than to stop and it is the term a linear model cannot carry.

### 2.3.6 Constants, Trimmed Start and Integration

Table 2.2 lists the constants. Eight are hidden and are never displayed, and they are the physical constants and the surface-layer constants that no scenario asks the reader to recover. Everything else is visible in the model panel and adjustable.

**Table 2.2** Model Parameters. Hidden Constants Are Never Displayed. Every Visible Constant Is Also Adjustable From the Model Panel.

| Symbol | Meaning | Value | Unit | Visibility |
|---|---|---|---|---|
| $V$ | envelope volume | 3000 | $m^3$ | visible |
| $m_{dry}$ | dry all-up mass | 640 | kg | visible |
| $m_{f0}$ | propane on board | 77 | kg | visible |
| $m_{b0}$ | ballast on board | 40 | kg | visible |
| $C_{fab}$ | fabric heat capacity | 145 | kJ/K | visible |
| $C_D$ | drag coefficient | 0.55 | — | visible |
| $A$ | frontal area | 250 | $m^2$ | visible |
| $k_a$ | added-mass coefficient | 0.50 | — | visible |
| $Q_{max}$ | burner rating | 2.80 | MW | visible |
| $\theta$ | burner transport delay | 4.0 | s | visible |
| $\tau_B$ | heat delivery lag | 0.8 | s | visible |
| $\tau_V$ | vent travel time | 1.0 | s | visible |
| $UA$ | envelope heat loss | 4600 | W/K | visible |
| $K_{vent}$ | vent loss, full open | 50000 | W/K | visible |
| $T_{fab}$ | fabric limit | 120 | °C | visible |
| $\tau_m$ | crown thermocouple lag | 2.0 | s | visible |
| $L$ | lapse rate | 0.0065 | K/m | visible |
| $z_{inv}$ | inversion altitude | 400 | m | visible |
| $w_{inv}$ | inversion thickness | 120 | m | visible |
| $z_{th}$ | thermal cut-off height | 60 | m | visible |
| $w_g$ | thermal strength | 2.0 | m/s | visible |
| $\tau_g$ | thermal decay time | 40 | s | visible |
| $v_{td}$ | stand-up touchdown limit | 1.5 | m/s | visible |
| $v_{hard}$ | hard-landing speed | 3.0 | m/s | visible |
| $u_{drag}$ | drag-out ground speed | 4.0 | m/s | visible |
| $\Delta m_b$ | ballast release size | 20 | kg | visible |
| LHV | propane heating value | 46.4 | MJ/kg | hidden |
| $\eta_b$ | burner efficiency | 0.92 | — | hidden |
| $p_0$ | sea-level pressure | 101325 | Pa | hidden |
| $R$ | gas constant of air | 287.05 | J/(kg K) | hidden |
| $c_p$ | specific heat of air | 1005 | J/(kg K) | hidden |
| $g$ | gravity | 9.80665 | $m/s^2$ | hidden |
| $z_0$ | surface roughness | 0.10 | m | hidden |
| $z_{ref}$ | wind reference height | 10 | m | hidden |

Every scenario opens trimmed. The altitude is set to the target, the envelope temperature to $T_{float}$ of Equation (2.11) clipped to the fabric limit, the climb rate to zero, and the burner position to the opening that matches the fabric loss exactly, as shown in Equation (2.13).

$$b(0) = \frac{UA\left(T_{float} - T_{amb}\right)}{Q_{max}} \tag{2.13}$$

which returns 12.65 % at the design point, or 354 kW of burner against 354 kW of fabric loss. On a day too hot to float the load the state opens at the fabric limit and the balloon sits on the ground. The engine of Section 1.6.1 integrates with classical fourth-order Runge-Kutta on a fixed 0.05 s step, the controllers execute every 0.5 s, the predictive controller executes on its own 2 s subclock and the historian logs five samples per simulated second. The model draws no random numbers, so a scenario replays identically. Touchdown is detected when $z$ reaches zero, the rigid body is then frozen, and the arrival is graded as a stand-up landing below 1.5 m/s, firm between 1.5 and 3.0 m/s, hard above that, and dragged when the ground speed of an arrival that is not hard exceeds 4 m/s.

### 2.3.7 Assumptions

Every simplification is given below beside the condition that defeats it.

*The envelope is one lumped air mass at one temperature.* A real envelope is stratified by twenty kelvin or more between crown and mouth, and it is the crown that reaches the fabric limit first, so the overtemperature warning here is optimistic about where the damage happens and approximately right about when.

*The transport delay is a fixed 4 s.* In a real balloon it depends on the burn rate, on the envelope size and on how much cold air the plume entrains, and it is shorter at full burn than at a trickle. Four seconds is a plausible figure for 3000 $m^3$, it is the number the whole application turns on, and it is the one an instrumented balloon should be used to measure.

*Buoyancy is computed at a constant envelope volume.* A real envelope breathes through its open mouth, but once inflated the shape is set by the load tapes and the volume error is under a percent.

*Vertical drag is one quadratic term with textbook sphere constants.* Neither $C_D = 0.55$ on 250 $m^2$ nor $k_a = 0.5$ has been measured on a balloon here, and $k_a$ in particular sets how long a climb rate takes to establish.

*Radiation is absorbed into UA.* Solar gain on the envelope, infrared exchange with the sky and ground, and the diurnal cycle that dominates a long-duration flight are all absent [15], which is defensible for a one-hour sport flight and not for a stratospheric one.

*The inversion is a temperature anomaly only.* Making it hydrostatically consistent with the column above would change the density inside the layer by well under half a percent at 8 K and 400 m, and would not change the sign of anything.

*Horizontal motion carries no direction, no turbulence and no navigation.* The drag-out verdict at touchdown is therefore about speed alone, and the real steering mechanism of a balloon, which is choosing an altitude whose wind points somewhere useful, is absent.

*The numerics repeat exactly.* A fixed step and a fixed controller period reproduce one trajectory on every machine, which is convenient for grading and is not what a real flight computer does.

## 2.4 The Chain From Blast Valve to Altitude

Scenario 2 is a clean open-loop step of 20 % on the blast valve from the trimmed condition, held. Table 2.3 reports what the application measures, with crossing times read at the 0.05 s physics step. The delivered heat begins to move 4.0 s after the step, which is the transport delay alone, and then rises through the 0.8 s delivery lag. The envelope then heats at up to 0.187 K/s, close to the 0.188 K/s of 560 kW into 2.972 MJ/K, and does not slow measurably over two minutes because the 646 s loss constant has barely begun to act. The variometer passes 0.1 m/s at 13.0 s, 0.5 m/s at 24.2 s and 1.0 m/s at 33.2 s, and at 120 s the envelope is 20.5 K hotter, the balloon is 264.0 m higher, and it is still accelerating. The unstepped balloon climbs as well, by 6.5 m over the same 120 s, because it is burning 8.3 g/s of its own load, so 257.5 m of the gain belongs to the step.

> **TRY IT IN THE APP** **2 · The Step Test the MPC Is Built From**
>
> Load the scenario and press Run. Watch the altitude. The blast valve steps up at 20 s, and for several seconds nothing moves. Two minutes after the step the balloon is 264.0 m higher and still accelerating, with no level at which a process gain could be read.

There is no steady state in Table 2.3 from which to read a process gain. Strictly the path from demand to altitude carries one pole at the origin, because altitude is the integral of climb rate, together with a delay of 4 s and three real poles at $1/\tau_B = 1.25/\text{s}$, at $UA/C_{env} = 1.55 \times 10^{-3}/\text{s}$ and at $2k_d|v_z|/m_{eff} = 3.1 \times 10^{-2}/\text{s}$ for a climb of 1 m/s. The thermal pole is so much slower than any maneuver that the plant behaves as a double integrator over the horizons that matter, which is what Table 2.3 shows. Evaluating that transfer function at $\omega = 0.01/\text{s}$, a period of about ten minutes, gives a phase of $-192°$, made up of $-90°$ from the integrator, $-81°$ from the thermal pole, $-18°$ from the

**Table 2.3** Open-Loop Response to a 20 % Blast-Valve Step From the Trimmed Design Point, Measured on the Shipped Model. Times Are After the Step. Altitudes in Parentheses Subtract the Unstepped Run.

| Quantity | Value |
|---|---|
| Delivered heat leaves its pre-step value | 4.0 s |
| Initial envelope heating rate | 0.187 K/s |
| Variometer reaches 0.1 m/s | 13.0 s |
| Variometer reaches 0.5 m/s | 24.2 s |
| Variometer reaches 1.0 m/s | 33.2 s |
| Variometer reaches 2.0 m/s | 49.8 s |
| Altitude gained at 30 s | 7.2 m (6.9) |
| Altitude gained at 60 s | 58.2 m (57.0) |
| Altitude gained at 120 s | 264.0 m (257.5) |
| Envelope temperature rise at 120 s | 20.5 K |

drag lag and only $-2.3^\circ$ from the delay. The plant on its own is already past $-180^\circ$ there, and at a period of five minutes it is at $-214^\circ$. A single proportional-integral-derivative loop from altitude error to burner position has to buy all of that phase back itself, and the derivative term is the only one that can see a climb rate while looking through the delay. Removing the derivative adds a second pole at the origin from the reset term, which takes the open loop to $-282^\circ$ at the same frequency before any lead is credited back, and the measured consequence is in Section 2.6.

The asymmetry compounds this. Descending is easy, because the vent removes heat at $-1.41\,\mathrm{K/s}$ and every kelvin removed is bought back at $+0.82\,\mathrm{K/s}$ through four seconds of dead time. The square root in Equation (2.12) then means that arresting a 3 m/s descent requires recovering 9.7 K while arresting a 1 m/s descent requires only 1.1 K. A controller that holds altitude well in cruise can still arrive at the ground much too fast, and that is what the challenge scenario grades.

## 2.5 Controllers

Four architectures manipulate the same demand $u$ of Equation (2.7). The demand is slew limited at 50 %/s and clipped to $[-1, +1]$ after every architecture, and once the balloon is on the ground the demand is forced to $-1$ with the vent fully open, which is what a crew does and which keeps the logged actuators inside their declared limits.

### 2.5.1 Manual and the Single Loop

Manual mode exposes the blast valve and the vent as sliders, with the B key held for full burner, which is how the aircraft is actually flown. Every PID loop of the application is the velocity-form algorithm of Section 1.6.2, Equations (1.14) to (1.17), executed every $\Delta t_c = 0.5\,\mathrm{s}$ with the derivative taken on the measurement and back calculation against the loop's own output limits with tracking time $T_t = \tau_I$. On the single loop the measurement passes through a first-order filter of 3 s before the derivative sees it. The demand slew limit is applied after every architecture and is not fed back to any loop, so each back calculation acts only against its own clip. In manual the integral is set so that the output equals the operator's slider, and every loop not currently driving the actuator is tracked in the same way, so transfer is bumpless in both directions and changing the architecture in flight moves nothing.

The single loop drives altitude error straight onto the blast valve with $K_c = 0.06\,\%/\mathrm{m}$, $\tau_I = 1200\,\mathrm{s}$ and $\tau_D = 130\,\mathrm{s}$, which is as well as one loop can be tuned on this plant, and the vent is handled by the high-altitude override of Equation (2.14).

$$v_{ovr} = \mathrm{clip}\left( \frac{z - z_{\mathrm{SP}} - z_{db}}{z_{span}},\ 0,\ 1 \right)$$
$$u = -v_{ovr} \quad \text{whenever } v_{ovr} > 0 \tag{2.14}$$

with a deadband $z_{db} = 40\,\mathrm{m}$ and a span $z_{span} = 60\,\mathrm{m}$. The override takes the burner away from the loop entirely while it acts, and the loop is not informed, so its back calculation still sees only its own clip to $[0, 1]$. That is the behavior that puts the balloon in a field on the graded landing of Section 2.6, where the override acts on 146 of the loop's 215 executions before touchdown.

### 2.5.2 Cascade

The cascade breaks the plant at each slow element. The altitude loop produces a climb-rate setpoint, the climb-rate loop produces an envelope-temperature setpoint, and the temperature loop drives the demand, as shown in Equation (2.15).

$$\begin{aligned} v_{z,\mathrm{SP}} &= \mathrm{clip}(\mathrm{PI}_z(z_{\mathrm{SP}} - z),\ -v_{lim},\ +v_{lim}) \\ T_{\mathrm{SP}} &= \mathrm{clip}\big(T_{float} + \mathrm{PI}_v(v_{z,\mathrm{SP}} - v_z),\ T_{amb} + 2,\ T_{fab} - 1\big) \\ u &= \mathrm{clip}\big(b_{trim} + \mathrm{PID}_T(T_{\mathrm{SP}} - \hat{T}_{env}),\ -1,\ +1\big) \end{aligned} \tag{2.15}$$

where $v_{lim}$ is the descent envelope of Equation (2.8), $\hat{T}_{env}$ is the crown thermocouple through its 2 s filter, and the two bias terms are feedforwards. The middle loop is biased by the float temperature of Equation (2.11), so it begins from the answer the ideal-gas law already gives and corrects only what the model misses. The inner loop is biased by the trim burner of Equation (2.13) evaluated on the measured temperature. Default gains are $K_{c,z} = 1.6\,(\mathrm{m/s})$ per 100 m with $\tau_{I,z} = 1200\,\mathrm{s}$, $K_{c,v} = 3.0\,\mathrm{K}$ per m/s with $\tau_{I,v} = 90\,\mathrm{s}$, and $K_{c,T} = 12\,\%/\mathrm{K}$ with $\tau_{I,T} = 60\,\mathrm{s}$ and no derivative.

Three choices in Equation (2.15) deserve their reasons. The general rule, that a loop rejects the disturbances entering inside its own measurement path, is the subject of Volume 3, Chapter 2. The inner loop is fast because it sees a single 646 s lag with a steady-state gain of 609 K per unit of demand through only a 4 s delay. The middle loop is deliberately slow, because a thermal lifts the balloon without changing its buoyancy by a single newton and a climb-rate loop faster than the disturbance will spend real heat fighting air that was going to stop pushing on its own. The outer loop is almost purely proportional, because an integrating process is offset-free under proportional control alone and reset action there would add a fourth pole at the origin for nothing. For a nonzero altitude target the output is clipped to the climb-rate envelope. When the target is the field ($z_{\mathrm{SP}} = 0$), the cascade instead requests $-\min(v_{max}, k_z z + v_{flare})$ directly and tracks the inactive altitude PI to that request. This keeps an asymptotic altitude command from bypassing the flare setting near the ground.

### 2.5.3 Step-Response Predictive Control

The predictive controller is dynamic matrix control in the original step-response form [6, 7], arranged so that everything expensive happens once. Figure 2.2 separates the offline path from the online path.

#### *Identification*

When the application loads, the controller integrates a reduced five-state copy of the same model, with the declared parameters of Table 2.2, from the trimmed design point of Table 2.1, twice over the horizon, once with the demand held at its trim value and once with a $+0.08$ step, and differences the two altitude trajectories, as shown in Equation (2.16).

$$a_i = \frac{z_i^{step} - z_i^{base}}{\Delta u}, \qquad a_0 = 0 \tag{2.16}$$

The reduced model carries altitude, climb rate, envelope temperature and the two actuator positions, plus the burner delay as a ring of past commands prefilled with the pre-step command, and it is integrated with explicit Euler at 0.2 s. Prefilling the ring with the stepped command instead would remove the dead time from the identified response entirely. The load and the wind are frozen, because a 52 s run burns about 430 g of propane, or 0.06 % of the all-up mass. The result at the design point is $a_1 = a_2 = 0$, which is the four seconds of dead time at a 2 s sample, $a_3 = 0.004\,\mathrm{m}$,

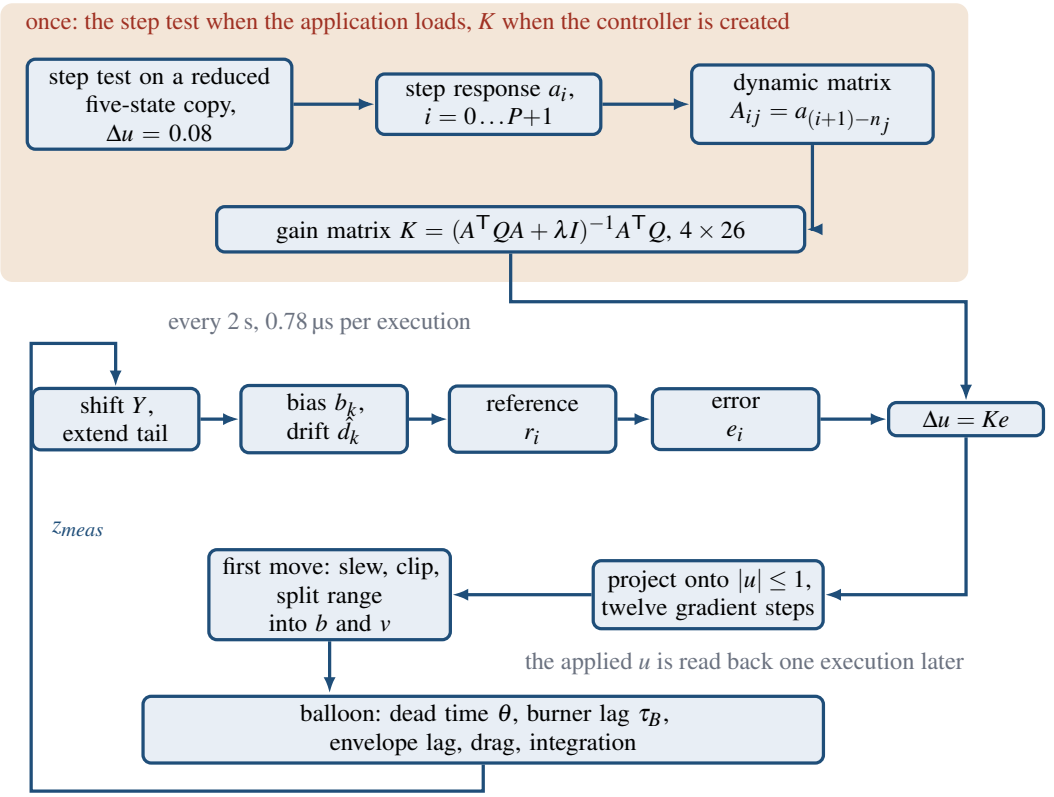



**Figure 2.2** Structure of the step-response predictive controller. The shaded path runs once. The step test is taken on the declared design when the application loads, and the constant gain matrix $K$ is produced when the controller is created. The unshaded path runs every 2 s and is one shift, one add, one $4 \times 26$ matrix-vector product and, on the small fraction of executions that violate the input box, twelve projected-gradient iterations. The prediction vector $Y$ is advanced only by the demand the actuator actually received, which is where the anti-windup lives.

$a_{12} = 15.14$ m at 24 s and $a_{26} = 210.3$ m at 52 s. The increments still grow, $a_{26} - a_{25} = 23.9$ m against $a_{13} - a_{12} = 5.2$ m, which is what an integrating output looks like and what a self-regulating one never does. The identification is not repeated for a run, so a parameter changed from its default, another ground temperature or another starting altitude is model error that the bias of Equation (2.21) must absorb.

### *Dynamic Matrix and Gain Matrix*

The prediction horizon is $P = 26$ points at 2 s, which is 52 s, and the move horizon is $M = 4$ after blocking at samples $n_j \in \{0, 1, 4, 9\}$, which is 0, 2, 8 and 18 s. Move blocking covers a horizon with a few decisions instead of one per sample and is standard practice [6]. The dynamic matrix and the gain matrix follow, as shown in Equation (2.17), which is Equations (1.3) and (1.5) written with $A$ for $G$, $P$ for $N$ and $\lambda$ for $R$.

$$A_{ij} = \begin{cases} a_{(i+1)-n_j}, & i+1 > n_j \\ 0, & \text{otherwise} \end{cases} \qquad K = \left(A^{\mathsf{T}} Q A + \lambda I\right)^{-1} A^{\mathsf{T}} Q \tag{2.17}$$

where $A$ is $26 \times 4$, $Q$ is diagonal with unit weight on the first eighteen points and terminal weight $w_T$ on the last eight, and $K$ is $4 \times 26$. The inverse is a $4 \times 4$ Gauss-Jordan elimination written out in the source. A unit step of demand is worth 210 m fifty-two seconds later, so the mean of the diagonal of $A^{\mathsf{T}}QA$ is $5.35 \times 10^5$ and a move suppression expressed in those units would need a slider spanning several decades. The application therefore exposes $\lambda$ as a dimensionless multiple of the mean diagonal, as shown in Equation (2.18).

$$\lambda = \lambda_{rel}\,\frac{1}{M}\,\mathrm{tr}\!\left(A^{\mathsf{T}}QA\right) \tag{2.18}$$

so $\lambda_{rel} = 1$ always means move suppression comparable to the tracking term. The defaults are $\lambda_{rel} = 0.02$, which is $\lambda = 1.07 \times 10^4$, and $w_T = 6$. Equation (2.17) is rebuilt only when $\lambda_{rel}$ or $w_T$ moves, and the step response is never rebuilt at all.

### *The Online Execution*

Each execution begins by accounting for what the actuator actually received, as shown in Equation (2.19).

$$Y_i \leftarrow Y_i + a_i\,(u_{now} - u_{prev}), \qquad i = 0 \ldots P \tag{2.19}$$

where $u_{prev}$ is the demand at the previous execution. Nothing is ever added to the prediction speculatively. A move that the clip halved, that the slew limiter is still ramping toward, or that never happened because a different architecture was flying, enters at its true size, and Equation (2.19) is therefore the whole of this controller's anti-windup. It is also what keeps the predictive controller warm while another architecture flies, so selecting it in mid-flight is bumpless. The vector is then shifted one sample and the tail extended, as shown in Equation (2.20).

$$Y_i \leftarrow Y_{i+1} \quad (i < P), \qquad Y_P \leftarrow Y_P + d_1 + d_2 \tag{2.20}$$

with $d_1 = Y_P - Y_{P-1}$ and $d_2 = d_1 - (Y_{P-1} - Y_{P-2})$, which preserves both the slope and the curvature of the last two intervals and is exact for a double integrator and close for this plant. Extending the tail at a constant value would be correct for a self-regulating process and extending it at a constant slope would be correct for a single integrator, and neither describes this output over a 52 s horizon.

The correction to the plant follows, as shown in Equation (2.21).

$$\begin{aligned} b_k &= z_{meas} - Y_0, \qquad Y_i \leftarrow Y_i + b_k \\ \hat{d}_k &= \hat{d}_{k-1} + \alpha_d \left( \frac{b_k}{t_m} - \hat{d}_{k-1} \right) \end{aligned} \tag{2.21}$$

where $t_m = 2$ s is the execution period and $\alpha_d = 0.33$ is a filter constant of about three executions. The constant $b_k$ is the standard output-disturbance model that makes dynamic matrix control offset-free and is the predictive counterpart of integral action [18], as Section 1.4.5 shows on a self-regulating tank. On an integrating output a constant is not enough, which is the subject of Section 2.6.4. Because $Y$ was re-anchored to the plant exactly one execution earlier and nothing else has touched it, $b_k/t_m$ is a direct measurement of the unmodeled drift rate, clipped to $\pm 6$ m/s. The rate estimate is not written back into $Y$, so the estimator cannot chase its own output.

The reference trajectory carries the terminal constraint, as shown in Equation (2.22).

$$\begin{aligned} r_0 &= z_{meas} \\ \dot{r}_i &= \mathrm{clip}\!\left( \frac{z_{SP} - r_i}{\tau_{ref}},\; -v_{lim}(r_i),\; +v_{lim}(r_i) \right) \\ r_{i+1} &= \max(r_i + t_m\,\dot{r}_i,\; 0) \end{aligned} \tag{2.22}$$

**Table 2.4** Scenario Sequence and the Mechanism Each Exercises

| # | Scenario | What it shows |
|---|---|---|
| 1 | Hand flying from trim | the delay, discovered rather than told |
| 2 | Blast-valve step test | integrating response, no gain to read |
| 3 | Single loop on altitude | why one gain cannot close it |
| 4 | Same level-off, cascade | one slow element per loop |
| 5 | Level flight, fuel burning | a load that never stops moving |
| 6 | Up 300 m and back | the asymmetric actuator |
| 7 | Thermal, then inversion | measurement against load disturbance |
| 8 | Dawn against afternoon | density altitude as a hard constraint |
| 9 | Ballast against sink | a fast actuator that is a finite resource |
| 10 | Predictive control, plan drawn | the horizon, the bias, the plan |
| 11 | Graded landing | a terminal constraint on arrival speed |

where $\tau_{ref} = 60\,\mathrm{s}$ and $v_{lim}$ is Equation (2.8) evaluated at the reference altitude. For a landing ($z_{\mathrm{SP}} = 0$), the clipped proportional request in Equation (2.22) is replaced by $\dot{r}_i = -v_{lim}(r_i)$, with the zero floor on $r_{i+1}$ retained. The flare setting then shapes the requested descent, while the actual velocity can still exceed the guidance profile because of tracking error. The tracking error is taken against the prediction with the drift ramp projected forward, as shown in Equation (2.23).

$$\begin{aligned} e_i &= r_{i+1} - \left[ Y_{i+1} + \hat{d}_k \min((i+1)t_m,\ c_k) \right] \\ c_k &= P t_m \left( 1 - \mathrm{e}^{-z_{meas}/z_{th}} \right) \end{aligned} \tag{2.23}$$

where the cap $c_k$ tapers the projection to nothing at the surface using the same profile Equation (2.10) uses for vertical air motion. A thermal or a patch of sink dies at the ground, so extrapolating one all the way to touchdown would make the controller arrive hot and balloon back up.

The unconstrained solution is one matrix-vector product, $\Delta u = Ke$. If the cumulative sum of that plan would leave $[-1, +1]$ or exceed the per-execution rate limit, twelve projected-gradient iterations run on the same precomputed Hessian $H = A^{\mathsf{T}} Q A + \lambda I$, as shown in Equation (2.24).

$$\Delta u \leftarrow \Pi\left[ \Delta u - \alpha \left( H \Delta u - A^{\mathsf{T}} Q e \right) \right], \qquad \alpha = \frac{1}{\max_j H_{jj}} \tag{2.24}$$

where $\Pi$ clips each blocked move to the rate limit and then clips the running sum of the demand to the box, in order. Twelve iterations is a constant rather than a convergence test, which bounds the worst-case execution time. The first blocked move is then applied and nothing is added to the prediction, because the next execution reads back what the actuator really did through Equation (2.19). Over five independent runs of 100000 executions the online path costs 0.78 µs per execution on an Apple M-series laptop, with the individual run means between 0.757 and 0.811 µs. On the graded landing the projection was not needed on any of the 102 executions before touchdown.

## 2.6 Results

Table 2.4 gives the eleven scenarios in the order a flight is taught. Each one fixes the architecture, the tuning, the atmosphere and any timed events.

### 2.6.1 One Loop, Three Loops and a Predictive Loop

Scenarios 3 and 4 apply the same 150 m level-off, a setpoint move from 300 to 450 m at $t = 60\,\mathrm{s}$, to the same balloon through different architectures. Table 2.5 adds the predictive controller on the identical maneuver, which is scenario 4 with the architecture switched to predictive control before the run, and reports every figure over the same 2400 s run, with times measured from the setpoint change. Each retuned case below changes one setting of scenario 3 before the run.

**Table 2.5** The Same 150 m Level-Off at $t = 60$ s Through Three Architectures, Read Over One 2400 s Run

| | Single PID | Cascade | Predictive |
|---|---|---|---|
| Overshoot (m) | 58.9 | 15.2 | 5.4 |
| Last exit from ±15 m (s) | 2340 | 221 | 220 |
| Peak-to-peak, 300–900 s (m) | 125.5 | 9.8 | 14.0 |
| Peak-to-peak, 1800–2400 s (m) | 201.1 | 0.8 | 6.8 |
| Error at 2400 s (m) | −142.2 | 1.1 | 2.8 |
| Demand travel (% of span) | 1087 | 296 | 457 |

The single loop never settles and gets worse. Its swing grows from 125.5 m peak to peak over the first fifteen minutes to 201.1 m over the last ten, and the balloon is 142.2 m below its setpoint when the run ends. Neither direction of retuning helps. Raising the gain from 0.06 to 0.25 %/m widens the 300 to 900 s swing from 125.5 to 357.3 m and the 1800 to 2400 s swing from 201.1 to 372.6 m, and it still does not settle, and removing the derivative term ends the run on the ground at 4.06 m/s after 344 s, which is the proportional-integral case of Section 2.4. Reducing the gain to 0.02 %/m with the derivative removed lands the balloon as well, at 3.72 m/s after 414 s. The failure is structural rather than numerical, and scenario 3 lets the reader produce it before scenario 4 shows the alternative.

Both structured controllers settle. The cascade is the better of the two on this maneuver, with 0.8 m of residual swing against 6.8 m and 296 % of demand travel against 457 %, and the honest reading is that prediction earns nothing in the cruise. Its 15.2 m overshoot against the predictive controller's 5.4 m is the price of having to see an error before acting on it, and on a level-off that price is small.

### 2.6.2 The Landing

Scenario 11 is where the architectures separate. The balloon starts at 400 m on a 24 °C afternoon with 2.5 m/s of wind at 10 m and 0.4 m/s of standing sink, is cleared to land at $t = 30$ s, and meets a further patch of sink at $t = 150$ s. Table 2.6 reports the arrivals with the suggested configuration applied and the architecture then chosen before the run. The case as loaded, with its slack single loop at $K_c = 0.02\,\%/\text{m}$ and no derivative, arrives in the same way as the first column, at 7.85 m/s after 108 s, more than five times the sink limit, and at 4.16 m/s across the ground, so it misses the ground-speed clause as well. The suggested configuration is predictive control with the cascade tuning and $v_{flare} = 0.70\,\text{m/s}$. It lands standing up at 1.39 m/s and 3.53 m/s after 202 s, with 0.11 m/s of sink and 0.47 m/s of ground speed to spare and a 104.9 °C peak, 15.1 K under the fabric limit. The challenge is judged on the case as given, so changing a model parameter or the air voids the attempt.

> **TRY IT IN THE APP** **Challenge · Land It Under 1.5 m/s**
>
> Load the challenge and press Run as it is loaded. The slack single loop arrives at 7.85 m/s after 108 s, a hard landing. Press Reset, apply the suggested configuration, which selects predictive control, and run again: the balloon lands standing up at 1.39 m/s after 202 s.

The single loop arrives at 7.85 m/s after 108 s, which is a wreck, and it does so for a reason no gain addresses. Its vent override of Equation (2.14) opens on the way down, takes the burner away from the loop, and lets the envelope fall toward ambient. There is nowhere in a single loop to put a constraint on arrival speed. The cascade arrives at 1.27 m/s and the predictive controller at 1.39 m/s, and both meet the challenge limits. During the landing the cascade requests the descent profile directly and the predictive controller builds its reference from it. These are guidance limits, not hard constraints on the true velocity. With $v_{flare} = 0.70\,\text{m/s}$, both arrivals are faster than the requested terminal rate because of tracking error. The predictive controller arrives 26.75 s sooner, while the cascade keeps more sink-rate margin, and both burn about 2.44 kg of propane. The cascade spends 48.1 s below 100 m, against 44.2 s for the predictive controller, and arrives with slightly less ground speed. A

**Table 2.6** Graded Landing From 400 m on a 24 °C Afternoon. A Stand-Up Landing Requires Sink Below 1.5 m/s and Ground Speed Below 4 m/s.

| | Single PID | Cascade | Predictive |
|---|---|---|---|
| Touchdown sink (m/s) | 7.85 | 1.27 | 1.39 |
| Ground speed (m/s) | 4.16 | 3.49 | 3.53 |
| Time to contact (s) | 108 | 229 | 202 |
| Propane burned (kg) | 0.32 | 2.44 | 2.44 |
| Peak envelope (°C) | 104.8 | 104.9 | 104.9 |
| Verdict | hard | stand-up | stand-up |

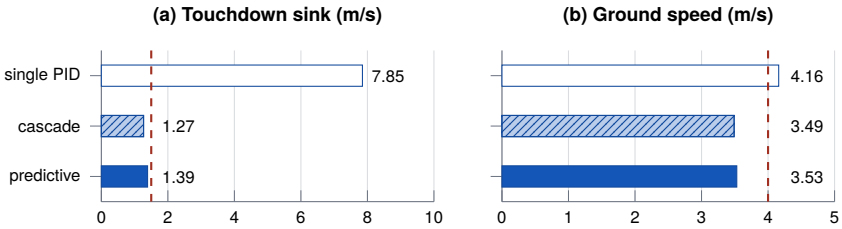



**Figure 2.3** The arrivals of Table 2.6: (a) the touchdown sink rate and (b) the ground speed, white bars for the single PID loop, hatched bars for the cascade and dark bars for the predictive controller, with the stand-up limits of 1.5 m/s and 4 m/s as dashed lines and the values of panel (b) printed beyond its limit. Notice that the single loop fails both limits, while the predictive controller passes the sink limit with only 0.11 m/s to spare.

**Table 2.7** Altitude Gain in m per Unit of Demand at Four Horizons and Four Step Sizes. A Linear Plant Would Give Identical Rows.

| $\Delta u$ | 26 s | 52 s | 90 s | 120 s |
|---|---|---|---|---|
| 0.02 | 20.5 | 222.3 | 1182.8 | 2557.9 |
| 0.08 | 20.4 | 210.3 | 950.5 | 1808.4 |
| 0.20 | 20.1 | 191.4 | 735.4 | 1300.8 |
| 0.50 | 19.6 | 160.7 | 525.6 | 884.3 |

flare sweep from 0.3 to 2.5 m/s changes the cascade touchdown from 0.92 to 2.08 and the predictive touchdown from 1.19 to 1.58 m/s. The measured verdict, and not the requested profile alone, decides whether a landing passes. Figure 2.3 draws the three arrivals against the stand-up limits.

**Table 2.8** Settled Altitude Hold at 300 m, Peak to Peak and Final Error Over 400 to 2000 s, for Two Prediction Horizons Across a Sweep of Move Suppression

| | $P = 26$ (52 s) | | $P = 60$ (120 s) | |
|---|---|---|---|---|
| $\lambda_{rel}$ | p-p (m) | error (m) | p-p (m) | error (m) |
| 0.004 | 1.4 | 2.4 | 170.2 | −42.4 |
| 0.02 | 2.9 | 3.9 | 214.4 | −99.3 |
| 0.1 | 17.0 | 12.0 | 229.9 | 47.3 |
| 0.5 | 24.2 | −5.3 | 234.4 | −107.5 |
| 2 | 28.7 | −9.7 | 252.7 | −82.5 |
| 5 | 35.2 | −12.0 | 294.3 | 118.0 |

**✦ DID YOU KNOW? The Hard Landing Burned the Least Fuel**

From the simulator: on the graded landing the single loop burns 0.32 kg of propane, against 2.44 kg for the cascade and for the predictive controller, and it is the only one of the three graded hard (Table 2.6). The heat it never bought is the heat that would have slowed the descent.

### 2.6.3 Choosing the Horizon by Where the Model Stops Being True

The step response of Equation (2.16) is identified at $\Delta u = 0.08$ and is then used as though the plant were linear. It is not. Quadratic drag makes the true climb rate saturate as $\sqrt{\Delta T}$ while the identified $a_i$ keeps growing, and rescaling the reduced model to other step sizes measures the disagreement directly. Table 2.7 normalizes the altitude gain by the step size, so a linear plant would return identical columns.

At 26 s the four rows of Table 2.7 agree to 4.3 %. At 52 s the model identified at $\Delta u = 0.08$ over-predicts a $\Delta u = 0.20$ move by 9.9 %, which is 3.8 m on a 38.3 m response. At 90 s the same comparison is 29 % and 43.0 m, and at 120 s it is 39 % and 101.5 m. The reason is visible in the climb rate. A $\Delta u = 0.20$ move has the balloon at 2.09 m/s at 52 s and 3.45 m/s at 90 s, and by then the square root in Equation (2.12) has taken hold while the linear model is still accelerating like $t^3$.

Weighting prediction points the model cannot predict is expensive. The general trade between horizon, move suppression and model error is set out in Section 1.4.4. The balloon adds a case the tank there does not show, a horizon too long for the model. Table 2.8 holds the balloon at 300 m under the predictive controller with no setpoint change and no disturbance, which is scenario 5 with the architecture switched to predictive control before the run, and reports the peak-to-peak altitude over 400 to 2000 s and the error $z - z_{SP}$ at 2000 s for two horizons and a sweep of move suppression. The application fixes $P = 26$, so the 120 s column is a counterfactual run of a copy of the source with only the horizon constant changed.

The 52 s horizon holds the balloon inside 2.9 m at the default weight, and its worst case anywhere in the sweep is 35.2 m. The 120 s horizon puts the loop into a 214.4 m limit cycle, and no value of $\lambda_{rel}$ over three decades removes it, because raising move suppression from 0.004 to 5 moves the swing from 170.2 to 294.3 m and never below. A limit cycle that does not respond to move suppression is the signature of a model error rather than an aggressive optimizer, and the remedy is to stop weighting the part of the horizon the model cannot predict. The cycle is not an artifact of the terminal weight sitting on the least trustworthy points either. Setting $w_T = 1$ at $P = 60$, which makes $Q$ the identity, leaves the swing between 142.8 and 293.7 m over the same move-suppression sweep, while at $P = 26$ raising $w_T$ from 1 to 40 moves the settled swing only from 1.6 to 4.4 m. Shortening the horizon to 52 s shortens it well below the settling time of the maneuvers the controller performs, which contradicts the usual advice, and it is correct here. The rule that survives is to choose the prediction horizon by where the identified model stops being true rather than by the settling time, and no tuning table contains it.

**Table 2.9** Effect of the Two Disturbance-Model Terms. Hold Is at 300 m With No Applied Disturbance, Read Over 400 to 2000 s. Landing Is Scenario 11 With the Suggested Configuration.

| Disturbance model | Hold p-p (m) | Hold error (m) | Touchdown (m/s) |
|---|---|---|---|
| Bias and drift ramp | 2.9 | 3.9 | 1.39 |
| Bias only | 117.0 | 132.9 | 2.18 |
| Neither | 996.7 | 605.6 | 2.51 |

🕮 **Pause.** Try Exercise 2.3 before reading on.

### 2.6.4 A Constant Bias Is Not Offset-Free on an Integrating Output

The output-disturbance model of Equation (2.21) is the standard route to offset-free predictive control, and the conditions under which it works are known [18, 19]. Adding a constant to every point of the horizon is correct when a sustained unmodeled load appears as a fixed output offset. On an integrating output it appears as a fixed offset *per sample*. The balloon supplies a permanent example without any disturbance being applied at all, because Equation (2.5) removes 8.3 g/s of load and the float temperature of Equation (2.11) therefore falls continuously.

Table 2.9 measures all three cases on copies of the shipped controller with one constant or one line changed, counterfactual runs that the application itself does not offer. With the bias alone the controller holds a 132.9 m steady offset and swings 117.0 m, because a constant added to every point of a horizon whose far end is already wrong by the drift rate multiplied by the horizon leaves the far end wrong by exactly that amount, the mechanism of Equation (1.6). Adding the two lines of Equation (2.21) that estimate the drift rate and the one term of Equation (2.23) that projects it forward takes the offset to 3.9 m and the swing to 2.9 m. The same three cases land the balloon at 1.39, 2.18 and 2.51 m/s against a 1.5 m/s stand-up limit, so only the full disturbance model passes. Removing the bias entirely leaves the controller dead reckoning off a step test taken on the declared design, and Table 2.9 prices that at 605.6 m of offset.

🕮 **Pause.** Try Exercise 2.2 before reading on.

### 2.6.5 Disturbances, Density Altitude and Ballast

Scenario 7 puts two disturbances through the same instrument. A thermal lifts the air and the balloon with it, so the variometer reads a climb while the buoyancy has not changed by a single newton, and the climb-rate loop drives the envelope 2.16 K below float chasing air that stops pushing on its own. An inversion is the opposite case. Eight kelvin of warm air at 400 m is air 2.7 % less dense, the float temperature at that height moves from 90.18 to 103.24 °C, and the controller is right to burn. The same reading on the same instrument means do nothing in one case and burn in the other, and the only defense the cascade has is a middle loop slower than the disturbance it should ignore.

Density altitude is a constraint no controller can relieve, and Table 2.10 gives its size. The same balloon and the same 757 kg floats at 300 m on 69.5 °C at dawn and needs 119.3 °C on a 33 °C afternoon, which is seven tenths of a degree from the fabric limit, and the ceiling collapses from 6743 m to 618 m. Every degree of ground temperature costs roughly a degree and a half of envelope temperature and two hundred meters of ceiling. A controller sitting against a limit it is not allowed to cross is behaving correctly, and the only handles are load and time of day.

Scenario 9 puts the envelope against that limit deliberately. On a 28 °C afternoon with 3 m/s of sink the cascade drives the envelope to 119.5 °C and the balloon is still sinking. One 20 kg sandbag is 196 N, which is 2.58 K of envelope temperature that is not available and 1.55 m/s of steady climb, and it is delivered with no dead time, no lag and no fuel. Releasing it turns the sink into a 0.19 m/s climb, and at 460 s it is at 262.3 m against 174.6 m for a counterfactual run with the scripted release removed, 88 m apart. There are exactly two sandbags, so the question stops being how to tune a loop

**Table 2.10** Density Altitude. Float Temperature at 300 m and Ceiling at the 120 °C Fabric Limit, for 757 kg All Up.

| Ground temperature | Float temperature | Ceiling |
|---|---|---|
| 2 °C | 69.5 °C | 6743 m |
| 15 °C | 90.1 °C | 5325 m |
| 24 °C | 104.6 °C | 3779 m |
| 28 °C | 111.1 °C | 2764 m |
| 33 °C | 119.3 °C | 618 m |

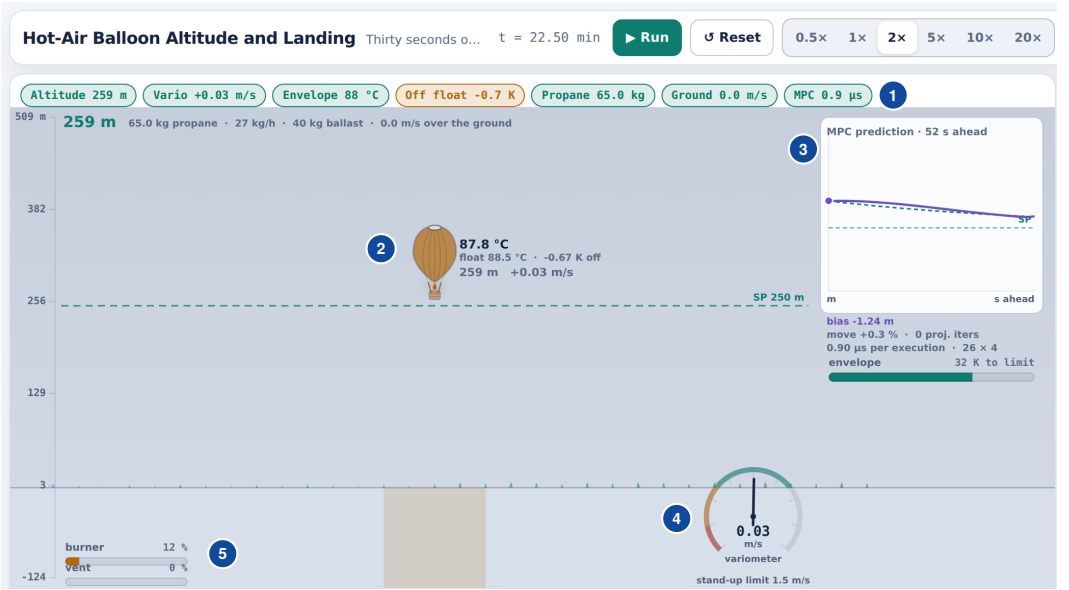



**Figure 2.4** Scenario 10 as shipped at $t = 22.5$ min, after the descent from 550 m to a 250 m setpoint, cropped to the header and the plant view. (1) The badges and (2) the balloon's label give the balloon at 259 m with the variometer at $+0.03$ m/s and an envelope of 87.8 °C that is 0.67 K below float. (3) The inset draws the controller's own predicted altitude over the next 52 s against the reference it is tracking, with the live bias of $-1.24$ m, the last move of $+0.3$ %, zero projection iterations and the execution time the page measures on the machine running it. (4) The variometer dial reads the climb rate against the stand-up limit, and (5) the actuator bars show the blast valve, commanded at 12.0 % and printed rounded, with the vent shut.

and becomes when to spend a finite resource, which is a supervisory decision rather than a regulatory one.

Scenario 5 makes the standing load visible without any disturbance at all. Over five minutes of level flight the cascade walks its temperature setpoint down by 0.32 K and holds the altitude within 0.06 m, while the hand-flown leg that follows, switched to manual at 300 s with the blast valve trimmed to 12.60 %, climbs 48.2 m in the next five minutes and is still going up at 0.38 m/s when the clock stops. Figure 2.4 shows the application under predictive control.

## 2.7 Technology in Practice

### What Measures It

The airworthiness standard for manned free balloons, 14 CFR Part 31, requires an altimeter and a rate of climb indicator on every balloon, and a fuel quantity gauge and an envelope temperature indicator on a hot-air balloon. The temperature indicator must show the maximum envelope skin temperature in operation and be marked with the limiting safe temperature of the envelope material [20]. In practice a cable plugged into the instrument pod runs to a sensor in the top of the envelope. Wireless links that carry the temperature signal to the basket remove the cable, but radio interference or the burner's thermal plume can degrade them, and the mechanical fuel gauges on the tanks are sometimes inaccurate and do not read from 0 to 100 % [1]. The chapter's TT is that crown sensor with a 2 s lag, and its ZT is the pod's altimeter and variometer.

> **TECHNOLOGY Sensor: Altimeter and Variometer**
>
> A pressure altimeter is an aneroid barometer. A stack of sealed, evacuated wafers expands as the static pressure falls with height, and a linkage turns that motion into an altitude. The vertical speed indicator, which balloonists call a variometer, feeds static pressure straight into a diaphragm and into its case through a calibrated leak, so a climb or descent leaves a pressure difference across the diaphragm. Its needle shows the trend at once but takes 6 to 9 seconds to settle on a new rate, a delay called the lag, and rough handling or turbulence lengthens the lag and makes the rate erratic. An altimeter that reads more than 75 feet off a surveyed field elevation goes for recalibration [21]. The chapter's ZT reads altitude and climb rate without a lag of its own.

### What Moves It

The standard also requires a means for the controlled release of hot air in flight [20]. In a balloon with a parachute top, opening the parachute valve partly is the normal way to cool the envelope, and the opening closes and seals itself under the upward pressure when the line is released. Manufacturers limit how long a parachute vent may stay open, and pilots time each opening [1]. The chapter's vent follows its command with a 1.0 s travel time and is never open with the burner, by the split range of Equation (2.7).

> **TECHNOLOGY Final Element: The Blast Valve of a Vaporizer Burner**
>
> Liquid propane is drawn from the bottom of the tanks in the basket. Opening the blast valve lets it into a heat exchange coil in the flame, where it vaporizes partly or completely, and the pilot light ignites it as it leaves the coil. The valve is meant to be opened and closed briskly and fully: a partly opened valve gives a soft yellow flame of unknown output, and four one-second burns give less heat than one four-second burn. Output falls by about four percent per thousand feet of altitude, propane that is too cold cannot supply enough pressure, and a butane mixture may go out when the valve is pulled [1]. The chapter's burner delivers a continuous fraction $b$ of a 2.80 MW rating, which a pilot can only approach by timing full burns.

### Where the Controller Runs

In a manned balloon the controller is the pilot. Training starts from a standard burn of four seconds, one every 25 to 40 seconds to hold level flight, and plans each maneuver in numbers of burns [1]. In a process plant a predictive controller sits above the basic loops: a survey of industrial practice draws it executing about every minute over regulatory PID control executing every second, and notes that a controller that must execute in a few milliseconds may not be able to solve a quadratic program each time, so a good suboptimal solution may be the only option [22]. The simulator's PID loops execute every 0.5 s and the predictive controller every 2 s, and its fixed twelve projected-gradient iterations are such a bounded suboptimal solution. Figure 2.5 draws the loops on balloon hardware.

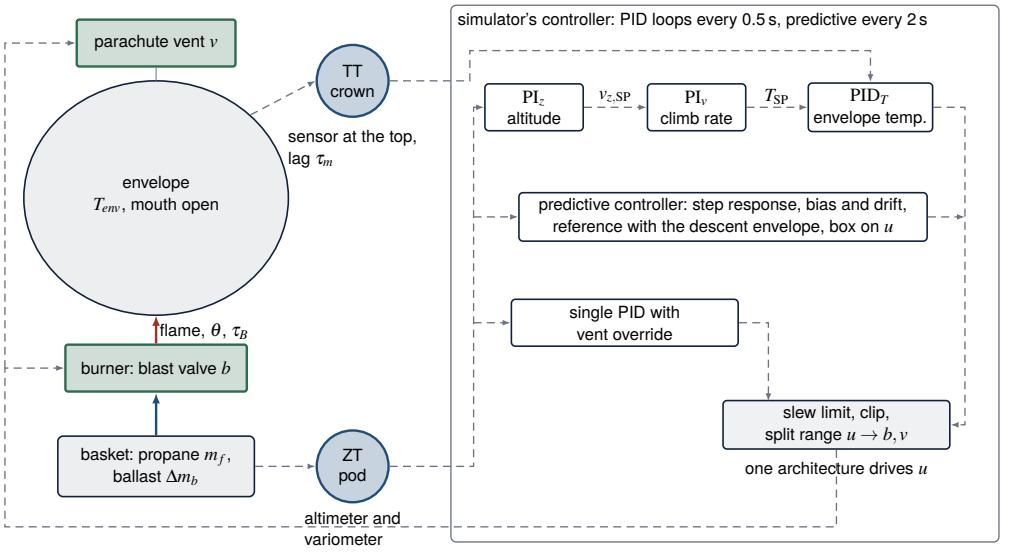



**Figure 2.5** The chapter's loops on balloon hardware. ZT, the pod's altimeter and variometer, feeds the altitude loop, the single loop and the predictive controller. TT, the crown sensor, feeds the inner temperature loop of the cascade. Whichever architecture flies sets one demand $u$, which is slew limited, clipped and split between the blast valve and the parachute vent. Notice that the burner's delay $\theta$ and lag $\tau_B$ sit between the split range and the envelope, outside every measurement.

> **IN PRACTICE** **Flying the Flare by the Numbers**
>
> Balloon training teaches a descent that slows as the ground comes up. A new pilot matches the vertical speed indicator to the altimeter, descending at 500 feet per minute from 500 feet above the ground and at 400 from 400, slows to 200 and then 100 feet per minute, and levels off with one or two standard burns of four seconds each. Below 200 feet the pilot stops reading instruments and looks down for obstacles, power lines above all. A climb follows the same rule: the rate of climb should not exceed the distance still to go [1]. The descent envelope of Equation (2.8) has the same shape, a permitted sink that falls with height.

### What Is Modernizing It

Electronic instrument systems are replacing the pressure-driven analog instruments of older balloons [1], and an instantaneous vertical speed indicator adds accelerometers to remove the variometer's lag [21]. Unmanned superpressure balloons in the stratosphere are already flown by a computer: a controller trained by reinforcement learning, which weighs cues such as wind speed and solar elevation against forecast errors and sparse wind measurements, has held balloons on station, including a 39-day controlled experiment over the Pacific Ocean [14]. The 2003 survey expected infinite prediction horizons to become a standard feature of predictive control for their theoretical properties [22]. The balloon shows that such a horizon needs a model that stays true that far ahead.

## 2.8 Limitations

The model is not a balloon flight model and cannot support a claim about a real balloon's climb performance, fuel burn or landing behavior. A single lumped envelope temperature, a fixed transport delay, textbook sphere aerodynamics and no radiation put it well short of the ascent codes used for mission planning [15], and nothing in it should be read as flight guidance.

The predictive controller carries three limitations worth naming beside its advantages. It identifies its step response by integrating a reduced copy of the same model at the declared design point, so it begins with a nearly perfect model of that point and is then flown against a different integrator, a

changing load, other altitudes and ground temperatures, and real disturbances. On a real plant the step test is where most of the model error comes from, and this application cannot show that. It has no state estimator and sees altitude alone, so neither the climb rate nor the envelope temperature reaches it and everything the linear model misses is absorbed into one bias and one drift rate. A state-space formulation with an observer and an augmented disturbance model would do better [19] and would not run in 0.78 µs. Volume 5, Chapter 2 carries the fuel in transit through its delay in the Kalman filter of such a controller. Output constraints are handled by the reference trajectory of Equation (2.22) and by the bounded projection of Equation (2.24) rather than as hard constraints in the optimization, so the descent envelope limits what the controller is asked to do and not what the balloon does, unlike the hard and soft output limits of Section 1.4.6. In the middle of a descent with unmodeled sink in the air the balloon exceeds Equation (2.8) by 0.35 m/s at 252 m on the graded landing, against 0.33 m/s for the cascade. A formulation with explicit output constraints and a terminal region would close that gap and would need a general quadratic-program solver [17].

Every upset is a commanded step or a timed event rather than the slow wandering of real air, and a trajectory that repeats to the last digit is itself unlike a flight. A thermal that arrives as a step of 2.0 m/s and decays with a 40 s time constant is tidier than the air over a plowed field, but it puts a measurement disturbance and a load disturbance on the same instrument within one run. A ground temperature that jumps from 2 to 33 °C between two legs of scenario 8 cannot occur in one flight, but it puts the two density altitudes side by side on one trend. Five features were kept because no controller for this plant can avoid them, which are the pure delay in front of the actuator lag, the double integration, the square-root climb characteristic, the split range that is not gain matched and the constraint on arrival speed. Section 2.3.7 states where each individual assumption breaks.

> ⚠ **OPEN CHALLENGE** **A Disturbance Model for an Output That Integrates**
>
> The constant output bias that makes a step-response controller offset-free fails on a purely integrating output, and industrial packages patch it ad hoc. One typical fix assumes a constant integrating disturbance and assigns a fraction of each prediction error to it through a rotation factor. A survey of industrial predictive control in 2003 found that factor hard to choose in practice, found the control of noisy integrating systems often poor, and called the design of sensible disturbance models a ripe area for research [22]. The simulator's altitude carries no noise, so its drift ramp has never been tested against noise. Progress would be a disturbance model that follows a drift without amplifying noise.

## 2.9 Discussion

Two results of this chapter transfer beyond balloons. The first is that a prediction horizon can be too long for a reason unrelated to computing cost. Table 2.8 shows a loop driven into a 214.4 m limit cycle by a horizon of 120 s on a plant whose maneuvers take three minutes, and shows that move suppression does not remove it. The diagnostic is useful in the field, because a cycle that persists across three decades of move suppression is a model problem and not a tuning problem, and shortening the horizon costs nothing that the terminal weight cannot restore. The second is the disturbance model. Dynamic matrix control is offset-free on self-regulating plants because the constant output bias captures a sustained load exactly [18], and the correct generalization for an integrating output is a ramp. Table 2.9 prices it at 129 m of steady offset and the difference between a stand-up landing and a firm one, obtained from two lines of arithmetic that add nothing measurable to the execution time.

The same two questions arise wherever an integrating output sits behind a delayed actuator that must arrive somewhere at a bounded speed. Ship steering integrates a yaw rate produced by a rudder whose hydraulics lag [2], building thermal control drives a room through a plant whose transport time is measured in minutes [3], and unmanned balloons steer entirely by choosing an altitude, which makes the loop in this application the inner loop of a navigation problem [14]. Each has an

actuator that is slow in one direction and fast in the other, each has a resource that is spent rather than modulated, and each ends against a constraint on a rate rather than on a position.

The equations of this chapter also support methods beyond the four architectures implemented. Nine states with two constrained inputs, a hard fabric-temperature limit, a finite fuel inventory and a discrete ballast release form a mixed-integer optimal control problem in which the decision of when to spend a sandbag is made by the optimizer rather than by the pilot. Moving horizon estimation on the same equations would recover the vertical air motion and the envelope temperature from altitude alone, which would replace the single bias and drift pair with a real disturbance state.

## 2.10 Conclusion

This chapter studied a hot-air balloon as an integrating output behind a delayed and lagged actuator that must arrive at the ground under a speed limit. The simulator integrates nine states with buoyancy, weight, quadratic drag and a potential-flow added mass of 1785 kg, an open-system envelope energy balance of 2.972 MJ/K against 4.6 kW/K of fabric loss, a 4 s burner transport delay, a split-range vent that removes heat 1.72 times faster than the burner supplies it at the design point, fuel burn at 29.9 kg/h and a standard atmosphere. An open-loop 20 % blast-valve step moves the delivered heat at 4.0 s, the variometer past 1.0 m/s at 33.2 s, and the balloon 264.0 m higher at 120 s with no steady state to read a gain from. A single loop on altitude swings 125.5 m peak to peak and grows to 201.1 m, while a three-deep cascade holds 0.8 m and a step-response predictive controller holds 6.8 m on the same maneuver. On the graded landing the three arrive at 7.85, 1.27 and 1.39 m/s against a 1.5 m/s stand-up limit. The cascade and the predictive controller use an explicit descent profile, and the actual arrival rate also depends on tuning and tracking error. Two implementation results were measured rather than assumed. Extending the prediction horizon from 52 to 120 s puts the loop into a 214.4 m limit cycle that move suppression from 0.004 to 5 does not remove, because the linear step response over-predicts a $\Delta u = 0.20$ move by 29 % at 90 s. Replacing the constant output bias with a filtered drift ramp takes the settled offset from 132.9 to 3.9 m and the touchdown from 2.18 to 1.39 m/s, at 0.78 µs per execution measured over 100000 executions. A fuller model would add envelope stratification, a burn-rate-dependent transport delay, solar and infrared terms in the energy balance, and a state-space predictive controller with an observer for comparison against the step-response form. The balloon is a plant on which to learn to predict an output that answers half a minute late.

**KEY POINTS**

- A balloon answers its burner late, and a burner set back to its trim holds whatever climb rate the balloon already has.
- One PID loop on altitude never settles, while a three-deep cascade gives each loop one slow element and settles.
- On the graded landing the cascade and the predictive controller arrive at 1.27 and 1.39 m/s, the single loop at 7.85 m/s.
- A 120 s horizon cycles at every move suppression, because the step response is wrong that far ahead.
- On an integrating output a filtered drift ramp takes the offset of a constant bias from 132.9 to 3.9 m.

**Where else this applies.** An output that integrates behind a wave travel time is also the plant of each reach of the canal in Chapter 5, where the delay sits in the water rather than in a burner plume.

## References

[1] Federal Aviation Administration, “Balloon flying handbook,” FAA-H-8083-11B, Chapters 2, 7 and Glossary, https://www.faa.gov/regulations_policies/handbooks_manuals/aviation/balloon_flying_handbook, 2024.

[2] J. Van Amerongen, “Adaptive steering of ships—A model reference approach,” *Automatica*, vol. 20, no. 1, pp. 3–14, 1984.

[3] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, pp. 15–27, 2012.

[4] K. J. Åström, C. C. Hang, and B. C. Lim, "A new Smith predictor for controlling a process with an integrator and long dead-time," *IEEE Transactions on Automatic Control*, vol. 39, no. 2, pp. 343–345, 1994.

[5] K. J. Åström and T. Hägglund, *Advanced PID Control*. Research Triangle Park, NC: ISA – The Instrumentation, Systems, and Automation Society, 2006.

[6] E. F. Camacho and C. Bordons, *Model Predictive Control*, 2nd ed., ser. Advanced Textbooks in Control and Signal Processing. London: Springer, 2007.

[7] C. E. Garcia and A. M. Morshedi, "Quadratic programming solution of dynamic matrix control (QDMC)," *Chemical Engineering Communications*, vol. 46, no. 1-3, pp. 73–87, 1986.

[8] B. W. Bequette, "Process control practice and education: Past, present and future," *Computers & Chemical Engineering*, vol. 128, pp. 538–556, 2019.

[9] J. Park, R. A. Martin, J. D. Kelly, and J. D. Hedengren, "Benchmark temperature microcontroller for process dynamics and control," *Computers & Chemical Engineering*, vol. 135, p. 106736, 2020.

[10] C. C. Gillispie, *The Montgolfier Brothers and the Invention of Aviation, 1783–1784*. Princeton, NJ: Princeton University Press, 1983.

[11] W. Marion, "Fuel requirements of a hot air balloon," in *7th Aerodynamic Decelerator and Balloon Technology Conference*. Reston, VA: American Institute of Aeronautics and Astronautics, 1981.

[12] *Balloon Flying Handbook*, U.S. Federal Aviation Administration, Washington, DC, 2022, FAA-H-8083-11B.

[13] M. S. Smith, "Demonstration of fine altitude control on stratospheric balloons to achieve a desired ground track," in *AIAA Balloon Systems Conference*. Reston, VA: American Institute of Aeronautics and Astronautics, 2017.

[14] M. G. Bellemare, S. Candido, P. S. Castro, J. Gong, M. C. Machado, S. Moitra, S. S. Ponda, and Z. Wang, "Autonomous navigation of stratospheric balloons using reinforcement learning," *Nature*, vol. 588, no. 7836, pp. 77–82, 2020.

[15] R. Farley, "BalloonAscent: 3-D simulation tool for the ascent and float of high-altitude balloons," in *AIAA 5th ATIO and 16th Lighter-Than-Air Systems Technology and Balloon Systems Conferences*. Reston, VA: American Institute of Aeronautics and Astronautics, 2005.

[16] S. K. Karanfilian and T. J. Kotas, "Drag on a sphere in unsteady motion in a liquid at rest," *Journal of Fluid Mechanics*, vol. 87, no. 1, pp. 85–96, 1978.

[17] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[18] K. R. Muske and T. A. Badgwell, "Disturbance modeling for offset-free linear model predictive control," *Journal of Process Control*, vol. 12, no. 5, pp. 617–632, 2002.

[19] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[20] Federal Aviation Administration, "Airworthiness standards: Manned free balloons," Code of Federal Regulations, Title 14, Part 31, 2024 edition, U.S. Government Publishing Office, https://www.govinfo.gov/content/pkg/CFR-2024-title14-vol1/pdf/CFR-2024-title14-vol1-part31.pdf, 2024.

[21] ——, "Pilot's handbook of aeronautical knowledge," FAA-H-8083-25B, Chapter 8, Flight Instruments, https://www.faa.gov/sites/faa.gov/files/regulations_policies/handbooks_manuals/aviation/phak/10_phak_ch8.pdf, 2016.

[22] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

## 2.11 Exercises

**2.1** Concept check · 1 min

In Scenario 1 the balloon floats trimmed at 300 m. Full burner is held for five seconds and the blast valve is then set straight back to its 12.65 % trim. What does the balloon do over the next seven minutes?

(a) It keeps climbing, because the trim holds the extra envelope temperature and so the climb rate

(b) It climbs for about a minute and returns to 300 m, because the valve is back at its trim

(c) It climbs about fifty meters and floats there, because the thinner air balances the hotter envelope

(d) It sinks back below 300 m, because the fabric loss removes the extra heat within a minute

**2.2** Concept check · 2 min

The predictive controller holds the balloon at 300 m with no applied disturbance. With the constant output bias alone, and the drift ramp removed, it settles 132.9 m from its setpoint. Which explanation is right?

(a) The bias is filtered too slowly to follow the fuel burn, so it lags the plant by a fixed amount

(b) The burning fuel makes the model fall behind the plant by a fixed amount each sample, which a constant cannot represent

(c) The step response was identified with full tanks, so its gain is wrong by the share of fuel since burned

(d) The projection clips the planned moves to the actuator box, so the correction is never applied

**2.3** Concept check · 2 min

With the prediction horizon lengthened from 52 to 120 s, the held balloon cycles by more than 140 m peak to peak at every move suppression from 0.004 to 5, and also with the terminal weight set to 1. What does that pattern point to?

(a) The optimizer is too aggressive, so a move suppression above 5 would remove the cycle

(b) The terminal weight pulls the plan toward the last eight points, where the predictions are least reliable

(c) The linear step response is wrong at the far end of the horizon, so every plan rests on a model error

(d) The longer horizon takes longer to compute, so the controller misses executions and acts late

**2.4** Theory and calculation · 6 min

At the design point the air at 300 m has $p_a = 97773\,\mathrm{Pa}$, $T_{amb} = 13.05\,^{\circ}\mathrm{C}$ and $\rho_a = 1.19012\,\mathrm{kg/m^3}$. The balloon has $V = 3000\,\mathrm{m^3}$, a load of 757 kg, $C_D = 0.55$ and $A = 250\,\mathrm{m^2}$, with $R = 287.05\,\mathrm{J/(kg\,K)}$ and $g = 9.80665\,\mathrm{m/s^2}$.

(a) Find the float temperature from Equation (2.11). *Checkpoint:* 90.06 °C.

(b) Find the lift per kelvin and the drag constant $k_d = \frac{1}{2}\rho_a C_D A$ at float, and the steady climb rate 1 K and 4 K above float. *Checkpoint:* 75.96 N/K, 81.82 N per $(\mathrm{m/s})^2$; 0.96 and 1.93 m/s.

(c) Find how far above float the envelope must be to arrest a steady descent of 3 m/s, and of 1 m/s. *Checkpoint:* 9.7 K and 1.1 K.

**2.5** Theory and calculation · 6 min

At float the envelope air has a density of $0.93778\,\mathrm{kg/m^3}$ at 90.06 °C, against 13.05 °C outside. Use $V = 3000\,\mathrm{m^3}$, $c_p = 1005\,\mathrm{J/(kg\,K)}$, $C_{fab} = 145\,\mathrm{kJ/K}$, $UA = 4600\,\mathrm{W/K}$, $K_{vent} = 50000\,\mathrm{W/K}$, $Q_{max} = 2.80\,\mathrm{MW}$, $\eta_b = 0.92$ and $\mathrm{LHV} = 46.4\,\mathrm{MJ/kg}$.

(a) Find the thermal capacity $C_{env}$ and the time constant $C_{env}/UA$. *Checkpoint:* 2.972 MJ/K and 646 s.

(b) Find the trim burner opening and the propane it burns. *Checkpoint:* 12.65 %; 8.30 g/s, or 29.9 kg/h.

(c) Find the rate of change of the envelope temperature from float at full burner and at a fully open vent, and their ratio. *Checkpoint:* $+0.82\,\mathrm{K/s}$, $-1.41\,\mathrm{K/s}$, 1.72.

**2.6** Theory and calculation · 7 min

Take the path from demand to altitude as one pole at the origin, a 4 s delay and real poles at 1.25/s, $1.548 \times 10^{-3}$/s (the envelope loss) and $3.056 \times 10^{-2}$/s (the drag lag at 1 m/s).

(a) Find the phase of the plant at $\omega = 0.01$/s and at 0.02/s. *Checkpoint:* $-192.1°$ and $-214.3°$.

(b) The single loop uses the ideal PID $K_c\,(1 + 1/(\tau_I s) + \tau_D s)$ with $\tau_I = 1200$ s and $\tau_D = 130$ s. Find the phase it adds at $\omega = 0.01$/s and the phase of the loop there. *Checkpoint:* $+50.6°$; $-141.5°$.

(c) Repeat (b) with $\tau_D = 0$, and explain the measured result that the loop without derivative puts the balloon on the ground. *Checkpoint:* $-4.8°$; $-196.8°$.

**2.7** App challenge · 8 min

Load Scenario 2, *The step test the MPC is built from*, at the default seed and change nothing: the balloon is in Manual and the Burner blast valve steps from 12.65 to 32.65 % at $t = 20$ s. Run to $t = 150$ s at the default speed 2 (about 75 s), pause and export the CSV file. Its time column is in minutes for a run this long, so convert it to seconds before reading.

(a) From `burner_cmd_pct` and `burner_delivered_pct`, find how long after the commanded step the delivered heat leaves 12.65 %. *Checkpoint:* 4.0 s (rows at 20.2 and 24.2 s).

(b) From `vz_mps`, find how long after $t = 20$ s the climb rate first reaches 0.5 and 1.0 m/s, to 0.3 s. *Checkpoint:* 24.2 s and 33.4 s.

(c) From `alt_m`, find the altitude gained between $t = 20$ and 72 s, and compare it with what the controller's identified step response, $a_{26} = 210.3$ m per unit of demand at 52 s, predicts for this step. Give the gain to 0.3 m. *Checkpoint:* 39.6 m measured against 42.1 m predicted.

**2.8** App challenge · 8 min

Load Scenario 3, *One PID loop on altitude, hunting*, at the default seed and change nothing: the Target altitude moves from 300 to 450 m at $t = 60$ s. Run to $t = 2400$ s at speed 20 (about 2 min) and export the CSV file. Repeat with Scenario 4, *The same level-off on the cascade*, as loaded. Convert the time column from minutes to seconds and read `alt_m`.

(a) For Scenario 3, give the peak-to-peak altitude over 300 to 900 s and over 1800 to 2400 s, and the altitude at 2400 s. *Checkpoint:* 125.5 m, 201.1 m; 307.8 m.

(b) For Scenario 4, give the highest altitude after the setpoint change and the last time the altitude is outside 435 to 465 m. *Checkpoint:* 465.2 m; 280.6 s, which is 220.6 s after the change.

(c) Explain why the cascade settles where the single loop does not.

**2.9** App challenge · 9 min

Load the challenge, *Land it under 1.5 m/s*, at the default seed and run it as given at speed 5 until it pauses at touchdown (about 22 s). Read the Touchdown tile and its hover text. Press Reveal suggested tuning, press Reset, run to touchdown again (about 40 s) and export the CSV file. Press Reset, set Control architecture to Cascade before Run (the suggested settings stay), run to touchdown and export again.

(a) Give the touchdown sink rate, the verdict and the ground speed as given, and say which clauses of the target fail. *Checkpoint:* 7.85 m/s, hard, 4.2 m/s; both.

(b) Give the sink rate, the ground speed and the time of touchdown under the suggested predictive controller, and the propane burned (77 kg less `fuel_kg` in the last row). *Checkpoint:* 1.39 m/s, 3.5 m/s, 202 s; 2.44 kg.

(c) Give the same for the cascade. Compare both sink rates with the suggested flare speed of 0.70 m/s, and say what the comparison means for how a landing is validated. *Checkpoint:* 1.27 m/s, 3.5 m/s, 229 s; 2.44 kg.

**2.10** Application area · 4 min

An operator plans a flight that must hold 300 m with the 757 kg load of Table 2.10. On a 33 °C afternoon the air at 300 m has $p_a = 97979\,\mathrm{Pa}$ and $\rho_a = 1.12206\,\mathrm{kg/m^3}$.

(a) From Table 2.10, give the margin between the float temperature and the 120 °C fabric limit at 28 and at 33 °C. *Checkpoint:* 8.9 K and 0.7 K.

(b) Find the load that floats at 300 m on 110 °C that afternoon, a 10 K margin, and how much must come off. *Checkpoint:* About 694 kg, so 63 kg less.

(c) Name the handles the operator has, and say why none of them is a controller setting.

**2.11** Application area · 3 min

The model holds the envelope at one temperature, while a real envelope is stratified and its crown reaches the fabric limit first (Section 2.3.7). Suppose the crown runs 15 K above the lumped mean.

(a) Find the mean temperature at which the crown reaches 120 °C, and say which ground temperatures of Table 2.10 still let 757 kg float at 300 m. *Checkpoint:* 105 °C; 2, 15 and 24 °C, the last with 0.4 K to spare.

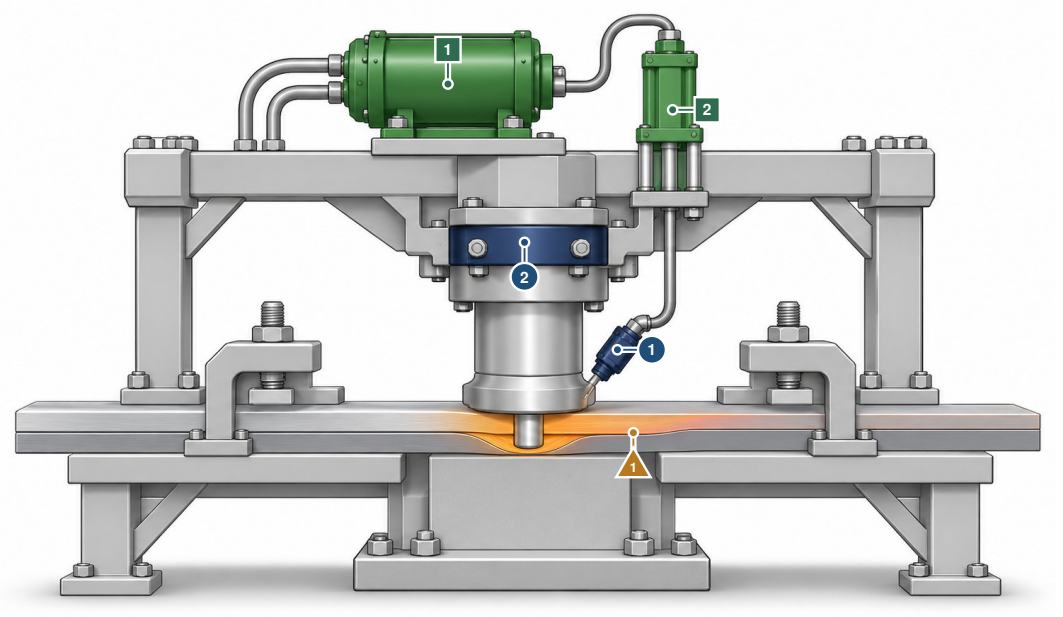
(b) Say what a pilot should read in flight and what the lumped model gets right.

**2.12** Review the agent · 5 min

An AI assistant reviews Scenario 3 and writes: “The single loop hunts because a derivative time of 130 s is far too long. On a plant this slow the derivative mostly amplifies measurement noise. Set PID derivative time to 0 and lower PID gain to 0.02 % per m, and the loop becomes a gentle PI controller that settles without overshoot.” Find the flaw and say how you would check the proposal.

CHAPTER 3

# Friction Stir Welding Simulator for Regime-Dependent Tuning and Predictive Control




**The controller moves** 1 spindle drive (spindle speed) 2 axial axis drive (plunge depth)

**The controller reads** 1 tool temperature thermocouple 2 axial load cell

**What disturbs it** 1 thin patch in the plate

*Notice that the spindle speed sets the heat, and that the weld answers it as one plant while the tool plunges and as another once it travels. Illustration, not to scale.*

### IN PLAIN TERMS Welding Metal Without Melting It

**The main idea.** A friction stir weld is made by a spinning tool that is pushed into the seam between two plates and dragged along it. Rubbing heats the metal until it is soft enough to be stirred shut. The spinning speed sets how hot the weld runs.

**The challenge.** While the tool sinks in, the weld answers a change of speed fast and strongly. Once the tool moves along, it answers slowly and weakly. An automatic speed controller that learns the weld at one of these times is wrong at the other.

**Why it matters.** The wrong choice leaves hollow or squeezed-out metal that must be cut away.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the fact in Section 3.1, the history in Section 3.2, the two try-it segments and the second fact in Section 3.6, and the key points |
| **Study** | 55 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 20 min | the first page, the results (Section 3.6), Technology in Practice (Section 3.7), Limitations (Section 3.8) and the open challenges |

**The control question** A step test in quasi-steady welding returns a gain 1.7 times smaller and a time constant ten times longer than the same test through the plunge of the same weld. Which test should fill the dynamic matrix, and what does the wrong regime cost in millimeters of defect?

**Learning objectives**

- Explain from the flow stress why heat generation is self-limiting and why the plunge is a different plant.
- Identify a first-order-plus-dead-time model and compute ITAE servo and regulator tunings from it.
- Compare servo and regulator tuning phase by phase, and explain how the entry into Auto changes the result.
- Predict how a step-response controller whose model gain is too high responds, and why move suppression does not rescue it.
- Choose position or force control of the axial axis from the root clearance as well as the heat.

**Prerequisites** Sections 1.4.2 to 1.4.5 (step response, gain matrix, move suppression, output bias), 1.6.2 and 1.6.4, and Chapter 2.

**The app** https://apmonitor.com/apps/friction-stir-welding/

**See also** Volume 7, Chapter 2 for heat that grows with temperature and Volume 7, Chapter 13 for a limit set by the spindle speed. Volume 1, Chapter 2 for a model that follows the operating point and Volume 1, Chapter 12 for identification from records.

## 3.1 Introduction

A friction stir weld is made by a rotating shouldered tool that is pushed into the joint line between two plates and then dragged along it. Nothing melts. Friction under the shoulder and plastic shear around the pin bring the aluminum to roughly four fifths of its absolute melting temperature, where the metal is soft enough to be extruded around the pin and forged shut behind it [1, 2]. The quantity that decides whether the joint is any good is the tool temperature. Below the window the metal does not flow fast enough to close the wake behind the pin and the weld carries a continuous tunnel void, and above it the metal extrudes sideways from under the shoulder as flash while the root softens and the tool wears [1]. In the simulator of this chapter that window is 435 to 505 °C at 120 mm/min and both of its edges move with the traverse speed.

**DID YOU KNOW? Half a Mile of Weld on Every Tank**

Each Space Shuttle External Tank carried about half a mile of welds. About 700 feet of it, eight longitudinal joints on the liquid hydrogen barrel and four on the liquid oxygen barrel, was made by friction stir welding, and a plasma arc process made the rest [3].

A block diagram of that loop is a first-order lag with dead time and a single manipulated variable, and it hides the one property that makes the weld hard to control. Heat generation is proportional to spindle speed times interface shear stress, and that shear stress follows the flow stress of the shear layer, which falls by a factor of eighteen between room temperature and 500 °C [4, 5]. Heat generation is therefore self-limiting, which is why a friction stir weld does not run away and melt the joint and why the linearization of the loop moves with the operating point. The exothermic reactor of Volume 7, Chapter 2 is the opposite case, a heat source that grows with temperature until the vessel can ignite. The misconception the simulator is built to correct is that a single well-tuned controller is the goal. Measured on the shipped model, the same loop returns 0.1425 °C/rpm with a 12.2 s time constant in quasi-steady welding and 0.240 °C/rpm with a 1.2 s time constant during the plunge of the same weld, so the controller that wins one phase is mistuned for the other and the choice is which phase to be wrong in.

The class of problem is general. It is the plant whose linearization changes materially between operating phases inside one batch, so that a model-based controller inherits the operating point of the

experiment that identified it. Volume 1, Chapter 2 meets the same inheritance on a tank whose gain and time constant both follow the level. A batch reactor between charging and reaction, a furnace between light-off and soak, an extruder between purge and running, and a crystallizer between seeding and growth all have that shape, and the standard answers are gain scheduling, a controller identified over several operating points, adaptive identification, or a nonlinear predictive formulation [6, 7]. Surveys of industrial predictive control report that the identification experiment, and not the optimizer, is where most of the commissioning effort and most of the later trouble sit [8]. What a simulator adds is the ability to run that experiment twice on the same plant, in two phases of one weld, and then run both resulting controllers over a weld and count the defect each one leaves.

This chapter works with a friction stir welding simulator whose two dynamic matrix controllers come from two step tests taken on the same weld in different phases, so that the price of identifying a model-based controller in the wrong regime is measured in defect length rather than asserted. Section 3.2 traces how the machine moved from position to force control and then to closed-loop temperature control. Section 3.3 sets out the nine-state thermal and mechanical model, with its instrument chain, its force-limited axial axis and its quality window, completely enough to be rebuilt from the text alone. Section 3.4 identifies three plants inside one weld, and Section 3.5 describes the five architectures that share the one actuator: the operator, two ITAE-tuned PI loops and two step-response predictive controllers. The predictor, the gain matrix and the output bias are those of Sections 1.4.2 to 1.4.5. Section 3.6 runs the scenario sequence, from a weld made by hand to a graded challenge in which force control on the axial axis is a trap. Section 3.7 describes the sensors, drives and controller platforms of real welding machines. Sections 3.8 and 3.9 state what the model cannot support and which methods it could carry next.

## 3.2 Friction Stir Welding and the Arrival of Temperature Control

The process was invented at The Welding Institute in Cambridge in 1991 and patented there, and the first published account aimed at industry set out what the solid-state route buys, which is a joint in alloys that cannot be fusion welded at all, with no porosity, no solidification cracking and very little distortion [9]. Adoption followed the alloys. Shipbuilding and rail took extruded aluminum panels first, because a long straight butt joint in 6xxx extrusion is the easiest weld the process makes, and aerospace followed with the 2xxx and 7xxx alloys whose fusion weldability is poor [1, 9]. By the time the process was reviewed at length it had moved from a laboratory curiosity to a production method for aircraft floor panels, rail carriage sides, fuel tanks and, later, the cooling plates under electric-vehicle batteries [1, 2].

> ⧖ HISTORY **1995: A Laboratory Curiosity Reaches a Rocket Tank**
>
> Friction stir welding was conceived in the United Kingdom, and when NASA took it up in 1994 it was little more than a laboratory curiosity. A laboratory system made the agency's first welds in the fall of 1995, and development began on a five-axis horizontal boring mill given over entirely to the work. Its engineers found that the process needs the control of only three variables, shoulder plunge depth, tool rotational speed and travel speed, and that once optimized they repeat weld after weld. A simulated hydrogen tank barrel 27.5 feet in diameter was welded in 1998, and in 2001 the process entered production of the Space Shuttle External Tank, for the longitudinal welds of its hydrogen and oxygen barrels [3]. Those three variables are the recipe that this chapter's loops act on.

Automation of the machine came before automation of the process. The first generation of production equipment held the tool at a programmed height above the machine table, because a milling machine already did that and because position is what a numerical control program naturally commands. Position control on a plate of varying thickness costs shoulder engagement directly, so the axial force and the heat input follow the rolled stock rather than the recipe, and the industry moved to force control on the axial axis for that reason [10]. The requirements that make force control

work against finite machine stiffness and a compliant fixture, chiefly the load-cell bandwidth and the axial stiffness of the head, were identified explicitly for robotic equipment where the compliance is largest [11]. Force control is the production default now, normally interlocked against a position limit rather than run alone, because a force loop answers a soft patch of plate by sinking a pin that has only a fraction of a millimeter of clearance to the backing anvil [10]. A robot that pushes a peg into a hole faces the same choice once contact fixes the part, as Volume 20, Chapter 13 shows.

Neither position nor force is the variable that sets weld quality. Temperature is, and closing a loop on it required a temperature measurement from inside a rotating tool. That measurement became practical with a thermocouple embedded a few millimeters behind the shoulder and telemetry or a slip ring off the spindle, and the first closed-loop demonstrations used it to hold a tool temperature by manipulating spindle speed, reporting that the loop removed the sensitivity to the fixturing and the plate condition that an open-loop recipe carries [12]. Model-based control of the same loop followed. A first-order-plus-dead-time model identified from a step in spindle speed supports a predictive controller that outperforms a PID loop on the transient parts of a weld, and the comparison has been reported for the initial traverse and for quasi-steady welding separately [13], and more recently with a distributed thermal model inside the controller [14]. Qualification practice has not moved with it. A friction stir weld is still qualified as a recipe, which is a spindle speed, a traverse speed and a plunge depth run open loop [10], and the scrap that practice produces is concentrated at the two ends of every weld, where the recipe is wrong.

## 3.3 Process Model

The simulated joint is a 6.35 mm AA6061-T6 butt weld 245 mm long in a 300 mm plate clamped to a steel backing anvil. The tool has a flat 12 mm shoulder and a cylindrical 4 mm pin 6.00 mm long, so a 0.15 mm shoulder plunge leaves 0.20 mm of unwelded parent metal under the pin. The machine program is a plunge to depth, a 12 s dwell, the traverse, a run-out at the plate edge and a retract. Figure 3.1 shows the arrangement, the four thermal nodes, the conductances between them and every symbol used below. The controlled variable is the tool temperature $T_t$ as reported by the rotating thermocouple, and the manipulated variable is spindle speed $\omega$ in every architecture the application offers.

What the loops interact with is classified in Table 3.1. The plant is written in the units a welding engineer uses, temperature in °C, spindle speed in rpm, traverse speed in mm/min, depth in mm, force in kN, torque in N m and time in seconds, with SI substituted inside the stress and heat expressions and converted at the boundary. Nine states are integrated. Four of them are temperatures, two are drive states, one is the axial position of the shoulder above the machine table, one is the distance welded and one is cumulative tool wear.

### 3.3.1 Contact Geometry

Everything mechanical follows from the tool height, so the geometry is written first, as shown in Equation (3.1).

$$\begin{aligned} z &= t + \Delta t - h_s, \qquad c = h_s - H_p \\ f_s &= \mathrm{clamp}(z/z_{eng},\,0,\,1), \quad L_p = \mathrm{clamp}(z + H_p,\,0,\,H_p) \\ f_{tip} &= \mathrm{clamp}((z + H_p)/0.2\,\mathrm{mm},\,0,\,1) \\ A_{sh} &= \pi\left(R_s^2 - R_p^2\right) f_s, \qquad A_{tip} = \pi R_p^2 f_{tip} \end{aligned} \tag{3.1}$$

where $z$ is the shoulder plunge below the plate surface in mm, $c$ is the unwelded parent metal left under the pin, $t$ is the nominal plate thickness, $\Delta t$ is the thickness deviation of the rolled stock, $h_s$ is the height of the shoulder face above the table, $H_p$ is the pin length, $f_s$ is the fraction of the shoulder in contact, $L_p$ is the buried length of the pin and $f_{tip}$ is the fraction of the pin face in contact. The shoulder radius is derated by wear, $R_s \leftarrow R_s(1 - 0.04W)$, which is one of the two ways wear reaches the heat balance.

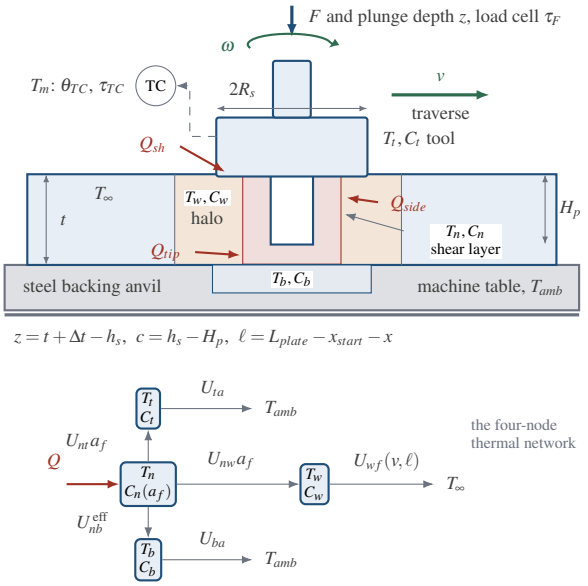



**Figure 3.1** Section through the weld, above, and the four-node thermal network it reduces to, below. The rotating tool generates $Q_{sh}$ under the shoulder, $Q_{side}$ along the buried pin and $Q_{tip}$ at the pin face, all into the shear layer at $T_n$, whose capacity and whose three conductances scale with the engaged contact fraction $a_f$. Heat leaves that node up the tool to $T_t$, sideways into the heated halo at $T_w$ and down into the anvil interface at $T_b$, and leaves the halo into cold plate at $T_\infty$ through a conductance that changes with traverse speed $v$ and with the remaining distance $\ell$ to the plate edge. The controller reads $T_m$ from a thermocouple in the rotating tool and manipulates $\omega$.

**Table 3.1** Variables the Loops Interact With, Read at the Design Weld of 900 rpm, 120 mm/min and 0.15 mm of Shoulder Plunge, 90 mm Along the Joint.

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $T_t$ | CV, measured | 471.4 °C | 380–560 SP |
| $T_m$ | measurement of $T_t$ | 471.4 °C | delayed, lagged |
| $\omega$ | MV | 900 rpm | 200–1500 |
| $z$ | MV in force mode | 0.15 mm | 0–0.45 |
| $F$ | inner CV, force mode | 9.50 kN | 2–18 SP |
| $v$ | operator setpoint | 120 mm/min | 20–300 |
| $\Delta t$ | DV, unmeasured | 0 mm | ±0.25 |
| $a_d$ anvil contact | DV, unmeasured | 1.0 | 0.15–1.2 |
| $T_\infty$ | DV, measured | 22 °C | −20–80 |
| $W$ tool wear | DV, unmeasured | 0 | 0–0.8 initial |
| $T_n$ | state | 516.6 °C | — |
| $T_t$ | state | 471.4 °C | — |
| $T_w$ | state | 395.6 °C | — |
| $T_b$ | state | 327.4 °C | — |
| $\omega_A$, $v_A$ | drive states | 900, 120 | — |
| $h_s$ | shoulder height | 6.20 mm | clamped |
| $x$ | distance welded | 90 mm | 0–245 |
| $W$ | wear state | 0 | 0–1 |

The engaged area also sets how large the shear layer is, as the fraction $a_f$ of fully engaged contact

area in Equation (3.2).

$$a_f = \mathrm{clamp}\left( \frac{f_s\left(R_s^2 - R_p^2\right) + R_p^2 f_{tip} + 2R_p L_p}{R_s^2 + 2R_p H_p},\ 0,\ 1.2 \right) \tag{3.2}$$

The three conductances out of the shear layer and its own thermal capacity are all scaled by $a_f$, so a pin tip touching cold metal is a small fast node and a fully engaged tool is a large slow one. That scaling is what makes the plunge a different plant rather than the same plant at a different temperature.

### 3.3.2 Flow Stress and the Interface Shear Stress

The flow stress of the shear layer is the single expression from which the regime change follows, and it is written as a decaying exponential in temperature with a residual, collapsed linearly through the solidus to liquidus range, as shown in Equation (3.3).

$$\begin{aligned} \sigma(T) &= \max(0.2\,\mathrm{MPa},\ \sigma_b(T)\,\phi(T)) \\ \sigma_b(T) &= \sigma_0 \mathrm{e}^{-T/T_\sigma} + \sigma_{res} \\ \phi(T) &= \mathrm{clamp}\left( \frac{T_{liq} - T}{T_{liq} - T_{sol}}, 0, 1 \right) \end{aligned} \tag{3.3}$$

where $\sigma_0$ is a pre-exponential in MPa, $T_\sigma$ is the decay temperature in °C and $\sigma_{res}$ is the residual strength near the solidus. The constants are fitted to hot-working flow stress for 6061 at the strain rates a friction stir shear layer sees, of order 100/s, rather than to the quasi-static yield stress [4]. Equation (3.3) returns 1165 MPa at 22 °C, 209 MPa at 300 °C, 115 MPa at 400 °C, 65 MPa at 500 °C and 42 MPa at the solidus; from room temperature to 500 °C it falls by a factor of eighteen.

The interface carries either the sliding shear stress or the shear yield of the metal, whichever is smaller, and the transition between the two is the plunge to weld handover. Equation (3.4) writes it as a smooth minimum so that no corner appears in the Jacobian at the crossover.

$$\begin{aligned} \tau_{stick} &= \frac{\sigma(T_n)}{\sqrt{3}} \left( \frac{\max(1,\omega)}{\omega_{ref}} \right)^m \\ \mathrm{sm}(a,b) &= \left(a^{-n_{sm}} + b^{-n_{sm}}\right)^{-1/n_{sm}} \\ \tau_{sh} &= (1 - w_{slip} W)\,\mathrm{sm}(\mu p_{sh},\ \tau_{stick}) \\ \tau_{pin} &= (1 - w_{slip} W)\,\mathrm{sm}(\mu p_{pin},\ \tau_{stick}) \end{aligned} \tag{3.4}$$

where $\tau_{stick}$ is the von Mises shear yield with a strain-rate exponent $m$ carried on spindle speed, $\mu = \mu_0(1 - 0.35W)$ is the sliding friction coefficient, $p_{sh}$ and $p_{pin}$ are the contact pressures of Equation (3.6), $n_{sm}$ sharpens the smooth minimum and $w_{slip}$ derates both branches for tool wear. Cold metal is strong and the indentation pressure under the shoulder is large, so $\mu p_{sh}$ is the smaller term and the contact slides. Once the weld is running the metal is soft and the shear yield is the smaller term, so the contact sticks [5]. Both regimes are logged, and at the design point the sticking branch is the smaller of the two, with $\tau_{stick} = 34.76\,\mathrm{MPa}$ against $\mu p_{sh} = 35.16\,\mathrm{MPa}$, so that Equation (3.4) returns $\tau_{sh} = 29.40\,\mathrm{MPa}$ under the shoulder.

### 3.3.3 Heat Generation and Spindle Torque

Frictional and plastic work is the Schmidt integral of surface velocity against interface shear stress over the three contact surfaces of a flat shoulder with a cylindrical pin, as shown in Equation (3.5) [5, 15].

$$\begin{aligned} Q_{sh} &= \tfrac{2}{3}\pi\,\tau_{sh}\,\Omega\left(R_s^3 - R_p^3\right) f_s \\ Q_{tip} &= \tfrac{2}{3}\pi\,\tau_{pin}\,\Omega\,R_p^3\,f_{tip} \\ Q_{side} &= 2\pi\,\tau_{pin}\,\Omega\,R_p^2\,L_p \\ Q &= Q_{sh} + Q_{tip} + Q_{side}, \qquad \mathscr{T} = Q/\Omega \end{aligned} \tag{3.5}$$

where $\Omega = 2\pi\omega/60$ is the spindle speed in rad/s and $\mathscr{T}$ is the spindle torque in N m. At the design point Equation (3.5) gives 1755 W and 18.62 N m, of which the shoulder makes 68.8 %, the pin side 28.1 % and the pin tip 3.1 %, which is 877 J into every millimeter of joint at 120 mm/min. Because $\tau$ in (3.5) falls with temperature through (3.3), the response to spindle speed is strongly sublinear. Doubling the spindle from 450 to 900 rpm raises the steady heat generation only from 1405 to 1755 W, an increase of 24.9 %, while the flow stress of the shear layer falls from 103.5 to 59.4 MPa and the interface shear stress under the shoulder from 47.0 to 29.4 MPa.

### 3.3.4 Contact Pressure, Axial Force and the Stalled Axis

The axial force is an indentation problem, and an indentation is resisted by the metal a few millimeters out and a few millimeters down rather than by the thin shear layer at the interface. The load-bearing temperature is therefore taken mostly from the halo, as shown in Equation (3.6).

$$\begin{aligned} T_{bulk} &= w_b T_w + (1 - w_b) T_n \\ p_{sh} &= C_{ind}\, \sigma(T_{bulk}) \left( 1 + \kappa \frac{\max(0, z)}{z_0} \right) \\ p_{pin} &= C_{pin}\, \sigma(T_{bulk}) \\ F &= p_{sh} A_{sh} + p_{pin} A_{tip} \end{aligned} \tag{3.6}$$

where $w_b$ is the weight of the halo temperature in the load-bearing metal, $C_{ind}$ and $C_{pin}$ are indentation pressures as multiples of the flow stress, $\kappa$ is the rise in shoulder pressure per reference plunge depth $z_0$, and $F$ is the axial force. When $F$ exceeds the machine limit $F_{max}$ both pressures are scaled by $F_{max}/F$, the force is held at $F_{max}$ and the downward motion of the axis is blocked. Otherwise the axis is throttled by the remaining force margin through the stall factor of Equation (3.7).

$$s = \operatorname{clamp}\left( \frac{F_{max} - F}{F_{stall}}, 0, 1 \right) \tag{3.7}$$

Equations (3.6) and (3.7) are what make the plunge a different plant. Cold 6061 under a 12 mm shoulder needs more than the 20 kN the machine has, so the axis stalls against its limit and sinks only as fast as the metal softens. A plunge commanded at 120 mm/min, three seconds of travel, takes 12.2 s with the axial force pinned at 19.99 kN and a peak spindle torque of 35.9 N m. Through those twelve seconds an extra revolution per minute buys depth as well as heat, and that positive feedback is the source of the extra plunge gain of Section 3.4. Setting the adjustable weight $w_b$ to zero moves the load onto the shear layer and removes the force-limited plunge and the regime change with it.

### 3.3.5 The Four Thermal Nodes

Energy balances on the shear layer, the tool, the halo and the anvil interface are Equations (3.8) to (3.11).

$$C_n(a_f) \frac{\mathrm{d}T_n}{\mathrm{d}t} = Q - q_t - q_w - q_b \tag{3.8}$$

$$C_t \frac{\mathrm{d}T_t}{\mathrm{d}t} = q_t - U_{ta} (T_t - T_{amb}) \tag{3.9}$$

$$C_w \frac{\mathrm{d}T_w}{\mathrm{d}t} = q_w - U_{wf} (T_w - T_\infty) \tag{3.10}$$

$$C_b \frac{\mathrm{d}T_b}{\mathrm{d}t} = q_b - U_{ba} (T_b - T_{amb}) \tag{3.11}$$

where the three heat flows out of the shear layer are $q_t = U_{nt} a_f (T_n - T_t)$, $q_w = U_{nw} a_f (T_n - T_w)$ and $q_b = U_{nb}^{\mathrm{eff}} (T_n - T_b)$, the capacities are in J/K and the conductances in W/K. Both sides of each equation carry W. The shear-layer capacity grows with engagement, $C_n(a_f) = C_{n,\min} + C_{n,stir}\, a_f$, so

it is 0.70 J/K with the pin tip alone in contact and 3.64 J/K at full engagement. At the design point the 1755 W generated leaves as 69.6 % into the plate, 22.6 % into the anvil and 7.7 % up the tool.

The two conductances that are not constant carry the moving heat source and the clamp-up, as shown in Equation (3.12).

$$\begin{aligned} U_{wf} &= U_{wf,0}\, g_v \left(1 - d_e \mathrm{e}^{-\ell / L_e}\right), \\ g_v &= \mathrm{clamp}\left(1 + c_V \left(\frac{v}{v_{ref}} - 1\right), 0.3, 3\right) \\ U_{nb}^{\mathrm{eff}} &= U_{nb}\, a_d\, \mathrm{clamp}\left(1 - \frac{c}{1.5\,\mathrm{mm}}, 0, 1\right) \mathrm{clamp}(a_f, 0, 1.2) \end{aligned} \tag{3.12}$$

where $\ell = \max(0, L_{plate} - x_{start} - x)$ is the distance from the tool to the trailing edge of the plate, $c_V$ is the rise in plate conductance per unit of traverse speed, $d_e$ is the loss of conductance at the edge and $L_e$ is the distance over which the edge is felt. The first factor is the moving source. A faster tool spreads its heat over more cold metal per millimeter of weld, so going faster makes the weld colder, and the steady tool temperature falls from 490.5 °C at 60 mm/min to 446.9 °C at 240 mm/min at a fixed 900 rpm. The second factor is the run-out. Between the middle of the joint, 155 mm from the plate edge, and the last millimeter of the weld, 25 mm from it, the plate conductance falls from 3.31 to 1.70 W/K and the steady tool temperature rises from 470.8 to 502.4 °C with nothing on the machine having changed. Scaling the anvil path by $a_f$ matters more than it appears, because the anvil takes almost a quarter of the heat in a running weld and leaving that path at full strength during the plunge would make the plunge a much stiffer plant than it is.

### 3.3.6 Machine Axes, Wear and the Instrument Chain

The spindle and traverse drives are first order, and the axial axis is a first order servo with separate limits in each direction and the stall factor of (3.7) on the way down, as shown in Equation (3.13).

$$\begin{aligned} \tau_\omega \frac{\mathrm{d}\omega_A}{\mathrm{d}t} &= \omega_{cmd} - \omega_A, \qquad \tau_v \frac{\mathrm{d}v_A}{\mathrm{d}t} = v_{cmd} - v_A \\ \dot{h}^* &= (h_{cmd} - h_s)/\tau_z \\ \frac{\mathrm{d}h_s}{\mathrm{d}t} &= \begin{cases} -s\,\mathrm{clamp}\left(-\dot{h}^*, 0, \dot{z}_{max}\right), & \dot{h}^* < 0 \\ \mathrm{clamp}\left(\dot{h}^*, 0, \dot{z}_{ret}\right), & \dot{h}^* \geq 0 \end{cases} \\ \frac{\mathrm{d}x}{\mathrm{d}t} &= \begin{cases} v_A/60, & x < L_{weld} \\ 0, & \text{otherwise} \end{cases} \\ \frac{\mathrm{d}W}{\mathrm{d}t} &= k_W \frac{\omega}{\omega_{ref}} \frac{\max\left(0,\ T_n - 450\,^\circ\mathrm{C}\right)}{100\,\mathrm{K}} \end{aligned} \tag{3.13}$$

where $\dot{z}_{max}$ and $\dot{z}_{ret}$ are the commanded plunge and retract rates in mm/s and $k_W$ is the wear rate. Wear accrues only above 450 °C in the shear layer, so running hot is expensive twice, and at half life the steady tool temperature falls from 471.4 to 436.7 °C, two degrees above the cold edge of the window.

The thermocouple sits in a rotating tool a few millimeters behind the shoulder, so its reading reaches the controller through a transport delay and its own bead time constant, as shown in Equation (3.14).

$$\tau_{TC} \frac{\mathrm{d}T_m}{\mathrm{d}t} = T_t(t - \theta_{TC}) + n(t) - T_m \tag{3.14}$$

where $\theta_{TC}$ is the telemetry delay off the spindle, $\tau_{TC}$ is the bead time constant and $n$ is an optional noise sequence made by filtering a Gaussian with a 0.4 s time constant and rescaling it so the stationary standard deviation matches the selected value. The chain is stepped at each execution of the regulatory controllers, so the reading is held for 0.2 s between executions. The lag uses the exact

hold expression $T_{m,k} = \alpha T_{m,k-1} + (1-\alpha)(\cdot)$, in which $\alpha = \exp(-0.2\,\mathrm{s}/\tau_{TC})$, and the transport delay is a ring buffer of those samples read by linear interpolation, which applies the 0.60 s delay exactly on that grid. Together the two account for most of the apparent dead time in the identified models, and both sit inside the loop the reader tunes. The axial load cell is a first order lag of time constant $\tau_F$ on $F$ with no delay.

**Pause.** Try Exercise 3.2 before reading on.

### 3.3.7 Weld Quality

The quality window is a band in temperature whose edges move with traverse speed, and the defect classification is a set of thresholds on the tool temperature and the root clearance, as shown in Equation (3.15).

$$\begin{aligned} T_{lo}(v) &= T_{lo,0} + s_{lo}\left(v - v_{ref}\right) \\ T_{hi}(v) &= \min\left(T_{sol} - 15,\; T_{hi,0} + s_{hi}(v - v_{ref})\right) \end{aligned} \tag{3.15}$$

Faster travel needs hotter metal to keep the flow closing behind the pin, so $s_{lo}$ is four times $s_{hi}$ and the window narrows as the production rate rises, from 70 °C wide at 120 mm/min to 52 °C at 240. A millimeter of weld is scored as sound when $T_{lo} < T_t < T_{hi}$ and the root clearance $c$ lies between the anvil strike limit of 0.05 mm and the lack-of-penetration limit of 0.45 mm. Otherwise it is recorded as a wormhole, as flash, as a root flaw or as an anvil strike, accumulated by length along the joint and drawn on the plan view in the color of that defect. The record does not heal, because the metal that should have filled the void is already behind the tool.

### 3.3.8 Parameters, Initialization and Numerics

The constants are collected in Table 3.2. Every one of them is exposed in the model panel and is adjustable, apart from the bulk aluminum properties, the sharpness of the smooth minimum and the mid-weld start position, which are held back as reference values.

A cold scenario starts with the tool clear of the plate at $h_s = t + H_p + 0.15\,\mathrm{mm}$ and every node at plate temperature, and the machine program runs from the plunge. A mid-weld scenario starts at the quasi-steady condition 90 mm along the joint, which is obtained by solving the shear-layer energy residual rather than by marching to it. That residual is strictly decreasing in $T_n$, because heat generation falls with temperature through (3.3) while every loss term in (3.8) rises, so the root is unique and sixty bisections of the interval between ambient and the liquidus locate it to a nanokelvin from any starting point. A damped fixed point was tried first and it oscillated above about 1200 rpm, where the solidus collapse in (3.3) makes the map non-contractive.

The engine of Section 1.6.1 advances the states with fourth-order Runge-Kutta on a fixed 0.02 s step. The regulatory controllers execute every 0.2 s and the predictive controllers every 1.0 s, and ten samples per simulated second are logged. The exported file holds the true tool temperature beside the measured one, both identified models, the predictive controller's own bias, the defect record along the joint and the measured cost per control execution. The noise seed is 20260906.

### 3.3.9 Assumptions

Detail is spent where the physics forces a control decision and withheld everywhere else, and every simplification below fails somewhere.

*Four thermal nodes, not a moving-source field solution.* The travelling thermal field appears only through the traverse-dependent conductance in (3.12), calibrated so that the steady temperature falls the right amount with traverse speed. There is no isotherm map, no heat-affected-zone width and no advancing to retreating asymmetry.

*Heat generation is the flat-shoulder integral.* Equation (3.5) assumes a flat shoulder and a cylindrical pin sweeping a full circle at the local surface velocity. Real tools are threaded or scrolled, and the effective slip at the interface is the largest single uncertainty in every published thermal model of the process [5, 15]. No slip factor and no shear-layer thickness appear here.

**Table 3.2** Model Parameters. All Values Are Adjustable in the Application Except the Three Marked Hidden, Which Are Never Displayed.

| Symbol | Meaning | Value | Unit | Visibility |
|---|---|---|---|---|
| $R_s$, $R_p$, $H_p$ | shoulder radius, pin radius, pin length | 6.0, 2.0, 6.00 | mm | adjustable |
| $t$ | nominal plate thickness | 6.35 | mm | adjustable |
| $L_{plate}$, $L_{weld}$ | plate and programmed weld length | 300, 245 | mm | adjustable |
| $x_{start}$ | plunge from the leading edge | 30 | mm | adjustable |
| $\sigma_0$, $T_\sigma$, $\sigma_{res}$ | flow stress: factor, decay, residual | 1329, 158.8, 8 | MPa | adjustable |
| $T_{sol}$, $T_{liq}$ | solidus and liquidus | 582, 652 | °C | adjustable |
| $m$, $\omega_{ref}$ | rate exponent, reference spindle | 0.12, 800 | rpm | adjustable |
| $\mu_0$, $w_{slip}$ | sliding friction, shear loss at wear | 0.52, 0.30 | — | adjustable |
| $w_b$ | halo weight in the load-bearing metal | 0.80 | — | adjustable |
| $C_{ind}$, $\kappa$, $z_0$ | shoulder pressure, rise, reference depth | 0.265, 1.49, 0.15 | mm | adjustable |
| $C_{pin}$, $z_{eng}$ | pin pressure, full shoulder contact | 2.10, 0.080 | mm | adjustable |
| $U_{nt}$, $U_{ta}$ | shear layer to tool, tool to spindle | 3.00, 0.302 | W/K | adjustable |
| $U_{nw}$, $U_{wf,0}$ | shear layer to halo, halo to plate | 10.1, 3.37 | W/K | adjustable |
| $U_{nb}$, $U_{ba}$ | shear layer to anvil, anvil to table | 2.42, 1.30 | W/K | adjustable |
| $c_V$ | plate conductance per traverse speed | 0.66 | — | adjustable |
| $d_e$, $L_e$ | edge conductance loss and length scale | 0.78, 55 | mm | adjustable |
| $C_{n,\min}$, $C_{n,stir}$ | pin-tip capacity, added at contact | 0.70, 2.94 | J/K | adjustable |
| $C_t$, $C_w$, $C_b$ | tool, halo and anvil capacity | 27.6, 30, 69 | J/K | adjustable |
| $F_{max}$, $F_{stall}$ | machine force limit, stall margin | 20.0, 4.0 | kN | adjustable |
| $\tau_\omega$, $\tau_v$, $\tau_z$ | spindle, traverse and z-axis lags | 0.25, 0.40, 0.15 | s | adjustable |
| $\dot{z}_{max}$, $\dot{z}_{ret}$ | commanded plunge and retract rate | 120, 300 | mm/min | adjustable |
| $z_{target}$, $t_{dwell}$ | programmed plunge, dwell | 0.15, 12 | mm, s | adjustable |
| $k_W$ | wear rate above 450 °C | $6 \times 10^{-5}$ | 1/s | adjustable |
| $T_{lo,0}$, $T_{hi,0}$ | quality window at $v_{ref}$ | 435, 505 | °C | adjustable |
| $s_{lo}$, $s_{hi}$ | window edge slopes | 0.20, 0.05 | °C min/mm | adjustable |
| $v_{ref}$ | reference traverse speed | 120 | mm/min | adjustable |
| $\tau_{TC}$, $\theta_{TC}$ | thermocouple bead, telemetry delay | 1.20, 0.60 | s | adjustable |
| $\tau_F$ | axial load cell | 0.15 | s | adjustable |
| $\rho$, $c_p$, $k$ | aluminum properties | 2700, 896, 167 | SI | hidden |
| $n_{sm}$ | smooth-minimum order | 4 | — | hidden |
| $x_{run}$ | mid-weld start | 90 | mm | hidden |

*The flow stress is two parameters and a residual.* Equation (3.3) is not a Zener-Hollomon or Johnson-Cook law. There is no strain or strain-rate history state, and the entire rate dependence is the single exponent $m$ in (3.4). The fit is to hot-working data at one order of magnitude of strain rate [4], and it is extrapolated outside it.

*The load-sharing split is a modeling decision.* Taking the axial load from $T_{bulk}$ in (3.6) and the interface shear from $T_n$ in (3.4) is defensible, because indentation is resisted by bulk material and shear by the interface layer, but the weight $w_b$ is calibrated rather than measured, and it produces the plunge force spike and most of the extra plunge gain.

*The quality window is an empirical band with straight edges.* Real defect formation depends on material flow around the pin, thread pitch and direction, and tilt angle, none of which is represented [1]. The window is calibrated to temperature ranges reported for 6xxx alloys and is a teaching device rather than a process map a weld could be qualified against.

*Tool wear is one scalar.* It shrinks the shoulder and derates the interface shear stress, and it accrues linearly above a threshold. Real wear in aluminum is slow, geometry dependent and dominated by loss of the pin thread.

*The numerics are more repeatable than a machine.* One trajectory comes back on every computer, whereas the controller in a real machine carries jitter, a coarser measurement resolution and a scan order that moves.

**Table 3.3** Identified First-Order-Plus-Dead-Time Models. The Top Four Rows Are Each a 100 rpm Step Differenced Against a Held Run. The Last Two Are the Models the Controllers Carry, Which Are Baselined Differently. The Residual Is the Root Mean Square Fit Error in the Same Units as $K_p$.

| Identified at | $K_p$ (°C/rpm) | $\tau_p$ (s) | $\theta_p$ (s) | residual |
|---|---|---|---|---|
| Traverse, +100 rpm | 0.1425 | 12.24 | 2.31 | 0.0006 |
| Traverse, −100 rpm | 0.1569 | 12.55 | 2.34 | 0.0006 |
| Settled dwell, 700 rpm | 0.1782 | 15.81 | 2.45 | 0.0008 |
| Plunge, shoulder contact | 0.2401 | 1.17 | 0.00 | 0.0179 |
| Traverse model in the controller | 0.1356 | 11.78 | 2.68 | — |
| Plunge model in the controller | 0.2552 | 1.92 | 2.10 | — |

## 3.4 Three Plants in One Weld

Identification is the whole subject of the application, so the experiments are described as experiments rather than as linearizations. Each one is an open-loop step test in spindle speed on the shipped nonlinear model, made in Manual, differenced against a second run held at the base speed so that the slow baseline drift cancels, normalized by the 100 rpm step size and fitted by least squares to the first-order-plus-dead-time form of Equation (3.16). The traverse steps are taken on the long plate of scenario 4 switched to Manual, stepped at 20 s and fitted over 300 s, which reproduces the step of scenario 3 to within 0.002 °C. The dwell step goes from 700 to 800 rpm after 120 s of the 240 s dwell of scenario 3b and is fitted to the end of the dwell. The plunge experiment compares two scenario 2 plunges at 900 and 1000 rpm, the faster speed set before the run so that its drive turns at 1000 rpm from the start, over the window from shoulder contact, 8.56 s into the slower run, to the start of the faster run's traverse 14.8 s later. The thermocouple reading is sampled every 0.04 s, and time is counted from the controller execution at which the new speed reached the drive.

$$\frac{\Delta T_m(s)}{\Delta \omega(s)} = \frac{K_p\, \mathrm{e}^{-\theta_p s}}{\tau_p s + 1} \tag{3.16}$$

The fit minimizes the sum of squared residuals over $K_p$, $\tau_p$ and $\theta_p$ jointly, with $K_p$ solved analytically inside a search over the other two. Results for four operating points are collected in Table 3.3 and the two extreme responses are plotted in Figure 3.2.

Three results in Table 3.3 matter for the controller. The first is the size of the change between the plunge and the traverse. The gain rises from 0.1425 to 0.2401 °C/rpm, a factor of 1.68, and the time constant falls from 12.24 to 1.17 s, a factor of 10.5, between two phases of one weld. The gain change comes from the stalled axis of (3.7), where an extra revolution per minute buys depth and therefore contact area, and the time-constant change from the capacity term $C_n(a_f)$ in (3.8), a fifth of its fully engaged value while the pin is still entering.

The second is the asymmetry within one operating point. Stepping the same condition down instead of up returns 0.1569 against 0.1425 °C/rpm, ten percent more gain for a step of the same size in the other direction, which is the plant reporting that it is not linear in either direction. The settled dwell at 700 rpm is a third point again, 0.1782 °C/rpm with a 15.81 s time constant, higher gain and a slower response than the weld because with no traverse to carry heat into cold metal the sink conductance in (3.12) is smaller and every watt is worth more degrees. That third measurement is worth having because it shows the fast 1.17 s time constant belongs to the plunge transient rather than to a stationary tool.

The third is the residual. The three fits taken at a settled operating point have residuals of 0.0006 to 0.0008 °C/rpm, and the plunge fit has 0.0179, twenty to thirty times larger. Its dead time is held at zero because the unconstrained optimum, −0.11 s from shoulder contact, is not physical. Figure 3.2 shows why. The differenced plunge response is not a monotone first-order rise. It climbs to 0.2610 °C/rpm

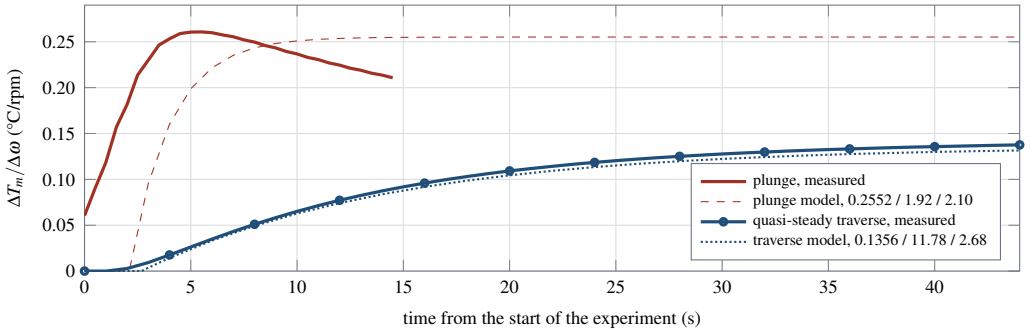



**Figure 3.2** The two step tests behind the predictive controllers. Each curve is the difference between a run stepped by one hundred revolutions per minute and a run held at the base speed, divided by the step size. The traverse experiment is timed from the step and the plunge experiment from shoulder contact. The measured plunge response is the thick solid line and the measured traverse response the thick line with circles. The two models whose step responses fill the dynamic matrices are the thin dashed line (plunge) and the dotted line (traverse).

5.3 s after the shoulder engages and then decays to 0.2098 by the end of the dwell as the halo warms and takes the extra pressure back out of (3.6). Fitting a first-order model to that shape is what a commissioning engineer would do with the trend in front of them, and it means the plunge-identified controller is handicapped by model structure as well as by operating point.

The two models the predictive controllers actually carry are the last two rows of Table 3.3, and the two rows are not baselined the same way. The traverse model comes from the same experiment fitted *without* the held-run difference, over a window long enough that the held run's own drift of –0.53 °C per 300 s biases the gain low and the apparent dead time high, which returns 0.1382 °C/rpm, 11.10 s and 2.60 s, within two percent of the stored gain. The plunge model cannot be read that way at all, because a plunge from a cold plate is a large transient with or without the step, so its stored gain of 0.2552 °C/rpm matches the peak of the differenced response in Figure 3.2, 0.2610, to within three percent rather than the least-squares asymptote of 0.2401. Both sets are defensible readings of the same trends and they differ by five percent on the traverse gain and six percent on the plunge gain, which changes nothing in the argument because the ratio between the two regimes is 1.9 on the stored constants and 1.7 on the differenced fits. The stored values are quoted wherever a controller's internal model is meant and the differenced values wherever the plant itself is meant.

## 3.5 Controllers

Tool temperature is the controlled variable and spindle speed is the manipulated variable in all five architectures, so that the comparison is between designs rather than between actuators. Traverse speed belongs to the operator, because it is also the production rate, and the axial axis is a second loop of its own. A friction stir weld is a sequence rather than an operating point, so the temperature program holds a soak setpoint of 330 °C through the plunge and the dwell and hands over to the welding setpoint of 470 °C when the tool starts moving. Without that program the plunge would be twelve seconds of saturated actuator and the two identified models could not be compared in the regime where they disagree.

🕮 **Pause.** Try Exercise 3.1 before reading on.

### 3.5.1 The Two PI Loops

Both PID controllers are the velocity-form algorithm of Section 1.6.2, with the derivative taken on the filtered measurement and anti-windup by back-calculation on the integral term alone. Equation (3.17)

writes it in the symbols of this loop, where $s_k$ is the running sum $S_k$ of Equation (1.15) and $c_k$ is the output before limiting, $v_k$ of Equation (1.16).

$$\begin{aligned} e_k &= T_{\mathrm{SP},k} - T_{m,k}, \qquad P_k = K_c e_k \\ D_k &= -K_c \tau_D \left. \frac{\mathrm{d}T_m^f}{\mathrm{d}t} \right|_k \\ s_k &= s_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \\ c_k &= s_k + I_{k-1}, \qquad u_k = \mathrm{clamp}(c_k,\ 200,\ 1500) \\ I_k &= I_{k-1} + \left[ \frac{K_c}{\tau_I} e_k + \frac{u_k - c_k}{T_t} \right] \Delta t_c \end{aligned} \tag{3.17}$$

where $u$ is spindle speed in rpm, $\Delta t_c = 0.2\,\mathrm{s}$ is the controller period, $c_k$ is the speed the algorithm demands, $I$ is the integral state and $T_t = \tau_I$ is the tracking time of the back-calculation, so a clipped command sheds $\Delta t_c / \tau_I$ of its excess at each execution. There is no bias term and no slew limit, and the clamp acts on the emitted command only. Both loops are reverse acting in the instrument sense, with the sign convention set by a direction flag of $+1$, since a tool below its setpoint calls for a faster spindle on a plant of positive gain. The derivative estimate uses a first-order filter of 2.0 s on the measurement and both defaults set $\tau_D = 0$, so derivative action is available but off. Under manual operation $I_k$ is instead back-solved as $u_k - s_k$, which pins the output to the operator's spindle speed and removes the need for any separate transfer logic in either direction. The loop that is not driving the spindle is fed the applied speed the same way every execution, so switching architecture mid-weld costs nothing. A reset is different. It seeds both loops with $s_0 = 0$ and $I_0 = 900\,\mathrm{rpm}$, the manual station value, so a weld started already in Auto from a cold plate adds the whole proportional term $K_c e_0$ at the first execution, with $e_0 = 308\,^{\circ}\mathrm{C}$ against the soak setpoint, and the command saturates at 1500 rpm. A loop that enters Auto from Manual has been tracking the station and takes over from 900 rpm without a bump, provided the switch is made after the run has started, since a switch made before the run is part of the starting conditions and is a start in Auto. Section 3.6 measures what that difference costs.

The two tunings are the ITAE setpoint and ITAE disturbance correlations for a first-order-plus-dead-time plant [16], applied to the stored traverse model of Table 3.3, as shown in Equations (3.18) and (3.19).

$$K_c K_p = 0.586 \left( \frac{\theta_p}{\tau_p} \right)^{-0.916}, \quad \tau_I = \frac{\tau_p}{1.03 - 0.165\,\theta_p/\tau_p} \tag{3.18}$$

$$K_c K_p = 0.859 \left( \frac{\theta_p}{\tau_p} \right)^{-0.977}, \quad \tau_I = \frac{\tau_p}{1.484} \left( \frac{\theta_p}{\tau_p} \right)^{0.680} \tag{3.19}$$

With $K_p = 0.1356\,^{\circ}\mathrm{C/rpm}$, $\tau_p = 11.78\,\mathrm{s}$ and $\theta_p = 2.68\,\mathrm{s}$, so that $\theta_p/\tau_p = 0.2275$, Equation (3.18) returns $K_c = 16.77\,\mathrm{rpm}/^{\circ}\mathrm{C}$ and $\tau_I = 11.87\,\mathrm{s}$, and Equation (3.19) returns $K_c = 26.91\,\mathrm{rpm}/^{\circ}\mathrm{C}$ and $\tau_I = 2.90\,\mathrm{s}$. The application ships those rounded to 16.8 with 11.9 s and to 26.9 with 2.9 s. The disturbance tuning carries sixty percent more proportional gain and four times the reset action, on the same plant and from the same step test, which is the whole of the difference between the two entries.

The axial force loop is a PI controller of the same form whose output is the commanded shoulder plunge depth, with $K_c = 0.012\,\mathrm{mm/kN}$, $\tau_I = 1.6\,\mathrm{s}$ and an output range of 0 to 0.45 mm. It is reverse acting for the same reason the temperature loop is, because more depth means more force. In position mode the depth command is the programmed value referenced to the machine table and the nominal plate thickness, which is what a position-controlled machine actually holds, and the force controller is tracked so that either mode can be selected without a bump.

### 3.5.2 Step-Response Predictive Control

Both predictive entries are dynamic matrix control in its original step-response form [7, 17], with the gain matrix formed once and reused. The two entries share every line of code, every weight and the same horizon. Only the three constants of the step response differ. Section 1.4.2 builds the same predictor from a step response, and the gain matrix of Equation (3.21) below is Equation (1.5) with $A$ in place of $G$, $\lambda$ in place of $R$ and the columns of $A$ set by move blocking.

Three quantities are built when the controller is created. Because the identified model is (3.16), its unit step response is written down rather than integrated, as shown in Equation (3.20).

$$a_k = \begin{cases} K_p\left(1 - \mathrm{e}^{-(kt_m - \theta_p)/\tau_p}\right), & kt_m > \theta_p \\ 0, & \text{otherwise} \end{cases} \qquad a_0 = 0 \tag{3.20}$$

where $t_m = 1.0\,\mathrm{s}$ is the execution period and $k$ runs from 0 to $P+1$. The prediction horizon is $P = 40$ points, which is 40 s, and the move horizon is $M = 3$ after blocking the moves at samples $n_j \in \{0, 2, 6\}$, so three decisions cover forty samples [7]. The dynamic matrix and the constant gain matrix of Equation (3.21) follow.

$$A_{ij} = \begin{cases} a_{(i+1)-n_j}, & i+1 > n_j \\ 0, & \text{otherwise} \end{cases}$$
$$K = \left(A^{\mathsf{T}} Q A + \lambda I\right)^{-1} A^{\mathsf{T}} Q \tag{3.21}$$

where $A$ is $40 \times 3$, the diagonal of $Q$ carries unit weight over the first thirty points and a terminal weight $w_T = 3.0$ over the last ten, and $K$ is $3 \times 40$. Inverting the $3 \times 3$ takes a hand-coded Gauss-Jordan pass and is the only linear algebra anywhere in the application. One revolution per minute is worth a seventh to a quarter of a degree, so the diagonal of $A^{\mathsf{T}} Q A$ runs from 0.58 to 0.71 for the traverse model and from 3.2 to 3.6 for the plunge model, five times larger. The move suppression is therefore scaled against the tracking term, as shown in Equation (3.22).

$$\lambda = \lambda_{rel} \frac{1}{M} \operatorname{tr}\left(A^{\mathsf{T}} Q A\right) \tag{3.22}$$

so that one slider position means the same aggressiveness for both models, which is what makes the comparison between them fair. The default $\lambda_{rel} = 0.08$ evaluates to $\lambda = 0.0523$ on the traverse model and $\lambda = 0.2754$ on the plunge model. Equation (3.21) is formed again only if $\lambda_{rel}$ or $w_T$ is moved, and Equation (3.20) is never recomputed.

Each execution begins by charging the prediction with the speed change the spindle received rather than the change that was requested, as shown in Equation (3.23).

$$Y_i \leftarrow Y_i + a_i\left(\omega_{cmd} - \omega_{cmd}^{prev}\right), \qquad i = 0 \ldots P \tag{3.23}$$

A move cut short by the clamp, one the rate limit has not finished delivering, or one that never occurred at all because another architecture held the spindle, enters Equation (3.23) at whatever size it really had, and nothing is ever charged in advance of being delivered. That single substitution supplies all of the anti-windup this controller has, and it also keeps the inactive model anchored to the plant, so either model can be selected in mid-weld without a bump. The prediction is then shifted one sample and its tail held at the last value, $Y_i \leftarrow Y_{i+1}$ for $i < P$, which is the correct extension for a self-regulating output and would be wrong for an integrating one. The plant correction and the reference trajectory are Equation (3.24).

$$\begin{aligned} b_k &= T_m - Y_0, & Y_i &\leftarrow Y_i + b_k \\ r_0 &= T_m, & r_{i+1} &= T_{\mathrm{SP}} + (r_i - T_{\mathrm{SP}})\,\mathrm{e}^{-t_m/\tau_{ref}} \\ e_i &= r_{i+1} - Y_{i+1}, & i &= 0 \ldots P-1 \end{aligned} \tag{3.24}$$

where $b_k$ is a constant output-disturbance estimate added across the whole horizon and $\tau_{ref} = 2.0\,\mathrm{s}$ shapes the approach to a new setpoint. The constant $b_k$ is what removes steady offset, as in Section 1.4.5, and it does for this controller what reset action does for the PI loops. On an integrating output the same constant bias leaves an offset, which Chapter 2 measures on a balloon's altitude. It is also the reason a controller carrying a gain error of nearly a factor of two produces a slowly corrected offset instead of ringing, and that robustness is measured in Section 3.6.

What remains is one matrix-vector product, $\Delta u = Ke$. The first blocked move is rate limited to $\Delta u_{max} = 250\,\mathrm{rpm}$ per execution, clamped to the drive envelope of 200 to 1500 rpm and applied, which projects the unconstrained move onto the limits rather than solving the constrained program of Section 1.4.3. Nothing is written back into $Y$ because the next execution reads the delivered change through (3.23). Both models execute on every cycle whatever architecture is selected. Timed over 100 000 executions on an Apple M-series laptop the online path costs 0.272 µs per execution for the traverse model and 0.274 µs for the plunge model. The application repeats that measurement at every reset and reports it on the scorecard, so the figure the reader sees is measured on the machine in front of them.

## 3.6 Results

Thirteen scenarios form a laboratory sequence, from a weld made by hand in scenario 1 to the graded challenge, with a second step test held in the dwell numbered 3b. Each of them presets the starting condition, the architecture, the three tunings, the disturbances and whatever timed events it needs, asks for a prediction before anything runs, and carries its own completion rule, as Section 1.6.4 describes. The challenge is judged on the case as given, so a run with a changed model parameter or disturbance does not count. Reset restarts a case from $t = 0$ and keeps every setting the reader changed, while reloading the case restores its shipped settings, so a setting chosen before the run and the same setting carried through Reset give one run.

Scenario 1 sets the problem before a controller is involved. Held at a fixed 900 rpm for the whole job the weld comes out 89.8 % sound with 24.9 mm of wormhole, all of it in the first 25 mm of the joint, because the tool leaves the dwell at 372 °C, 63 °C below the cold edge of the window, and takes 12.9 s of traverse to climb into it. A spindle speed that suits the plunge does not suit the weld, and the two identified models of Table 3.3 say by how much.

> **TRY IT IN THE APP** **1 · Weld It by Hand (MANUAL)**
>
> Load the scenario and press Run without touching the spindle. Watch the weld track drawn behind the tool. The tool leaves the dwell at 372 °C, below the window, and needs 12.9 s of traverse to climb into it, so all 24.9 mm of wormhole lies in the first 25 mm of the joint.

### 3.6.1 Five Architectures Over a Complete Weld

Table 3.4 runs all five architectures over the identical job, a cold plate, a plunge, a 12 s dwell, 245 mm of traverse at 120 mm/min and the run-out at the edge, and scores each phase separately. In the top five rows each architecture is selected and switched to Auto from the 900 rpm manual station of scenario 1 at the first controller execution after the run starts, 0.2 s in. A switch made before the run is part of the starting conditions and gives the start in Auto of the last two rows. The soak measure is integrated absolute error between the tool and the soak setpoint over the settled part of the dwell, from eight seconds after the axis reaches depth to the start of the traverse, which is a window of about four seconds. The traverse and run-out measures are integrated absolute error over those phases, 108 and 15 s long. The defect lengths and the sound fraction are read at the end of the weld and need no window.

The servo and regulator tunings change places between two phases of one weld, on the same loop and the same plant, when each enters Auto from the manual station. The plunge and the dwell are

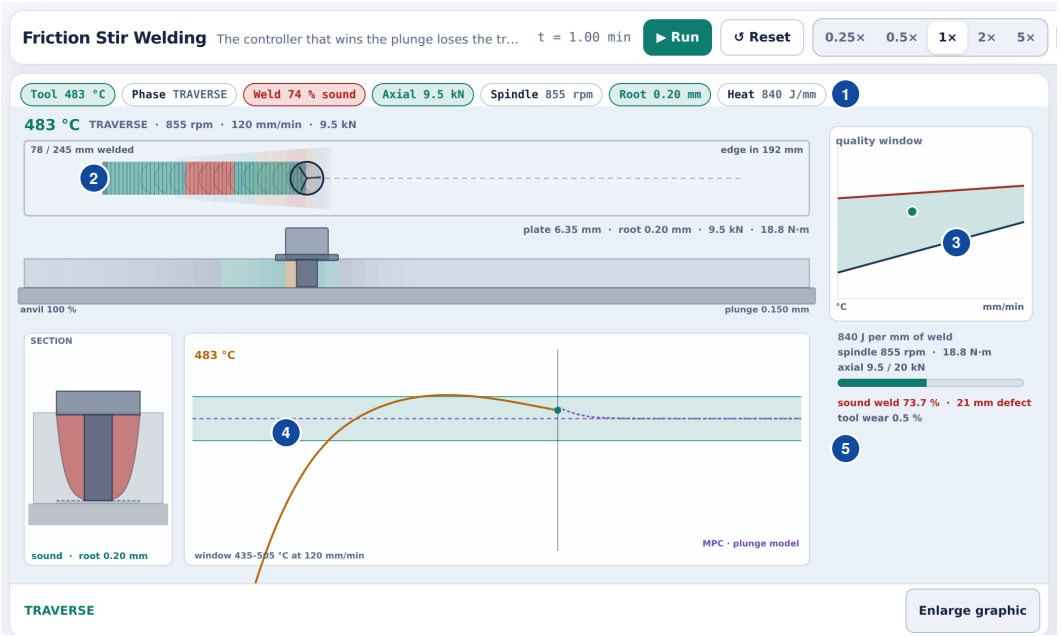



**Figure 3.3** Scenario 9, the plunge-identified controller meeting the traverse, read 60 s into the run with the weld still being made, cropped to the header and the plant view. The tool stands at 483 °C against a setpoint of 470 and a window of 435 to 505 °C, and the controller is pulling the spindle back to 855 rpm, below the 891 rpm that holds 470 °C in quasi-steady welding. (1) The badges carry the tool temperature, the spindle speed, the root at 0.20 mm and a heat input of 840 J/mm. (2) The plan view records the consequence, 78 of 245 mm welded, with the flash drawn where the tool ran hot and the wormhole made as the traverse began, and (5) the readout under the inset gives 73.7 % of the weld sound and 21 mm of defect, of which the run's record puts 18 mm in flash and 2.5 mm in wormhole. (3) The inset places the operating point inside the quality window, and (4) the temperature plot shades the window behind the tool temperature. The scorecard and the trend strips, outside the crop, record a peak of 507 °C, a command that has spent 38.3 % of the run against a limit and accumulated 1316 rpm of travel, and an axial force of 9.55 kN against a 9.50 setpoint.

**Table 3.4** Per-Phase Comparison Over One Complete Weld From a Cold Plate. In the Top Five Rows Each Architecture Enters Auto From the 900 rpm Manual Station Just After the Run Starts, and Each Column Has a Different Winner, Shown in Bold. The Last Two Rows Start the PI Loops Already in Auto, as Scenario 5 Ships.

| Arch. | Soak IAE (°C s) | Into window (s) | Trav. IAE (°C s) | Run-out IAE (°C s) | Cold, hot (mm) | Sound weld (%) |
|---|---|---|---|---|---|---|
| Manual, 900 rpm | 82 | 12.9 | 1453 | 186 | 24.9, 0 | 89.8 |
| PI, servo tuned | **18** | 8.1 | 817 | 34 | 15.4, 0 | 93.7 |
| PI, regulator tuned | 213 | 4.2 | **777** | **17** | 7.7, 0 | 96.9 |
| DMC, traverse model | 232 | 2.5 | 974 | 59 | 4.1, 0 | **98.3** |
| DMC, plunge model | 239 | **0.0** | 1311 | 74 | **2.5**, 18.2 | 91.6 |
| PI servo, Auto start | 217 | 3.5 | 637 | 34 | 6.3, 0 | 97.4 |
| PI reg., Auto start | 212 | 4.2 | 773 | 17 | 7.6, 0 | 96.9 |

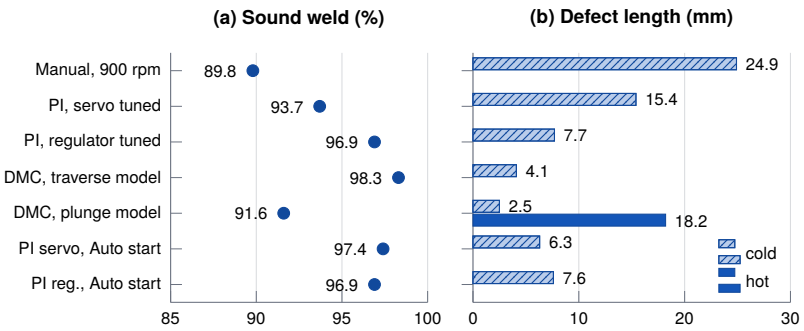



**Figure 3.4** The weld-level columns of Table 3.4: (a) the sound fraction of the weld, as dots on an axis that starts at 85 %, and (b) the cold defect (hatched bars) and the hot defect (dark bars) of each architecture, in the table's row order. Notice that the plunge-model DMC leaves the shortest cold defect and the only hot one.

a setpoint change, and the ITAE setpoint tuning holds the soak to 18 °C s against the disturbance tuning's 213, a factor of twelve. The start of the traverse is a load change, because the tool moves into plate that has no thermal field in it, and there the disturbance tuning reaches the window in 4.2 s against 8.1, leaving 7.7 mm of cold defect against 15.4. A 0.08 mm thin patch in quasi-steady welding is a load change of the same kind. Set 15 s into scenario 4, it costs the regulator tuning 109 °C s of error over the following sixty seconds with a peak deviation of 6.1 °C, against the servo tuning's 160 and 8.3 °C. A setpoint step is not a load change, and stepping the setpoint from 470 to 500 °C at the same instant overshoots by 10.9 °C with the regulator tuning against 1.5 °C with the servo tuning. The disturbance tuning also pays in damping. After the step into the traverse it rings with a 16.7 s period, each swing about two thirds of the one before, from +17.2 °C at 34 s to ±2.7 °C by 109 s, inside the window throughout, while the servo tuning passes its setpoint once, by 1.3 °C. There is no single best tuning. If one PI loop has to be shipped, the one that is shipped is the one that is wrong in the phase that costs less, and that is a production decision rather than a control one. Figure 3.4 draws the sound fraction and the defect lengths of all seven rows.

How a loop enters Auto matters as much as how it is tuned. Scenario 5 starts the same weld already in Auto, so the first execution adds the whole proportional term on a 308 °C soak error to the 900 rpm station, as described with Equation (3.17), and both loops hold the spindle at 1500 rpm until about 16 s, while the lagging thermocouple catches up. Neither then holds the soak, at 217 and 212 °C s in the last two rows of Table 3.4. The tool leaves the dwell at 400 °C under servo tuning and 393 °C under regulator tuning, so the servo tuning now reaches the window first, in 3.5 s against 4.2, with 6.3 mm of cold defect against 7.6 and a weld 97.4 % sound against 96.9 %. That ordering comes from the hotter start rather than from better load rejection.

✦ DID YOU KNOW? **A Bad Start That Made a Better Weld**

From the simulator: started already in Auto, the servo-tuned PI loop saturates and spoils the soak, 217 °C s of error against 18 when it enters Auto after the run begins. Yet it leaves the sounder weld, 97.4 % against 93.7 % (Table 3.4).

### 3.6.2 The Controller That Wins the Plunge Loses the Traverse

The two DMC rows of Table 3.4 are the result the application exists for. The two controllers are the same code with three different numbers in Equation (3.20), and scenario 9 as shipped, which starts the plunge model already in Auto, reproduces its row to within 1.4 °C s and 0.2 mm.

The plunge-identified controller wins the plunge outright. It puts the tool inside the quality window at the very first millimeter of the traverse, which is 0.0 s against 2.5 for the traverse-identified controller, 4.2 for the regulator PI and 8.1 for the servo PI, and it leaves 2.5 mm of cold defect at the start against 4.1 and 15.4. For the first thirty millimeters of joint it is the best controller in the application.

It then loses the traverse, and it loses it as flash. Its internal model says the temperature will arrive where it was asked for in two seconds while the plant takes twelve, so it keeps adding spindle speed while very little appears to be happening on the thermocouple. Its integrated error through the traverse is 1311 °C s against 974, its peak tool temperature is 508 °C against 496, and it is the only one of the five architectures that makes a hot defect at all, 18.2 mm of it, laid down between 33 and 51 mm along the joint. That is not a metric that degraded. It is metal extruded out from under the shoulder and lost from the joint, visible on the plate and on the plan view of the animation in the color of the defect, as shown in Figure 3.3 sixty seconds into the run, and the weld finishes 91.6 % sound against 98.3 % for the identical controller carrying the other model.

The failure is a slowly corrected offset rather than an oscillation. Sampling the bias term of (3.24) at $t = 30$ s, nine seconds into the traverse, and at $t = 80$ s, the plunge model reads 5.54 °C and then −1.23 °C while the traverse model reads 4.91 and 0.15 °C. A constant output-disturbance estimate absorbs a gain error of 1.7 to 1.9 as an offset that decays over about a minute, which is the underrated robustness property of this controller form and also the reason a badly identified predictive controller can stay in service for years without drawing attention [8]. Every column of the top five rows of Table 3.4 has a different winner, and that is the argument for gain scheduling, for identification over several operating points, and for a nonlinear predictive formulation [6, 14].

🕮 **Pause.** Try Exercise 3.3 before reading on.

On the setpoint step of scenario 8, 470 to 500 °C at 15 s, the traverse-model controller is slower than the servo PI given the same step in scenario 4, and gentler. It overshoots by 0.8 °C against 1.5 °C and needs 21.2 s to stay inside ±1.5 °C against 8.5 s, and that ordering is not a fault. The servo PI answers the 30 °C step by throwing the spindle from 901 to 1458 rpm and pulling it back to 1122 rpm, while the traverse-model controller lifts it only to 1206 rpm and settles at 1123 rpm. On a machine where every excess revolution is tool wear above 450 °C, the distinction is worth having. The plunge-model controller overshoots the same step by 6.0 °C and takes 58 s to settle. The ordering turns against both predictive controllers on the thin patch, where they accumulate 315 and 410 °C s against the regulator PI's 109, because a reference trajectory shapes the approach to a setpoint and does nothing for a disturbance the model never saw.

### 3.6.3 Position, Force and the Root

Scenario 7 removes 0.10 mm of plate thickness with the temperature loop switched off, so that the disturbance can be seen directly. Under position control the axis holds its programmed height above the table and never learns that the plate got thinner. The shoulder plunge falls from 0.15 to 0.05 mm, the engaged shoulder area from 100 to 62 %, the heat generation from 1755 to 1470 W and the steady tool temperature from 471.4 to 414.3 °C, which is twenty-one degrees below the cold edge of the

window and a continuous wormhole. In the run itself the tool reaches 422.7 °C at 56 s before the approaching plate edge warms it again.

Force control holds the heat instead. The loop sinks the tool 0.098 mm to restore the contact pressure, so the steady heat generation is 1761 W and the tool sits at 470.2 °C, and in the run the tool dips only to 466.6 °C and is back at 471 °C by 40 s. What it spends is the root, and there is only 0.20 mm of clearance between the pin and the backing anvil. The clearance falls to 0.086 mm thirty seconds after the patch arrives and to 0.011 mm by sixty seconds. It passes the 0.05 mm anvil strike limit at 60.6 s and reaches zero at 74.3 s, with the tool at 491 °C, because the tool is entering the run-out where the metal is softer and more depth is needed to hold the force setpoint. Neither mode is safe alone, which is why production machines interlock one against the other [10, 11].

### 3.6.4 Disturbances Nobody Measures, and the Run-Out

A plate that lifts off its backing bar removes most of a path that carries almost a quarter of the heat leaving the shear layer. At an anvil contact of 0.30 the steady tool temperature rises from 471.4 to 485.7 °C at mid-plate, and at 0.15 it reaches 492.9 °C. Nothing on the machine measures it. At a contact of 0.30 the spindle speed that holds 470 °C is 88 rpm lower. The regulator PI gets there through a lightly damped swing, from 891 rpm at the event in scenario 10 to a trough of 616 rpm 39 s later and 803 rpm at 60 s, and the tool never passes 474.7 °C, against 507.4 °C and 5.4 mm of flash open loop at 900 rpm with the run-out on top of it. A temperature loop turns an unmeasured heat-sink failure into a visible actuator movement, which is an argument for closing the loop on a process that runs perfectly well open loop most of the time.

The run-out is the opposite kind of disturbance, because it is entirely predictable from a position the machine already knows. Every closed-loop architecture handles it. Over the complete weld of Table 3.4 the run-out costs 17 °C s under the regulator PI, 34 under the servo PI, 59 under the traverse-identified controller and 74 under the plunge-identified one, the reverse of the soak ordering for the two PI loops. Started at 170 mm in scenario 11, each loop begins in Auto with the tool 6 °C above the setpoint, and the regulator PI is still swinging by 4 °C from that start when the run-out begins at 22.6 s. There the same four score 45, 32, 53 and 49 °C s, and the servo PI is the best of them. Open loop at a fixed 900 rpm the tool reaches 487.6 °C at the last millimeter, which is still inside the window and leaves no margin for anything else, and adding the loose clamp to it puts flash into the end of the joint. Closed loop the margin can be spent instead of lost, and a scenario 11 run under the traverse-identified controller with its setpoint raised to 500 °C has the run-out carry the tool to 506.4 °C against a window top of 505, leaving 8.5 mm of flash over the last 8.5 mm of an otherwise sound joint.

### 3.6.5 The Graded Challenge

The challenge is a complete weld at 190 mm/min with a 0.09 mm thin patch arriving at 50 s, handed over with a servo loop detuned to $K_c = 2\,\text{rpm}/^\circ\text{C}$ and $\tau_I = 60\,\text{s}$. As handed it finishes 46.1 % sound with 132.1 mm of wormhole and fails by 48.9 points and 117.1 mm. The target is 95 % sound, under 15 mm of defect, and the pin still clear of the anvil.

The answer has three parts and only one of them is tuning. The traverse speed comes down, because at 190 mm/min the cold edge of the window has risen to 449 °C while the tool at a fixed spindle speed has fallen to 455.6 °C, which is 6.6 °C of margin against 31.8 °C at 130 mm/min, and the run-out still has to be survived at the top of the window. The temperature goes on the traverse-identified controller. With the axis in position it finishes 97.7 % sound with 5.6 mm of defect and a 0.200 mm root at 130 mm/min, passing by 2.7 points, 9.4 mm and 0.150 mm whether that configuration is revealed before the run or carried through Reset, while the regulator PI with its ITAE tuning at that speed reaches 96.4 % and 8.8 mm and also passes.

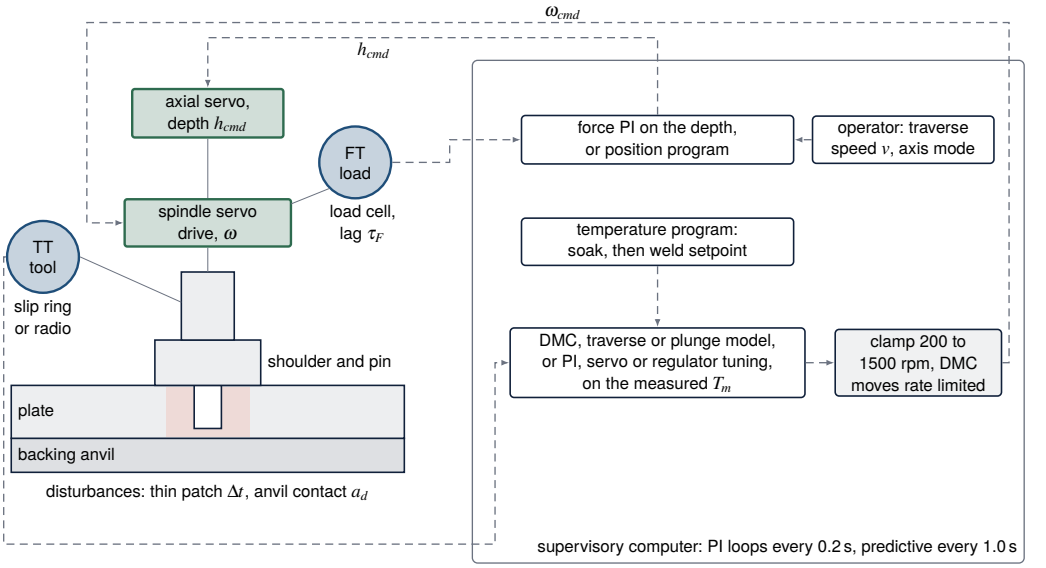



**Figure 3.5** The chapter's loops on welding hardware. TT, a thermocouple in the rotating tool, feeds whichever temperature controller is selected, and its command sets the speed of the spindle servo drive. FT, the axial load cell, feeds the force loop, which moves the depth command of the axial servo, or the axis holds a programmed position. Notice that the thin patch and the anvil contact reach the controllers only through the tool temperature and the force.

> **TRY IT IN THE APP** **Challenge · A Complete Weld**
>
> Load the challenge and press Run as it is handed over. The detuned servo loop at 190 mm/min finishes 46.1 % sound with 132.1 mm of wormhole. Press Reset, apply the suggested configuration, which puts the traverse-identified predictive controller on the spindle at 130 mm/min with the axis in position, and run again: 97.7 % sound with 5.6 mm of defect.

The axial axis is the interesting choice, and the safe answer is to leave it in position. Force control is the obvious move, because it removes the thin patch as a heat disturbance entirely, and it finishes 96.2 % sound with 9.4 mm of defect, 5.7 mm of it at the end of the joint where the pin is inside the anvil strike limit. By the run-out, where the metal is softest and the force loop has to sink furthest to hold its setpoint, the root clearance has fallen to 0.040 mm against an anvil strike limit of 0.050. Force control fails the challenge on the root with a weld that would have passed every surface inspection.

## 3.7 Technology in Practice

### What Measures It

Weld temperature is read in three ways. Thermocouples set in the workpiece give no signal a controller can use during the weld and cannot be reused. Thermocouples set in the tool can. The tool-workpiece thermocouple uses the voltage that arises between the steel tool and the aluminum workpiece and averages the temperature over the whole contact. Pyrometers and thermal cameras are troubled by the reflective aluminum and see only the edge of the shoulder [18]. After the weld, phased-array ultrasonics and eddy-current arrays look for the flaws that the chapter scores, wormholes and lack of penetration among them, and film radiography is kept for wormholes and larger flaws [3]. The chapter's TC is a thermocouple in the tool, and its defect record stands in for that inspection.

> **TECHNOLOGY** **Sensor: A Thermocouple in the Rotating Tool**
>
> The usual sensor is a thermocouple set in a hole drilled in the tool, in the shoulder or in the center of the pin, as close to the tool-workpiece interface as it can go so that it answers quickly. A slip ring or a wireless link carries its signal off the spinning tool. The steel around the junction and its ceramic coating make it read low: in one comparison a shoulder thermocouple read 125 °C and a pin thermocouple 70 °C below a reading averaged over the whole tool contact. The hole makes the tool costly, a small tool may have no room for it, and a junction that close to the interface can break during a weld [18]. The chapter's TC carries a 0.60 s telemetry delay and a 1.20 s bead lag.

### What Moves It

The spindle sets the heat, and the axial axis sets how deep the shoulder rides. Depth was once held by roller assemblies or by actuators fed from external depth sensors, whose reading is only an indirect sign of depth. Feedback on the penetration force has proved a robust and reliable way: on a tapered panel, a force that varied by 890 N kept the depth to a variation of 0.1 mm [19]. On a compliant robot under force control, metal that grows too hot can collapse under the tool, a meltdown that a rigid position-controlled machine does not risk [20]. The chapter's axial axis holds a position or, through its PI loop, a force read by a load cell with a 0.15 s lag.

> **TECHNOLOGY** **Final Element: The Spindle Servo Drive**
>
> The tool turns on a spindle driven by an electric servo motor through a gearbox, and the servo drive sits on a fieldbus that carries its speed setpoint in and its measured speed out [20]. Conventional practice holds the spindle at constant speed by feedback control of the motor. Neglecting the losses through the pin and the spindle, the heat put into the work is the product of torque and rotational speed, so one proposal regulates the motor current to hold the torque constant and lets the speed vary [21]. With a small tool the torque changes little with temperature, which makes it a poor variable to control [20]. The chapter's spindle follows its command with a 0.25 s lag, and its torque is the heat divided by the spindle speed.

### Where the Controller Runs

In one robotic cell the force loop runs inside the robot controller, and the temperature loop is a software PI controller on a separate measurement computer that sends a spindle-speed setpoint over a fieldbus. Its loop runs at 5 Hz, limited by the iteration speed of the robot program, and its first-order model came from a step in spindle speed during steady welding [20]. The chapter's PI loops execute every 0.2 s, the same period, and its predictive controllers every 1.0 s, a layer that would sit on such a supervisory computer. Figure 3.5 draws the loops on that hardware.

> **IN PRACTICE** **Qualifying the Recipe, Not the Loop**
>
> Welds in crewed spaceflight hardware are qualified by statistics. Once the optimal parameters are set for a material, a joint design and a thickness, a large population of test welds is made and tested, and the strength data must show with 95 percent confidence that 90 percent of the weld strengths fall within the structural design criteria. On the tank the friction stir welds earned a greater margin of safety than the arc welds before them, because their test data scattered less [3]. The Discussion of this chapter notes that a sacrificial run-out tab is cheaper than requalifying a recipe.

### What Is Modernizing It

The tool-workpiece thermocouple needs no sensor inside the tool. A PI loop on its reading, moving spindle speed, has held the welding temperature within 10 °C on workpieces of changing shape and has started the traverse when the plunge reached its temperature, and its authors left more advanced models and control schemes to future work [20]. For the welds of a launch vehicle, a physics-based flow model predicts input power and wormhole defects, and a reinforcement-learning agent adjusts spindle speed, travel speed and load during the weld to hold input power in a target window [22]. The

self-reacting tool, with a second shoulder under the plate, removes the expensive tooling that reacts the forging load from below [3], which is the backing anvil against which this chapter measures its root clearance.

## 3.8 Limitations

The application is not a friction stir welding process model. With four thermal nodes, a calibrated load-sharing split and an empirical quality window, no claim about a real weld's peak temperature, its nugget geometry, its defect map or its mechanical properties can rest on it, and the effective slip at the tool interface, the largest uncertainty in published thermal models, is not represented at all [5]. What it is evidence for is regime-dependent plant behavior, the servo against regulator trade with numbers attached, step-response predictive control from identification through to measured execution cost, and what a model of the wrong operating point costs.

Three scope limits follow from the control design rather than from the physics. The predictive controllers are single input, manipulating spindle speed only, because traverse speed is the production rate and an optimizer free to slow the machine down needs an economic objective the application does not carry, so multivariable move allocation cannot be demonstrated. Both identified models are first order plus dead time while the real response has at least three thermal modes in it, and the plunge residual of Table 3.3 reports that mismatch, which means the plunge-model controller is handicapped by structure as well as by operating point. The axial axis offers position or force but not the interlocked combination that production machines run [10], which is why the challenge can be failed on the root. Section 3.3.9 states where each physical assumption breaks.

Three elements of the scenarios are stylized on purpose. A plate whose thickness changes by 0.10 mm in one instant is not something rolled stock does, but only a step gives a response that can be read straight off a trend. A tool that reaches half its wear life inside one weld wears far faster than a real tool in aluminum, which puts a drift a shop discovers over weeks inside a single run. The 330 °C soak setpoint turns a saturated actuator into a control problem, which is what lets the two identified models be compared where they disagree most.

> ⚠ **OPEN CHALLENGE** **No Agreed Target for a Sound Weld**
>
> Weld temperature is agreed to bear on quality, yet there is no agreement on the best weld temperature. Reported targets run from 60 to 90 percent of the solidus temperature in degrees Celsius, and a reading at one point of the tool is questioned because the temperature varies across it [18]. In the welds of a launch vehicle, abnormally low strengths have appeared sporadically, and an investigation tied quality to input power instead, which suggests holding a power window [22]. The chapter's window is a teaching device. Progress would be a quality window measured on production welds and stated in a variable a machine can hold.

## 3.9 Discussion

A friction stir weld runs open loop on a recipe because the middle of a long weld is genuinely forgiving, and Table 3.4 shows where that forgiveness ends. Started at quasi-steady conditions in scenario 4, each of the four automatic architectures holds the tool within 0.1 °C of setpoint, and over a complete weld every millimeter of defect any of the five leaves lies in the first 52 mm of the joint, so the whole spread between the best and the worst of them, 89.8 against 98.3 % sound weld, is made where the traverse begins. Two of the disturbances are unmeasured on a real machine and both are ordinary. Tool wear moves the steady tool temperature by 34.7 °C between a new tool and half life at the same spindle speed, which is why recipes are written per tool, and a plate lifting off the backing bar is worth 14.3 °C. Closing the temperature loop is how the two ends of the weld are recovered and how both drifts become visible in the manipulated variable, which is the case the closed-loop literature has been making since the rotating-tool measurement became practical [12, 13].

**OPEN CHALLENGE When Is a Model Accurate Enough?**

A survey of industrial predictive control in 2003 found model uncertainty poorly handled. Products could be detuned for robustness, but the trade against performance was not clear, and until that link is made, the survey concluded, it is not possible to tell when a model is accurate enough. It expected test signals to be designed to explore the regions where a model is inadequate for control [8]. The two step tests of Table 3.3 meet both questions inside one weld. Progress would be a test that finds where a model must be accurate.

Methods well beyond the five architectures implemented here fit the model of this chapter. The four-node balance is small enough for nonlinear predictive control with the shear-layer temperature as a state, which would remove the regime problem rather than trade between its two halves, and that is the direction recent friction stir welding work has taken [14]. A gain schedule on the engagement fraction $a_f$ of (3.2) is a smaller step and suits a plant whose gain moves with a measured geometric quantity [6]. Moving horizon estimation on the same equations could recover the anvil contact and the tool wear from the tool temperature and the axial force together. The run-out is the easiest of all, because the machine knows the distance to the plate edge and the rise from (3.12) is a known function of it, so a feedforward spindle taper cancels most of 31.6 °C of predictable disturbance. Almost nobody implements it, because a sacrificial run-out tab is cheaper than requalifying a recipe.

## 3.10 Conclusion

This chapter studied a friction stir weld as a laboratory for plants whose linearization changes inside a single batch. The simulator integrates nine states covering the shear layer, the tool, the heated halo, the anvil interface, the two drives, the axial position, the distance welded and cumulative tool wear, with Schmidt heat generation over three contact surfaces, an interface shear stress taken as a smooth minimum of a sliding and a sticking branch, a flow stress falling by a factor of eighteen from room temperature to 500 °C, an axial force limited to 20 kN, and a plate conductance that changes with traverse speed and collapses at the plate edge. The design point is 471.4 °C at the tool with 9.50 kN on the axis and 1755 W of frictional and plastic work, which is 877 J into every millimeter of joint. A step test in quasi-steady welding identifies 0.1425 °C/rpm with a 12.24 s time constant, and the same experiment through the plunge returns 0.2401 °C/rpm with 1.17 s, ratios of 1.68 and 10.5 within one weld. Entering Auto from the manual station once the run is under way, ITAE setpoint tuning holds the plunge soak to 18 °C s against 213 for disturbance tuning and then overshoots a setpoint step by 1.5 °C against 10.9, while a loop started already in Auto saturates and holds neither soak. The predictive controller built on the traverse step test finishes a weld 98.3 % sound, and the identical controller built on the plunge step test enters the window fastest of the five and then finishes 91.6 % sound with 18 mm of flash, both at 0.27 µs per execution. A fuller model would add a second manipulated variable with an economic objective, a gain schedule on the engagement fraction, an interlocked axial axis, and a distributed thermal model against which the four-node network could be checked. On this weld the price of choosing the wrong regime for an identification experiment is paid in millimeters of scrap.

**KEY POINTS**

- Heat in a friction stir weld falls as the metal softens, so the weld does not run away.
- The plunge has 1.7 times the gain and a tenth of the time constant of quasi-steady welding.
- The traverse-model DMC finishes a weld 98.3 % sound, the plunge-model DMC 91.6 % with flash.
- Servo and regulator tunings change places between the soak and the traverse.
- Force control holds the heat against a thin plate and spends the root clearance.

**Where else this applies.** A tank in Volume 1, Chapter 2 also hands its operating point to the model, because its gain and time constant both follow the level.

## References


[1] R. S. Mishra and Z. Y. Ma, "Friction stir welding and processing," *Materials Science and Engineering: R: Reports*, vol. 50, no. 1–2, pp. 1–78, 2005.

[2] P. L. Threadgill, A. J. Leonard, H. R. Shercliff, and P. J. Withers, "Friction stir welding of aluminium alloys," *International Materials Reviews*, vol. 54, no. 2, pp. 49–93, 2009.

[3] J. Ding, B. Carter, K. Lawless, A. Nunes, C. Russell, M. Suites, and J. Schneider, "A decade of friction stir welding R&D at NASA's Marshall Space Flight Center and a glance into the future," NASA Marshall Space Flight Center, NASA Technical Reports Server preprint 20080009619, 2006. [Online]. Available: https://ntrs.nasa.gov/citations/20080009619

[4] T. Sheppard and A. Jackson, "Constitutive equations for use in prediction of flow stress during extrusion of aluminium alloys," *Materials Science and Technology*, vol. 13, no. 3, pp. 203–209, 1997.

[5] H. Schmidt, J. Hattel, and J. Wert, "An analytical model for the heat generation in friction stir welding," *Modelling and Simulation in Materials Science and Engineering*, vol. 12, no. 1, pp. 143–157, 2004.

[6] D. J. Leith and W. E. Leithead, "Survey of gain-scheduling analysis and design," *International Journal of Control*, vol. 73, no. 11, pp. 1001–1025, 2000.

[7] E. F. Camacho and C. Bordons, *Model Predictive Control*, 2nd ed., ser. Advanced Textbooks in Control and Signal Processing. London: Springer, 2007.

[8] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[9] W. M. Thomas and E. D. Nicholas, "Friction stir welding for the transportation industries," *Materials & Design*, vol. 18, no. 4–6, pp. 269–273, 1997.

[10] B. T. Gibson, D. H. Lammlein, T. J. Prater, W. R. Longhurst, C. D. Cox, M. C. Ballun, K. J. Dharmaraj, G. E. Cook, and A. M. Strauss, "Friction stir welding: Process, automation, and control," *Journal of Manufacturing Processes*, vol. 16, no. 1, pp. 56–73, 2014.

[11] W. R. Longhurst, A. M. Strauss, and G. E. Cook, "The identification of the key enablers for force control of robotic friction stir welding," *Journal of Manufacturing Science and Engineering*, vol. 133, no. 3, p. 031008, 2011.

[12] A. Fehrenbacher, N. A. Duffie, N. J. Ferrier, F. E. Pfefferkorn, and M. R. Zinn, "Toward automation of friction stir welding through temperature measurement and closed-loop control," *Journal of Manufacturing Science and Engineering*, vol. 133, no. 5, p. 051008, 2011.

[13] B. S. Taysom, C. D. Sorensen, and J. D. Hedengren, "A comparison of model predictive control and PID temperature control in friction stir welding," *Journal of Manufacturing Processes*, vol. 29, pp. 232–241, 2017.

[14] S. P. Meyer, S. Fuderer, and M. F. Zaeh, "A holistic, model-predictive process control for friction stir welding processes including a 1D FDM multi-layer temperature distribution model," *Metals*, vol. 11, no. 3, p. 502, 2021.

[15] Y. J. Chao, X. Qi, and W. Tang, "Heat transfer in friction stir welding—experimental and numerical studies," *Journal of Manufacturing Science and Engineering*, vol. 125, no. 1, pp. 138–145, 2003.

[16] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[17] C. E. Garcia and A. M. Morshedi, "Quadratic programming solution of dynamic matrix control (QDMC)," *Chemical Engineering Communications*, vol. 46, no. 1–3, pp. 73–87, 1986.

[18] A. C. F. Silva, J. De Backer, and G. Bolmsjö, "Temperature measurements during friction stir welding," *The International Journal of Advanced Manufacturing Technology*, vol. 88, no. 9–12, pp. 2899–2908, 2017.

[19] NASA Marshall Space Flight Center, "Controlling force and depth in friction stir welding," NASA Tech Briefs, November 2005, pp. 21–22, MFS-31310, NTRS 20110016397, 2005. [Online]. Available: https://ntrs.nasa.gov/citations/20110016397

[20] J. De Backer, G. Bolmsjö, and A.-K. Christiansson, "Temperature control of robotic friction stir welding using the thermoelectric effect," *The International Journal of Advanced Manufacturing Technology*, vol. 70, no. 1–4, pp. 375–383, 2014.

[21] NASA Marshall Space Flight Center, "Heat control via torque control in friction stir welding," NASA Tech Briefs, September 2004, p. 19, MFS-31834, NTRS 20110020531, 2004. [Online]. Available: https://ntrs.nasa.gov/citations/20110020531

[22] P. W. C. Northrop, A. Maenza, S. Tashakkor, R. Amaro, and J. Anders, "Improving friction stir welding reliability through machine learning control and physics-based modeling," Poster, 2025 Jamboree within the Marshall 65 Spring Showcase, Huntsville, AL, NTRS 20250003564, 2025. [Online]. Available: https://ntrs.nasa.gov/citations/20250003564

## 3.11 Exercises

**3.1** Concept check · 2 min

The Temperature control selector offers five architectures: Manual, two PI loops and two predictive controllers. Which controlled variable and which manipulated variable do all five share?

(a) The axial force, held by moving the commanded plunge depth
(b) The tool temperature, held by moving the traverse speed
(c) The tool temperature, held by moving the plunge depth
(d) The tool temperature, held by moving the spindle speed

**3.2** Concept check · 1 min

Which of these quantities does the simulated machine measure and pass to a controller?

(a) The tool temperature, read by a thermocouple in the rotating tool
(b) The deviation of the plate thickness from its nominal value
(c) The contact between the plate and the steel backing anvil
(d) The wear of the tool shoulder since the start of the weld

**3.3** Concept check · 2 min

The predictive controller identified during the plunge carries a model gain about 1.9 times the stored traverse gain and a time constant about a sixth as long. Run over a complete weld from a cold plate, what does it do in the traverse?

(a) It oscillates with a growing amplitude until the spindle reaches a limit
(b) It holds a slowly corrected offset that runs the tool hot and lays down flash
(c) It enters the window late and leaves a long cold defect at the start
(d) It matches the traverse model, because the output bias removes model error

**3.4** Theory and calculation · 6 min

The flow stress of the shear layer follows Equation (3.3) with $\sigma_0 = 1329\,\text{MPa}$, $T_\sigma = 158.8\,^\circ\text{C}$ and $\sigma_{res} = 8\,\text{MPa}$, and $\phi = 1$ below the solidus.

(a) Find the flow stress at 22, 400 and 500 °C. *Checkpoint:* 1165, 115 and 65 MPa.

(b) Find the ratio of the flow stress at 22 °C to that at 500 °C, the tool temperature of a running weld. *Checkpoint:* About 18.

(c) Raising the spindle from 450 to 900 rpm lowers the interface shear stress under the shoulder from 47.0 to 29.4 MPa. Use Equation (3.5) at a fixed contact to estimate the change in shoulder heat, and compare it with doubling the heat. *Checkpoint:* A rise of about 25 %, not 100 %.

**3.5** Theory and calculation · 7 min

The stored traverse model of Table 3.3 is $K_p = 0.1356\,^\circ\text{C/rpm}$, $\tau_p = 11.78\,\text{s}$ and $\theta_p = 2.68\,\text{s}$, and the stored plunge model has $K_p = 0.2552\,^\circ\text{C/rpm}$.

(a) Find $\theta_p/\tau_p$ and the ITAE setpoint tuning of Equation (3.18). *Checkpoint:* 0.2275, $K_c = 16.77\,\text{rpm}/^\circ\text{C}$, $\tau_I = 11.87\,\text{s}$.

(b) Find the ITAE disturbance tuning of Equation (3.19) and its ratios to the setpoint tuning. *Checkpoint:* $K_c = 26.91\,\text{rpm}/^\circ\text{C}$, $\tau_I = 2.90\,\text{s}$, with 1.60 times the gain and 4.1 times the reset rate.

(c) Find the loop gain $K_c K_p$ of the shipped servo tuning, 16.8 rpm/°C, on the stored traverse and plunge models. *Checkpoint:* 2.28 and 4.29.

**3.6** Theory and calculation · 6 min

The quality window follows Equation (3.15) with $T_{lo,0} = 435\,^{\circ}\mathrm{C}$, $T_{hi,0} = 505\,^{\circ}\mathrm{C}$, slopes $s_{lo} = 0.20$ and $s_{hi} = 0.05\,^{\circ}\mathrm{C}\,\mathrm{min/mm}$, $v_{ref} = 120\,\mathrm{mm/min}$ and $T_{sol} = 582\,^{\circ}\mathrm{C}$.

(a) Find both edges and the width of the window at the challenge's 190 mm/min, and the margin to the cold edge of a tool at 455.6 °C. *Checkpoint:* 449 and 508.5 °C, 59.5 °C wide, and 6.6 °C of margin.

(b) At a fixed 900 rpm the steady tool temperature is 471.4 °C at 120 mm/min and 446.9 °C at 240 mm/min. Find the margin to the cold edge at each speed and estimate by linear interpolation the traverse speed at which the open-loop margin vanishes. *Checkpoint:* 36.4 °C and −12.1 °C, zero at about 210 mm/min.

(c) The design weld generates 1755 W. Find the heat input per millimeter of joint at 120 and at 190 mm/min if the heat stayed the same. *Checkpoint:* About 877 and 554 J/mm.

**3.7** App challenge · 8 min

Load scenario 2, *The plunge: 20 kN and a stalled axis*, default seed, and run it to 26 s at the shipped 900 rpm. Export the CSV.

(a) Read the time at which the axis reaches depth (the first row whose `phase` column reads `DWELL`), the Peak force tile, and the tool temperature `Tt_C` in the last `DWELL` row, at the end of the dwell. *Checkpoint:* About 12.2 s, 20.0 kN, 372 °C.

(b) Reset, set Spindle speed to 1300 rpm before Run, and repeat. *Checkpoint:* About 9.0 s, 20.0 kN, 405 °C.

(c) Reload the scenario, set the halo weight $w_b$ (Weight of the halo temperature in the load-bearing metal) to 0 in Show Model, and repeat at 900 rpm. Say what changed and why. *Checkpoint:* About 3.8 s, 18.1 kN, 319 °C.

Give the times to 0.2 s and the temperatures to 1 °C.

**3.8** App challenge · 8 min

Load scenario 5, *Servo tuning through the plunge*, default seed. It starts from a cold plate with Temperature control at PID · servo already in Auto. Run to 170 s at speed 5, past WELD COMPLETE, and export the CSV.

(a) Read In window, Defect length and the tool temperature `Tt_C` in the last row whose `phase` is `DWELL`. *Checkpoint:* 97.4 %, 6 mm, 400 °C.

(b) Reset, set Temperature control to PID · regulator before Run, and repeat. *Checkpoint:* 96.9 %, 8 mm, 393 °C.

(c) The soak error at the first execution is $330 - 22 = 308\,^{\circ}\mathrm{C}$. Find the proportional kick of each tuning and explain why the servo tuning wins this weld, when Table 3.4 shows the regulator tuning reaching the window first after a bumpless entry.

Give the percentages to 0.1 % and the temperatures to 1 °C.

**3.9** App challenge · 9 min

Load scenario 9, *The plunge MPC meets the traverse*, default seed. It runs the plunge-identified predictive controller in Auto from a cold plate. Run each part to 170 s at speed 5 and read In window, Defect length and Peak tool T.

(a) Run it as loaded. *Checkpoint:* 91.6 %, 21 mm, 507 °C.

(b) Reload, switch to Advanced, and set MPC move suppression to 1.0 and MPC reference time constant to 12 s before Run. *Checkpoint:* 82.5 %, 43 mm, 513 °C.

(c) Reload and set MPC move suppression to 0.02, its lowest value, with the reference time constant left at 2 s. Explain why detuning made this controller worse and less suppression made it better. *Checkpoint:* 94.7 %, 13 mm, 506 °C.

Give the percentages to 0.2 %, the lengths to 1 mm and the temperatures to 1 °C.

**3.10** Application area · 3 min

A fabricator welds 6.35 mm plate with a 6.00 mm pin at 0.15 mm of shoulder plunge, which leaves 0.20 mm of root. The rolled stock carries thin patches of a tenth of a millimeter.

(a) Say what a position-controlled axial axis gives up on a 0.10 mm thin patch, and what a force-controlled axis gives up.

(b) Say how production machines combine the two, and which limit protects the anvil.

**3.11** Application area · 4 min

A friction stir weld is still qualified as a recipe run open loop: a spindle speed, a traverse speed and a plunge depth.

(a) Say why that works for most of a long weld, and where the scrap of an open-loop recipe concentrates.

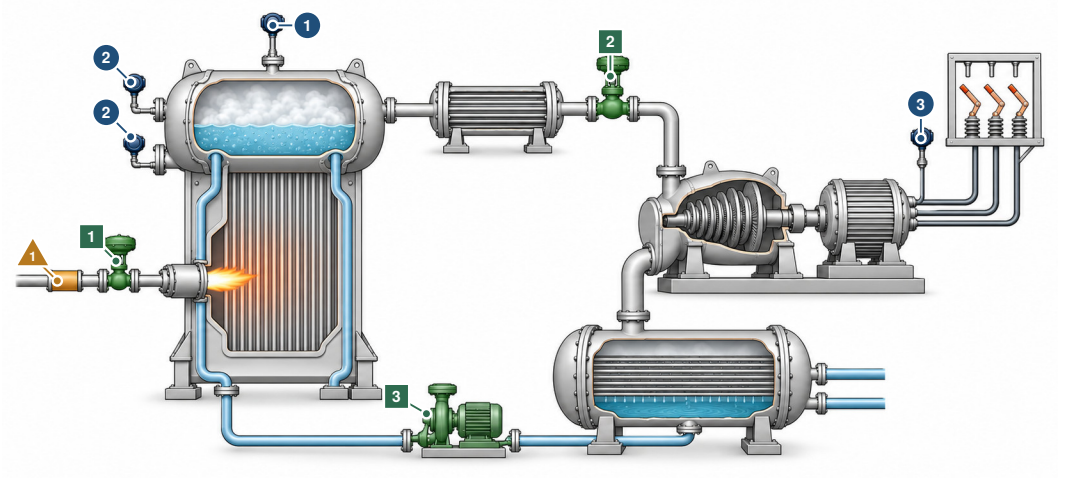
(b) Tool wear and a plate lifting off the backing bar are not measured on a real machine. Say how much each moves the steady tool temperature in the model, and what a closed temperature loop shows an operator when either happens.

**3.12** Review the agent · 5 min

An AI assistant reviews the graded challenge and writes: “Reveal the suggested configuration (traverse-identified MPC at 130 mm/min) and then set the Z axis to Force. Force control removes the 0.09 mm thin patch as a heat disturbance, so the weld can only get sounder, and the challenge passes with a wider margin.” Find the flaw, and say how you would check the claim in the app.

CHAPTER 4

# Boiler-Turbine Simulator Where Fast Megawatts Are Borrowed From Drum Pressure




**The controller moves** 1 fuel valve to the burner 2 turbine steam valve 3 feedwater pump

**The controller reads** 1 drum pressure 2 drum level, two taps 3 electrical output

**What disturbs it** 1 fuel heating value

*Opening the turbine's steam valve gives megawatts within seconds by drawing down the drum pressure, which only more fuel can restore. Illustration, not to scale.*

IN PLAIN TERMS **Borrowing Power From a Boiler**

**The main idea.** A power plant boils water in a drum to make steam that turns a turbine. Opening the steam valve gives more electricity within seconds, paid for by heat stored in the boiler's hot water and steel.

**The challenge.** That quick power is borrowed, because the steam pressure falls and only more fire can pay it back over minutes. A controller that plans ahead must also keep the pressure under its safe limit when the fuel burns hotter or cooler than it expects.

**Why it matters.** Plants built for steady output must now follow a grid whose wind and solar supply keeps changing.

THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the fact in Section 4.1, the history in Section 4.2, the try-it segments in Sections 4.4 and 4.6, the second fact, and the key points |
| **Study** | 40 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 15 min | the first page, the results (Section 4.6), Technology in Practice (Section 4.7), Limitations (Section 4.8) and the open challenge |

**The control question** A drum boiler stores energy in its saturated water and steel, so the turbine valve can deliver megawatts within seconds by drawing down a pressure that only the fuel can restore. Who should repay that loan, and what must a predictive controller know about a fuel it cannot measure to hold its setpoint and its pressure limit?

**Learning objectives**

- Compute the steady map and the linearized pressure time constant of the published drum unit, and explain where a valve step's first megawatts come from.
- Compare boiler-following, turbine-following and coordinated control on a load ramp, and explain what the feedforward lead adds under sliding pressure.
- Explain why a nominal MPC settles with an offset under an unmeasured fuel change, and how an input-disturbance state in the estimator removes it.
- Describe how tube MPC tightens a limit by a bound on the fuel's effect, and state what operating range the guarantee costs.
- Weigh the heat-rate credit of sliding pressure against its slower load response and the unit's ceiling below the nameplate.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.3 (the quadratic program), 1.4.5 (offset-free tracking), 1.4.6 (soft constraints and tightening), 1.4.7 (steady-state decisions) and 1.6.4 (how an exercise names a run).

**The app** https://apmonitor.com/apps/boiler-turbine/

**See also** Volume 6, Chapter 2 for choosing a pairing before any loop is tuned. Volume 17, Chapter 2 for one fuel valve whose two products arrive on different clocks, and Volume 17, Chapter 7 for a fast store and a slow source sharing a load under droop.

## 4.1 Introduction

A thermal unit asked for more power can answer before its fire changes. In the simulated 160 MW drum unit at its nominal point, 66.65 MW at a drum pressure of 108 $kg/cm^2$, opening the turbine's steam valve from 69 to 74 % by hand raises the electrical output 6.6 MW in 40 s while the pressure falls 1.5 $kg/cm^2$ in 100 s. The extra steam is flashed from water and metal that cool with the saturation temperature. Two more points of fuel stop the fall 26 s later, and the pressure recovers only 0.7 $kg/cm^2$ over the next 280 s. Units built for steady output are now asked to follow load and to cycle with variable renewable supply [1], so how a unit spends and refills that stored energy has become an operating question as well as a design one.

> ✦ DID YOU KNOW? **Built for Baseload, Started 1,760 Times**
>
> At one North American coal station, units designed to run at full output and to start cold only a few times a year have averaged 1,760 starts each, 523 of them cold, and at times have cycled on and off four times a day. Most utilities keep minimum generation at 40 to 50 percent of rated capacity [2].

The dynamics behind this trade have been modeled for half a century, from a simplified nonlinear model of a drum unit's pressure and output [3] to a fourth-order physical model of shrink and swell [4], and boiler-turbine dynamics have been reviewed for the power-plant control audience [5]. A third-order model of a 160 MW unit, fitted to plant data and published with its program listings and a basic coordinated controller [6], became a common test plant, analyzed for its nonlinearity and carried with a seven-point operating table [7]. Control designs tried on boiler-turbine units range from practical optimal control [8] and genetic-algorithm tuning [9] to pressure setpoint scheduling [10], fuzzy model predictive control [11] and robust predictive control [12]. The predictive controllers rest on a theory of constrained receding-horizon control [13], on disturbance models that remove steady offset [14, 15], on industrial practice that does the same with a bias update [16], and on tubes that hold constraints under a bounded disturbance [17].

The simulator is built to correct three beliefs. The first is that a fast actuator gives its effect for free. In a drum unit the valve's megawatts are borrowed from the pressure, so a structure that lets the valve hold the output must also decide who repays the loan, and when. The second is that a model predictive controller with a correct optimizer reaches its setpoint. A fuel 10 % poorer than the model assumes leaves the nominal MPC of case 6 1.3 $\mathrm{kg/cm^2}$ below its pressure setpoint, and the cure is a state in the estimator rather than a change in the quadratic program. The third is that a limit written into the program is a limit the plant respects. In case 7 the nominal MPC, whose pressure limit is a soft constraint at 145 $\mathrm{kg/cm^2}$, reaches 145.35 $\mathrm{kg/cm^2}$ under a 10 % richer fuel, and only a plan tightened by a bound on the fuel's effect stays under it.

These are features of a class rather than of boilers. Any plant whose fast handle draws on an inventory that a slow handle must refill has the same shape, among them the line pack of a gas pipeline (Volume 10, Chapter 12), the thermal mass of a building, a reservoir behind a turbine and the holdup of a distillation column. Any plant whose model carries an unmeasured gain error shows the same offset under a nominal predictive controller [14], and any hard limit approached under bounded uncertainty raises the same question of how much operating range a guarantee costs [17]. The drum unit makes all three visible on one plant whose equations are published.

The material a student usually meets treats these pieces apart. The benchmark appears in papers as a set of equations and a table, the structures of boiler-turbine practice appear in handbooks as block diagrams, and offset-free and tube MPC appear in the literature on linear examples whose disturbance has no physical name. This chapter sets the classical pairings, coordinated control and three predictive controllers side by side on one browser simulator of the published drum unit, sharing its plant, instruments and limits, and measures what each buys and costs. The model is the benchmark unchanged [6], and the methods are established ones. The simulator hides the fuel quality as truth behind a boundary of declared measurements, adds an offset-free estimator and a tube whose bound is recomputed at every decision, and states every equation, constant and default it uses. The analysis also records that the first point of the published operating table is not a steady state of the equations. Eight cases and a graded challenge, whose completion is judged from both the learner's actions and the unit's response, carry the measurements.

Section 4.2 follows the history of the unit and its control. Section 4.3 states the plant, its actuators, instruments and parameters. Section 4.4 shows what the unit does before any control: a pressure that integrates fire minus steam, a level that moves the wrong way first, a ceiling below the nameplate and the heat-rate credit of sliding pressure. Section 4.5 describes the unit master, its PI loops, two estimators, the predictive controller and the tube. Section 4.6 measures the eight cases and the challenge, Section 4.7 describes the instruments, valves and control platforms of real units, Section 4.8 states the limitations and Section 4.9 draws the results together. The exercises in Section 4.10 run the same cases in the simulator.

## 4.2 From a Peak-Load Unit to a Control Benchmark

The benchmark comes from a 1987 report on a peak-load unit, as the history segment below tells [6]. Its listings are the source of every coefficient in Section 4.3.

> ⧖ HISTORY **1969: Data Logs From a Peak-Load Unit**
>
> The plant behind the benchmark is a 160 MW oil-fired natural-circulation drum unit whose report lists 140 kg/s of steam at 140 $\mathrm{kg/cm^2}$ at rated load, superheated steam at 535 °C, a 40 $\mathrm{m^3}$ drum, and 40 000 kg of water against 2000 kg of steam in the circuit. By then it ran only for peak load, and it was chosen for the extensive dynamic data logged on it in 1969. A simplified nonlinear model that predicted its drum pressure and electrical output appeared in 1972 [3, 6]. A 1979 extension added the drum level, and the 1987 report brought the order back down with a simple evaporation-rate model, fitted six open-loop logs in which one input at a time was varied by pseudo-random steps, three near 70 MW and 110 $\mathrm{kg/cm^2}$ and three at 135 to 140 MW and 125 to 135 $\mathrm{kg/cm^2}$, and released the models, the data and the listings together [6].

The report also fixed the two features that make the model a control problem rather than a simulation task. Each actuator is written as a lag, a rate limit, an integrator and a bound, with the fuel limited to 0.007 of its range per second, the feedwater to 0.05 per second and the valve to 0.2 per second opening and 2 per second closing. A basic coordinated controller of three PI loops, with interactions between them and anti-windup by limiting the integral terms, is drawn beside it [6]. Both choices reflect the practice the unit was built in. A turbine valve that opens freely while the fuel moves slowly is the reason the classical pairings exist. Boiler-following holds the output with the valve and leaves the fuel to restore the pressure, turbine-following does the reverse, and coordinated control keeps one pairing while feeding the load demand to both handles. Optimal control of a drum boiler power plant was being formulated for practical use by the early 1980s [8], and the dynamics of the boiler and turbine together were later reviewed as the basis of power-plant control [5].

Sliding pressure came from economics. A unit held at rated pressure at part load throttles its steam across a partly closed valve, while a unit whose pressure slides with the load keeps the valve wider and saves fuel, at the price of filling and emptying the drum's stored energy with every load change. Scheduling the pressure setpoint inside a coordinated controller was framed as a multiobjective trade in the early 2000s [10]. The seven-point operating table that the control literature carries with the benchmark lies on such a line, with the pressure rising 10.8 $\mathrm{kg/cm^2}$ and the valve 0.069 from one point to the next [7].

Model predictive control reached the unit through the process industries, where commercial controllers remove steady offset by a bias between the model and the measurement [16]. The theory followed with disturbance models placed in the estimator, so that a persistent mismatch is explained as an integrating state and the prediction becomes right [14, 15]. Robust MPC for bounded disturbances took the form of a nominal plan surrounded by a tube that an ancillary feedback keeps the real state inside [17], and both fuzzy and robust predictive controllers have since been reported for boiler-turbine units [11, 12]. The fuel's heating value, which changes with every delivery of coal, biomass or residual oil, is the physical disturbance these methods are built for, and the simulator makes it the hidden variable of cases 6 and 7.

### 4.3 The Unit as Modeled

The simulated plant is the benchmark's drum, downcomers, risers, steam valve, turbine and alternator, with three actuators and a fuel of uncertain quality, as shown in Figure 4.1. Fuel burns in a furnace lined with water-wall tubes. The heated risers return a steam and water mixture to the drum, unheated downcomers outside the furnace return water to the lower header, and the steam leaves through the superheater in the gas pass, the turbine's control valve and the turbine to the condenser. A feed pump returns the water to the drum. The figure is a process drawing only. The model has no superheater, condenser or pump dynamics, and these appear because the simulator's Unit view draws them.

Table 4.1 gives the role of each variable. Six states are integrated, the drum pressure $p$, the electrical output $E$, the mean fluid density $\rho_f$ and the three actuator positions $a_1$, $a_2$ and $a_3$, while the drum level is algebraic. Time is in seconds, pressure in the benchmark's $\mathrm{kg/cm^2}$, power in MW, flows in kg/s and densities in $\mathrm{kg/m^3}$. The actuator commands and positions are fractions of their ranges.

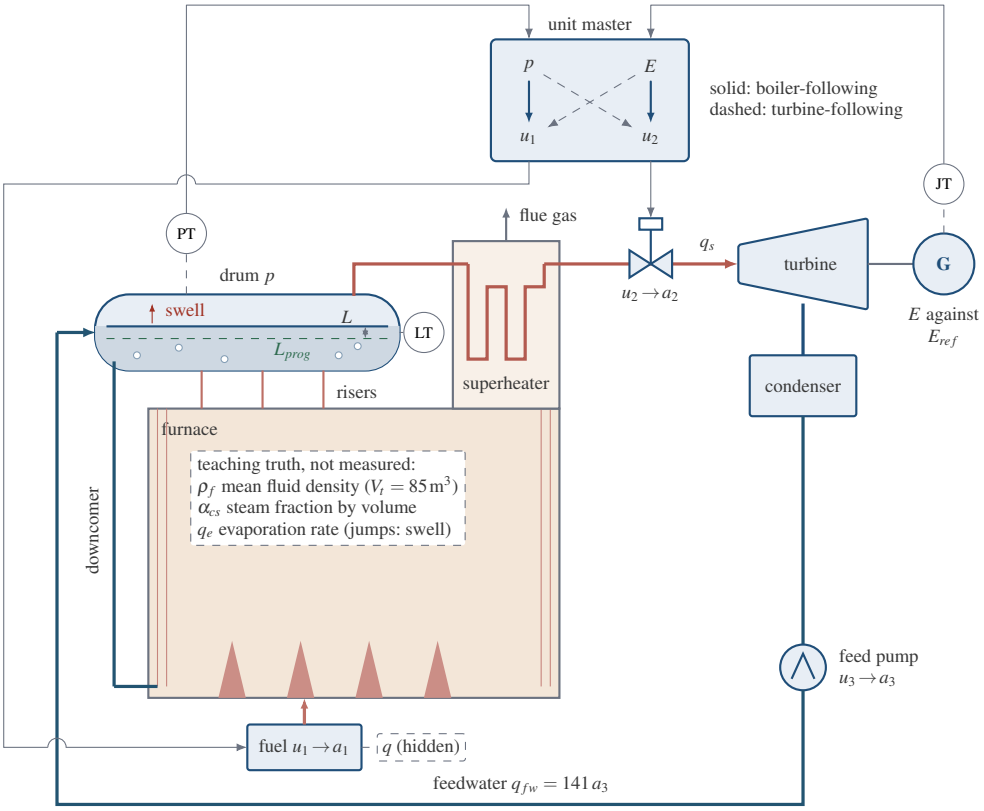



**Figure 4.1** Drum unit as modeled, after the simulator's Unit view. The fuel command $u_1$ moves the fuel actuator $a_1$, whose heat is scaled by the hidden fuel quality $q$. The drum at pressure $p$ holds a level $L$ that is read from its program $L_{prog}$ and swells when the valve opens. The fluid density $\rho_f$, the steam fraction $\alpha_{cs}$ and the evaporation rate $q_e$ of the risers and downcomer are teaching truth. Steam $q_s$ passes the superheater, the valve $u_2 \to a_2$ and the turbine, whose generator delivers $E$ against its reference $E_{ref}$. The feed pump $u_3 \to a_3$ returns $q_{fw} = 141\,a_3$ from the condenser. The unit master pairs the output with the valve and the pressure with the fuel in boiler-following (solid) and crosses the pairings in turbine-following (dashed). Not to scale.

**Table 4.1** Variables by Role, at the Nominal Point #4 of the Operating Table

| Variable | Role | Nominal | Range |
|---|---|---|---|
| $E$ | CV, measured output | 66.65 MW | demand 15–160 MW |
| $p$ | CV, measured drum pressure | 108 kg/cm$^2$ | SP 70–145 |
| $L$ | CV, measured level from program | 0 mm | SP ±150 mm |
| $u_1$ | MV, fuel command | 0.3402 | 0–1 (14 kg/s oil at 1) |
| $u_2$ | MV, steam valve command | 0.6900 | 0–1 |
| $u_3$ | MV, feedwater command | 0.4358 | 0–1 |
| $q$ | DV, fuel quality (hidden) | 1 | 0.8–1.2 |
| $\Delta f$ | DV, grid frequency deviation | 0 Hz | ±0.5 Hz |
| noise | DV, instrument noise | 100 % | 0–300 % |
| $\rho_f$ | state, fluid density (hidden) | 427.9 kg/m$^3$ | — |
| $a_1, a_2, a_3$ | states, actuator positions | as $u$ | 0–1 |
| $q_s$ | output, steam flow | 61.45 kg/s | — |
| $q_e$, $\alpha_{cs}$ | outputs, evaporation, steam fraction | truth | — |

### 4.3.1 Pressure, Output and Fluid Density

The three balances of the benchmark are written for the actuator positions, with the fuel quality multiplying the fuel's heat, as shown in Equation (4.1).

$$\begin{aligned}\frac{\mathrm{d}p}{\mathrm{d}t} &= -0.0018\,a_2\,p^{9/8} + 0.9\,q\,a_1 - 0.15\,a_3 \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= \frac{(0.73\,a_2 - 0.16)\,p^{9/8} - E}{10} \\ \frac{\mathrm{d}\rho_f}{\mathrm{d}t} &= \frac{141\,a_3 - q_s}{85}, \qquad q_s = (1.1\,a_2 - 0.19)\,p\end{aligned} \tag{4.1}$$

where $p$ is in kg/cm$^2$, $E$ in MW, $\rho_f$ in kg/m$^3$, 85 m$^3$ is the volume $V_t$ of the drum, downcomers and risers, and $q_s$ is the steam flow in kg/s. The fuel adds heat, the steam leaving through the valve takes it in proportion to $p^{9/8}$, and the cold feedwater takes a little. The turbine and alternator answer the valve through a 10 s lag, and the term $-0.16\,p^{9/8}$ is the heat passed to the condenser and feed heaters. The fuel quality $q$ is the simulator's one addition to the equations. It is 1 for the nominal fuel and scales the heat released per unit of fuel position, so a fuel 10 % poorer has $q = 0.9$.

### 4.3.2 Drum Level, Steam Fraction and Evaporation

The drum level of the benchmark is an algebraic output built from the density, the steam fraction and the evaporation rate, as shown in Equation (4.2).

$$\begin{aligned}L &= 0.05\left(0.13073\,\rho_f + 100\,\alpha_{cs} + \frac{q_e}{9} - 67.975\right) \\ \alpha_{cs} &= \frac{(1 - 0.001538\,\rho_f)(0.8\,p - 25.6)}{\rho_f\,(1.0394 - 0.0012304\,p)} \\ q_e &= (0.854\,a_2 - 0.147)\,p + 45.59\,q\,a_1 \\ &\quad - 2.514\,a_3 - 2.096\end{aligned} \tag{4.2}$$

where $L$ is in meters, $\alpha_{cs}$ is the steam quality by volume written out with the water's specific volume $v_w = 0.001\,538$ m$^3$/kg and the steam density $0.8p - 25.6$ fitted to steam tables, $0.13073 = v_w V_t$, and $q_e$ is the evaporation rate in kg/s with the report's constants multiplied out [6]. The term $q_e/9$ jumps at once with the valve, the fuel and the feedwater, which is the swell and shrink, while the density term carries the slow inventory. The form in meters with zero at the nominal point is the one the control literature uses [7]. It differs from the report's listing, which writes the level in millimeters about a different datum, by a shift of 16 mm.

The level is a deviation variable valid near the operating line. Along the seven-point table it runs from −0.97 to +0.98 m, at 140 kg/cm$^2$ and 120 MW no density gives $L = 0$ (the minimum is +0.017 m), and below a density $\rho^* = \sqrt{100/(v_w V_t (1/\rho_s - v_w))}$, which is 276.1 kg/m$^3$ at 140 kg/cm$^2$, the level falls as the density rises. The simulator therefore holds and shows the level relative to a level program $L_{prog}$, the steady $L$ at the table's density for the present load reference and pressure setpoint, followed through a 120 s lag. Over the shipped cases the density comes no closer to $\rho^*$ than a ratio of 1.46.

### 4.3.3 Actuators and Fuel Quality

Each actuator follows its command through a first-order lag whose speed is capped by the report's rate limits, as shown in Equation (4.3).

$$\frac{\mathrm{d}a_i}{\mathrm{d}t} = \mathrm{clip}\left(\frac{u_i - a_i}{\tau_i},\ -r_i^{\downarrow},\ r_i^{\uparrow}\right), \qquad 0 \le a_i \le 1 \tag{4.3}$$

where the rates are 0.007 per second for the fuel, 0.05 for the feedwater, and 0.2 opening and 2 closing for the valve [6]. Later papers often quote 0.02 per second for the valve's opening limit,

**Table 4.2** Parameters. F Fitted (Fixed), T Teaching Value, A Adjustable in Advanced Mode

| Symbol | Meaning | Value | Unit | |
|---|---|---|---|---|
| — | coefficients, Eqs. (4.1)–(4.2) | report | — | F |
| $V_t$ | drum, downcomer and riser volume | 85 | $m^3$ | F |
| $\tau_1, \tau_2, \tau_3$ | fuel, valve, feed lags | 20, 2, 10 | s | T, A |
| $r_1$ | fuel rate limit | 0.007 | $s^{-1}$ | F, A |
| $r_2^{\uparrow}, r_2^{\downarrow}$ | valve open, close limits | 0.2, 2 | $s^{-1}$ | F, A |
| $r_3$ | feedwater rate limit | 0.05 | $s^{-1}$ | F, A |
| $p_{max}$, $p_{min}$ | drum pressure limits | 145, 70 | $kg/cm^2$ | T, A |
| $\Delta L$ | level band | ±0.25 | m | T, A |
| $E_r$ | rated output | 160 | MW | A |
| $R$ | governor droop | 5 | % | T, A |
| LHV | fuel heating value | 40 | MJ/kg | T, A |
| $\sigma_p, \sigma_E, \sigma_L$ | noise, 1σ | 0.03, 0.1, 3 | $kg/cm^2$, MW, mm | T |
| $\tau_n$ | noise correlation time | 2 | s | T |
| $q$ | fuel quality | 1 | — | hidden |
| $p_r$ | rated pressure, credit basis | 140 | $kg/cm^2$ | F |

and the simulator follows the report. The lags $\tau_1 = 20\,\text{s}$, $\tau_2 = 2\,\text{s}$ and $\tau_3 = 10\,\text{s}$ are teaching values, because the report draws the structure without numbers. Fuel position 1 is 14 kg/s of oil and feedwater position 1 is 141 kg/s, the density balance's coefficient. The fuel quality is the product of a slider (80 to 120 %), the steps that events apply (±10 %) and a drift of −10 % over ten minutes. Every event fires once, at the start of a physics step, with a mark in the historian. The governor adds $-(\Delta f/50\,\text{Hz})/R \times 160\,\text{MW}$ to the output reference with a droop $R$ of 5 %, so the challenge's 0.1 Hz dip asks for 6.4 MW more. The frequency offset that such a droop leaves on a grid is the subject of Volume 1, Chapter 8.

### 4.3.4 Instruments and the Truth Boundary

One function per physics step fills the only packet the controllers may read. It holds $p$, $E$ and $L$ with first-order colored noise of 2 s correlation and standard deviations of 0.03 $kg/cm^2$, 0.1 MW and 3 mm at the default noise setting, the three actuator positions, the steam flow computed from the valve and the measured pressure as a flow computer would, the feedwater flow $141\,a_3$ and the frequency deviation. The fuel quality and the fluid density are truth, logged in columns marked as truth and drawn dashed in the scene, and a test of the application searches the controller, the estimators, the MPC and the tube for any read of a truth field and finds none. The noise draws come from their own stream seeded by the case, so Reset reproduces a run exactly (Section 1.6.1).

### 4.3.5 Parameters, Initialization and Numerics

Table 4.2 lists every parameter. The plant coefficients are the report's fitted values and are fixed. The actuator, limit and unit parameters are adjustable in the simulator's Advanced mode, and each limit is a teaching value. A real drum's pressure limit comes from its design pressure and safety-valve settings under the boiler code, and its level trips from the drum design. Pressures are read as absolute. The report writes only $kg/cm^2$, and its steam-density line is a least-squares fit to steam tables, which are absolute. Reading them as gauge would move the saturated steam density by 1.1 to 1.6 % and the saturation temperature by under 1 °C between 75.6 and 145 $kg/cm^2$, less than the line's own misfit of up to 4.08 $kg/m^3$ along the table.

In Auto each case starts lined out at its load demand and the pressure its structure asks for. The steady map follows from setting Equation (4.1) to zero, with the valve $u_2 = (E/p^{9/8} + 0.16)/0.73$, the feedwater $u_3 = q_s/141$ and the fuel $u_1 = (0.0018\,u_2 p^{9/8} + 0.15\,u_3)/(0.9\,q)$, and the density is then found by bisection so that the level sits on its setpoint. In Manual the stations set the starting state. At point #4 the map gives a fuel of 0.3402, a valve of 0.6900, a feedwater of 0.4358 and a density of 427.9 $kg/m^3$, and every derivative there is below $3 \times 10^{-17}$. Integration is the classical

fourth-order Runge–Kutta method at 0.1 s, the unit master runs every 1 s, the MPC every 5 s, and the log takes a row every 0.5 s, on the three clocks of Section 1.6.1.

### 4.3.6 Assumptions

The model is the published fit and inherits its range. Its data came from a peak-load oil-fired unit at two loads, and the report notes its poorest fit at high load for valve moves [6]. There are no superheater or reheater temperatures, no attemperation, no air and flue-gas side and no mills, so the fuel's heat reaches the drum through the fuel actuator's lag alone, and a coal unit's mills would add minutes. The level is a deviation variable and breaks down far from the operating line, which is why the simulator programs its setpoint. The actuator lags, the fuel-quality multiplier, the noise, the droop, the heating value and every limit are teaching values, declared as such in the simulator's Model dialog.

## 4.4 What the Unit Does Before Any Control

### 4.4.1 A Pressure That Integrates Fire Minus Steam

Linearized at point #4, the pressure loses $0.002\,51\ \mathrm{s}^{-1}$ of its own excess, a time constant of 399 s, so over the minutes a load change takes the drum is nearly an integrator of fuel minus steam. The output answers the valve through the 2 s actuator and the 10 s turbine, with $0.73\,p^{9/8} = 141.6\,\mathrm{MW}$ of output per unit of valve at fixed pressure. In case 1 the valve step from 69 to 74 % at 20 s gives 6.55 MW in 40 s while the pressure falls $1.50\,\mathrm{kg/cm^2}$ in 100 s and keeps falling, as shown in Figure 4.2a. Raising the fuel from 34.0 to 36 % at 120 s stops the fall 25.9 s later, through the 20 s fuel lag, and the pressure climbs back only $0.73\,\mathrm{kg/cm^2}$ in the next 280 s. The valve is a fast handle on output and on pressure. The fuel is a slow handle on both.

> **TRY IT IN THE APP** **1 · Open Loop: The Valve Spends Stored Energy**
>
> Load the scenario and press Run. Open the Steam valve from 69 to 74 %, as the instructions ask, and watch the output and the drum pressure. The output rises 6.55 MW within 40 s while the pressure falls $1.50\,\mathrm{kg/cm^2}$ in 100 s and keeps falling until the fuel is raised.

### 4.4.2 The Level Moves the Wrong Way First

The same valve step raises $q_e$ at once, and the level swells 22.6 mm within 6.8 s before the draining inventory takes over and leaves it 82 mm below its start at 400 s, with the feedwater held. In case 5 a 20 MW demand step under boiler-following swells the level 38 mm. A level controller that reads only the level sees too much water at the moment the drum is losing it. Volume 18, Chapter 4 studies the same inverse response in the drum of a steam generator.

### 4.4.3 A Ceiling Below the Nameplate

At the rated $140\,\mathrm{kg/cm^2}$ a fully open valve gives 148.0 MW, at the $145\,\mathrm{kg/cm^2}$ teaching limit 154.0 MW, and the nameplate's 160 MW would need $150.04\,\mathrm{kg/cm^2}$. The published table itself ends at 128.9 MW. The equations reproduce its fuel and feedwater within 0.65 %, its levels within its printed 0.01 m and its outputs at points #2 to #7 within 0.01 %. Point #1 is the exception. At its pressure of $75.6\,\mathrm{kg/cm^2}$ and valve of 0.483 the equations give 25.00 MW where the table prints 15.27 MW. At the table's own state $\frac{\mathrm{d}E}{\mathrm{d}t} = 0.97\,\mathrm{MW/s}$ while the other two derivatives are below $10^{-4}$, so the fuel and feedwater entries agree with the valve and the output entry is the one in error (15.27 MW would need a valve of 0.380). The simulator uses the equations.

### 4.4.4 Sliding Pressure and the Heat-Rate Credit

The table's pressures against the outputs the equations give at its valves form the sliding-pressure line, from 25.0 MW at $75.6\,\mathrm{kg/cm^2}$ to 128.9 MW at $140.4\,\mathrm{kg/cm^2}$, linear between points and flat

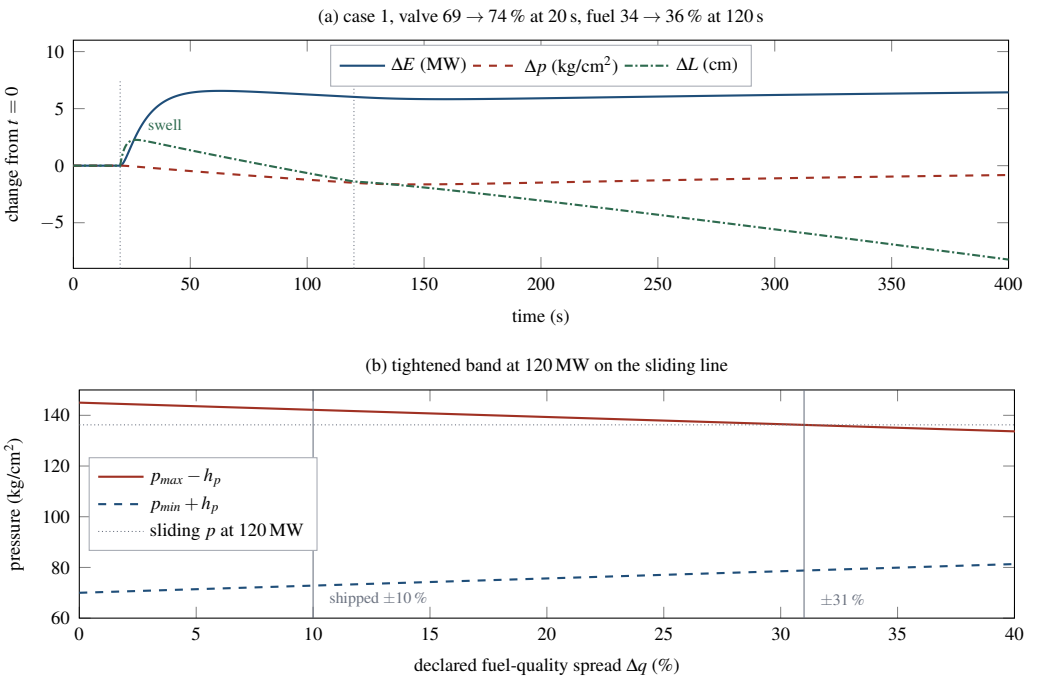



**Figure 4.2** Stored energy and the price of a guarantee. (a) Case 1 in Manual, from the simulator's truth: the output (solid) answers the valve at once while the pressure (dashed) falls, and the level (dash-dotted) swells 22.6 mm before draining. (b) The tube's tightened pressure band at 120 MW on the sliding line, from the simulator's bound with the shipped ancillary gains, against the declared fuel-quality spread: the tightened upper limit (solid), the tightened lower limit (dashed) and the sliding pressure (dotted).

outside them. The heat-rate credit compares the steady fuel at the present output and pressure with the same output at the rated pressure, as shown in Equation (4.4).

$$c = 1 - \frac{u_{1,ss}(E,p)}{u_{1,ss}(E,140)}, \qquad HR = \frac{14\,a_1\,q\,\mathrm{LHV}}{E} \tag{4.4}$$

where $u_{1,ss}$ is the steady fuel of the map above and the heat rate $HR$ is in kJ/kWh after unit conversion. A run's credit is the time average of $c$ while $E > 5$ MW. On the line the credit is 7.76 % at 66.65 MW and 0.58 % at 120 MW. It comes from the model's $-0.16\,p^{9/8}$ loss term, which grows with pressure, while a real unit's credit comes from throttling losses and feed-pump power. A run-integrated fuel comparison is not used, because the stored energy that a sliding ramp moves dominates it (−5.5 % for sliding pressure on the challenge). Choosing the pressure along the line is a steady-state decision of the kind Section 1.4.7 sets out.

## 4.5 The Unit Master as Implemented

The unit master executes every $\Delta t_c = 1$ s on the instrument packet. The load reference $E_{ref}$ ramps toward the demand at the automatic generation control (AGC) ramp rate (8 MW/min by default, zero meaning a step), the governor's share is added to give the target $E_t$, and the pressure setpoint is either the slider's value or the sliding line at $E_{ref}$. Six structures share the fuel and valve, and in every automatic structure the feedwater holds the level.

### 4.5.1 The PI Loops

Every proportional-integral loop is the velocity-form PID of Section 1.6.2 with the derivative off, run every $\Delta t_c$ on a reference $r$ and a measurement $y$. A feedforward bias $b$ takes the place of the output bias $u_b$ of Equation (1.16), the output is limited to the range 0 to 1, and anti-windup comes from back-calculation whose tracking time is $\tau_I$. Every loop not in charge solves its integral for the command actually sent, so a change of structure and a switch between Manual and Auto are bumpless. With the noise off, every switch around the six structures at a steady 60 MW moves no command by more than 0.001 in the following 2 s.

Boiler-following sets the valve from the output error at $K_c = 0.005$ per MW (0.5 % of valve per MW) and $\tau_I = 12\,\text{s}$, and the fuel from the pressure error at 2 % per kg/cm$^2$ and 240 s. Turbine-following crosses the pairing, the fuel on the output at 0.2 % per MW and 40 s, and the valve on the pressure, reverse acting, at 30 % per kg/cm$^2$ and 30 s. How a pairing is chosen before any loop is tuned is the subject of Volume 6, Chapter 2. Coordinated control keeps boiler-following's loops and gives them biases from the model's steady map, as shown in Equation (4.5).

$$\begin{aligned} b_1 &= u_{1,ss} + \frac{T_{lead}}{\tau_f}\left(u_{1,ss} - \bar{u}_{1,ss}\right) \\ \tau_f \frac{\mathrm{d}\bar{u}_{1,ss}}{\mathrm{d}t} &= u_{1,ss} - \bar{u}_{1,ss} \\ b_2 &= \text{clip}\big(u_{2,ss}(E_t, \max(20, p_m)),\ 0,\ 1\big) \end{aligned} \tag{4.5}$$

where $u_{1,ss}$ is the steady fuel at $(E_t, p_{ref})$ with the nominal fuel, $\tau_f = 20\,\text{s}$, $T_{lead} = 60\,\text{s}$, and $p_m$ is the measured pressure. The fuel bias is a lead-lag $(1 + (\tau_f + T_{lead})s)/(1 + \tau_f s)$ on the steady fuel, meant for the energy the drum must store when the pressure rises with the load.

The level loop uses the same law at 2 % of feedwater per 10 mm and $\tau_I = 150\,\text{s}$ on $L$ against $L_{prog} + \Delta L_{SP}$. With the steam-flow feedforward on (three-element control) its bias is the measured steam flow through a 15 s filter divided by 141, and without it (single-element) the bias is zero.

### 4.5.2 Two Estimators

Two constant-gain Kalman filters run every 1 s in every structure, so either MPC takes over warm. The nominal one estimates $\hat{x} = [p, E]$, and the offset-free one adds an input disturbance $d$ on the fuel, the disturbance state of Section 1.4.5, as shown in Equation (4.6).

$$\begin{aligned} \frac{\mathrm{d}p}{\mathrm{d}t} &= -0.0018\,a_2\,p^{9/8} + 0.9\,(a_1 + d) - 0.15\,a_3 \\ d_{k+1} &= d_k + w_k \\ \hat{x}_k &= \hat{x}_k^- + L_K\,(y_k - C\hat{x}_k^-) \end{aligned} \tag{4.6}$$

where $\hat{x}^-$ is the prediction over 1 s with the measured actuator positions held (four Runge–Kutta substeps), $y = [p_m, E_m]$, and the gain $L_K$ comes once from the steady-state Riccati iteration on the model linearized at point #4, with process noise variances of $0.02^2$, $0.2^2$ and $0.001^2$ per step and measurement variances of $0.03^2$ and $0.1^2$. The nominal gain is $[[0.480, 0.000], [0.003, 0.824]]$ and the offset-free gain adds a row $[0.0235, 0]$ for $d$, with a steady standard deviation of 0.49 % of fuel. With $d$, a fuel of quality $q$ appears as $d = (q-1)a_1$, and the pair of the augmented model and the two measurements has rank 3, so the disturbance is observable [15]. The fluid density enters neither predicted balance and is not estimated.

### 4.5.3 The Predictive Controller

The MPC decides the fuel and valve every 5 s with the feedwater left to the level loop. Its model is the pressure and output balances with the two actuator lags, linearized at the estimate with the exact affine remainder and discretized for a zero-order hold, so its first prediction step is right to first order

wherever the unit is. The quadratic program is Equation (4.7), the objective of Equation (1.19) with the pressure band softened by one slack priced as in Equation (1.7) without the linear term.

$$\begin{aligned} \min \ & \sum_{k=1}^{N} w_k \left[ w_E (E_k - E_{t,k})^2 + w_p (p_k - p_{ref,k})^2 \right] \\ & \quad + \sum_j \left( S_1 \Delta u_{1,j}^2 + S_2 \Delta u_{2,j}^2 \right) + 10^4 s^2 \\ \text{s.t. } & 0 \le u_j \le 1, \qquad |\Delta u_{1,j}| \le 0.035 \\ & p_{min} - s \le p_k \le p_{max} + s, \qquad s \ge 0 \end{aligned} \tag{4.7}$$

where $w_k = 1$ except $w_N = 5$, $w_E = w_p = 1$, $S_1 = 2000$, $S_2 = 50$, the references are the unit's own ramped load reference projected ahead (with the governor's share) and the pressure setpoint or the sliding line along it, and the 240 s horizon is $N = 48$ steps in twelve move blocks, four of one step, then 2, 2, 3, 4, 6, 9 and 15 steps, and a last block of 3. The fuel bound 0.035 is the fuel's rate limit over one decision. The nominal MPC predicts with $d = 0$ and the offset-free MPC with $\hat{d}$. A dual active-set solver [18] from the simulators' shared modules computes the plan, and only its first move is applied, the receding horizon of Section 1.4.3 [13]. Any plan the solver does not return as optimal, or whose move is not finite or leaves the unit interval, is discarded and the previous commands stay in force. Over the suggested challenge the offset-free MPC solved 180 programs, rejected none and never took more than 4 active-set iterations, while in case 7 the tube's plans took up to 35 iterations with a pressure slack of up to 0.109 kg/cm$^2$.

### 4.5.4 Tube MPC

The tube MPC keeps a nominal unit $z = [p, E, a_1, a_2]$, simulated with the nominal fuel from the state where the tube took over and never reset to the measurements. The same quadratic program plans the nominal commands $v$ with tightened limits, and the commands sent to the plant add an ancillary feedback on the measured gap, $u_1 = v_1 - K_p(p - z_p)$ and $u_2 = v_2 - K_E(E - z_E)$, with $K_p = 0.02$ per kg/cm$^2$ and $K_E = 0.005$ per MW. To first order the gap $e = x - z$ obeys $e^+ = \Phi_K e + \Gamma w$, where $\Phi_K$ is the closed ancillary loop over 1 s including both actuator lags and $w = 0.9(q - 1)a_1$ is the fuel's missing or extra heat. Starting from $e = 0$ every reachable gap lies in the box of Equation (4.8) [17], and the limits are tightened by it, as Section 1.4.6 describes.

$$\begin{aligned} |e_i| \le h_i &= \bar{w} \sum_{k \ge 0} \left| (\Phi_K^k \Gamma)_i \right| \\ \bar{w} &= 0.9 \, \Delta q \, \min(1, \bar{a}_1) \\ p_{min} + h_p \le z_p &\le p_{max} - h_p, \quad h_{a,i} \le v_i \le 1 - h_{a,i} \end{aligned} \tag{4.8}$$

where $\Delta q$ is the declared fuel-quality spread (10 % by default) and $\bar{a}_1$ starts as the larger of the nominal unit's fuel position and its command, plus 0.05. The bound is recomputed at every decision at the nominal state, the sum stopping when a term adds less than $10^{-10}$ of the total, and a second pass raises $\bar{a}_1$ to the plan's largest fuel plus its own bound when that is higher. For a scalar disturbance the box is exact on the linear error model. A nominal unit that starts above its tightened limit is brought under it along a declared 0.02 kg/cm$^2$ per second ramp rather than in one decision, which only the valve could do. The display calls the plant inside the tube when each measured gap is within $h_i$ plus three standard deviations of the declared instrument noise.

### 4.5.5 Three Models Kept Apart

The plant is Equations (4.1)–(4.3) with the true fuel. The identified model a learner may fit is the response of case 1. The controllers' models are the nominal balances of Equations (4.6) and (4.7), which know neither the fuel quality nor the level, and the tube's linear error model, which leaves out the actuators' rate limits, the plant's nonlinearity away from the nominal unit and the instrument noise.

**Table 4.3** The Cases

| # | Case and events | Teaching point |
|---|---|---|
| 1 | Manual at #4, valve 69 to 74 %, then fuel to 36 % | stored energy |
| 2 | Boiler-following, ramp 66.65 to 85.06 MW at 8 MW/min from 30 s | fast, drifting |
| 3 | Turbine-following, same ramp | steady, slow |
| 4 | Coordinated, same ramp, then sliding, then no lead | the compromise |
| 5 | Single-element level, +20 MW step at 30 s, then three-element | swell |
| 6 | Nominal MPC, fuel −10 % at 60 s, offset-free from 600 s | offset |
| 7 | MPC at 144 $kg/cm^2$, 100 MW, −20 MW and fuel +10 % at 300 s, then tube | a hard limit |
| 8 | Offset-free MPC at 140 $kg/cm^2$, 60 to 80 MW at 60 s, then sliding, then 160 MW | credit, ceiling |
| 9 | Challenge: 60 to 120 MW at 8 MW/min, fuel −10 % at 285 s, −0.1 Hz at 420–480 s | the whole problem |

**Table 4.4** Measured Ramps, Cases 2 to 4. Largest Output Gap and Pressure Gap, and Loop IAE on the Measured Signals From 0 to 360 s

| Structure | $\lvert E - E_t \rvert$ (MW) | $\lvert p - p_{ref} \rvert$ ($kg/cm^2$) | $IAE_E$ (MW s) | $IAE_p$ ($kg/cm^2$ s) |
|---|---|---|---|---|
| boiler-following | 2.69 | 3.51 | 372 | 682 |
| turbine-following | 9.67 | 0.10 | 1608 | 17 |
| coordinated | 0.81 | 1.21 | 69 | 195 |
| coordinated, sliding | 0.81 | 2.96 | 69 | 423 |
| coordinated, sliding, no lead | 0.81 | 4.09 | 70 | 633 |

## 4.6 Eight Cases and a Challenge, Measured

The cases appear in Table 4.3. All share the default tunings, three-element level control, 100 % instrument noise and the nominal fuel, and depart from them only as the rows say. Each number quoted here was produced by running the shipped engine without a browser at the case's seed, advanced with the browser's physics step, with settings entered at the moments the case's own instructions give. Loading a case first, after another case, or by Reset gives identical runs.

### 4.6.1 Three Ways to Follow a Ramp

Cases 2 to 4 ramp the demand from point #4 to point #5 over 138 s, and their measured results appear in Table 4.4. Boiler-following keeps the output within 2.69 MW of its reference and lets the pressure fall 3.51 $kg/cm^2$ below its setpoint, because the valve draws the extra steam before the fuel loop sees any error. Turbine-following holds the pressure within 0.10 $kg/cm^2$ and falls 9.67 MW behind, because the valve may open only as the fire raises the pressure. Coordinated control holds both, 0.81 MW and 1.21 $kg/cm^2$, since the static map starts the fuel with the ramp. With sliding pressure the setpoint climbs from 108.0 to 118.8 $kg/cm^2$ during the ramp and the fire must also fill the drum. The pressure then trails by up to 2.96 $kg/cm^2$ with the 60 s lead and 4.09 $kg/cm^2$ without it, while at constant pressure the same lead slightly overfires (1.21 against 0.99 $kg/cm^2$).

> ✦ DID YOU KNOW? **The Steadiest Pressure, the Largest Output Error**
>
> From the simulator: on the same ramp, turbine-following accumulates an output IAE of 1608 MW s, against 372 for boiler-following and 69 for coordinated control, while its pressure IAE is only 17 $kg/cm^2$ s against 682 and 195 (Table 4.4). Each classical pairing buys one error with the other.

### 4.6.2 Swell and the Three-Element Loop

On the 20 MW step of case 5 the level swells 38 mm under either level loop, since the pressure falls within seconds whatever the feedwater does. What differs is the recovery. The single-element loop takes the swell for surplus water and cuts the feedwater, and the level then falls 107 mm below its

setpoint. With the steam-flow feedforward the feedwater rises with the steam from the start and the dip is 73 mm.

🕮 **Pause.** Try Exercise 4.1 before reading on.

### 4.6.3 The Offset and the Estimator That Removes It

In case 6 the fuel's heating value falls 10 % at 60 s under the nominal MPC, which settles 1.25 $\mathrm{kg/cm^2}$ below its setpoint at 600 s and 1.27 $\mathrm{kg/cm^2}$ at 840 s. Its model predicts that the present fuel will raise the pressure, so it holds fuel back against a rise that never comes, and the estimator's correction each step balances the model's error only with the pressure off its setpoint. A predictive controller has no integral term of its own, so it removes a steady error only through a disturbance model, as Section 1.4.5 shows on the tank. The offset-free estimator has run all along. Its $\hat{d}$ is −3.65 % of fuel at 600 s against a true bias $(q-1)a_1$ of −3.76 %, and after a switch to the offset-free MPC at 600 s the pressure returns within 0.2 $\mathrm{kg/cm^2}$ of its setpoint in 55 s. Both controllers solve the same program. Only the estimate they start from differs.

> **TRY IT IN THE APP** **6 · MPC and the Offset: A Poorer Fuel**
>
> Load the scenario and run it to 600 s. The fuel's heating value falls 10 % at 60 s, and the nominal MPC settles 1.25 $\mathrm{kg/cm^2}$ below its setpoint. Then set Control structure to Offset-free MPC: the pressure returns within 0.2 $\mathrm{kg/cm^2}$ of its setpoint in 55 s.

🕮 **Pause.** Try Exercise 4.2 before reading on.

### 4.6.4 A Limit Held for Every Fuel

Case 7 holds 144 $\mathrm{kg/cm^2}$ at 100 MW, 1 $\mathrm{kg/cm^2}$ under the limit, and at 300 s cuts the demand 20 MW as a step while a fuel 10 % richer arrives. The nominal MPC reaches 145.35 $\mathrm{kg/cm^2}$ and stays over the limit for 558 s, settling 1.2 $\mathrm{kg/cm^2}$ over its setpoint, the offset of case 6 with the opposite sign. The offset-free MPC peaks at 145.25 $\mathrm{kg/cm^2}$ and is over for 27.7 s while its estimate catches up. The tube MPC computes a bound $h_p$ of 2.56 $\mathrm{kg/cm^2}$ at 100 MW and 2.24 $\mathrm{kg/cm^2}$ at 80 MW, plans its nominal unit under the tightened 142.76 $\mathrm{kg/cm^2}$ and peaks at 144.65 $\mathrm{kg/cm^2}$. Its price is range. It cannot hold 144 $\mathrm{kg/cm^2}$, because its nominal unit may not rise above the tightened limit, and its usable pressure range at 900 s is 70.53 of 75 $\mathrm{kg/cm^2}$. With fuels of quality 0.9 and 1.0 in the same case the nominal MPC peaks at 144.35 and 144.92 $\mathrm{kg/cm^2}$ and the tube at 144.00 $\mathrm{kg/cm^2}$, so only the richest fuel of the set breaks the nominal plan.

🕮 **Pause.** Try Exercise 4.3 before reading on.

The tightened band against the declared spread at the challenge's top, 120 MW on the sliding line at 136.24 $\mathrm{kg/cm^2}$, is shown in Figure 4.2b. The bound grows in proportion to $\Delta q$, from 2.83 $\mathrm{kg/cm^2}$ at the shipped 10 %, which leaves a usable range of 69.3 $\mathrm{kg/cm^2}$. The tightened limit falls below the sliding pressure at 120 MW only at a spread of 31 %. Along the whole sliding line at 10 % the usable range stays between 69.3 and 71.6 $\mathrm{kg/cm^2}$, so the tube's price at the declared set is a few $\mathrm{kg/cm^2}$ near the limit rather than lost load.

### 4.6.5 Sliding Pressure and the Nameplate

In case 8 the offset-free MPC takes a step from 60 to 80 MW at 60 s. At the constant 140 $\mathrm{kg/cm^2}$ the output settles inside its ±2 MW band 8.5 s after the step with a credit of 0.05 % over 300 s. With sliding pressure it takes 20.0 s, because the fire must raise the pressure from 103.5 to 115.8 $\mathrm{kg/cm^2}$ as the load rises, and the credit is 6.70 %. Asked for 160 MW, the output stops at 153.96 MW with the valve wide open and the pressure riding its soft limit at up to 145.05 $\mathrm{kg/cm^2}$, over it for 120 s of the 600 s after the request. A logged sweep of 72 runs (six structures, both pressure policies, setpoints of 140 and 145 $\mathrm{kg/cm^2}$ and ramps of 0, 8 and 16 MW/min) found none that holds 160 ± 2 MW for

**Table 4.5** The Challenge Over 900 s. Largest Pressure Gap, Level Gap, Final Output and Pressure Gaps, Credit, and Seeds Passed of 11

| Configuration | $\Delta p$ (kg/cm$^2$) | $\Delta L$ (m) | $E_{fin}$ (MW) | $p_{fin}$ (kg/cm$^2$) | $c$ (%) | Pass |
|---|---|---|---|---|---|---|
| BF, constant (shipped) | 8.4 | 0.07 | 0.25 | 1.07 | 0.6 | 0 |
| BF, sliding | 10.0 | 0.07 | 0.18 | 0.73 | 3.7 | 0 |
| TF, sliding | 0.8 | 0.16 | 2.90 | 0.04 | 3.4 | 0 |
| coordinated, sliding | 4.6 | 0.07 | 0.09 | 1.81 | 3.1 | 0 |
| MPC, sliding | 2.7 | 0.09 | 0.29 | 2.05 | 3.2 | 0 |
| offset-free MPC, constant | 1.4 | 0.08 | 0.13 | 0.10 | 0.0 | 0 |
| offset-free MPC, sliding (sugg.) | 1.6 | 0.09 | 0.13 | 0.12 | 3.0 | 11 |
| tube MPC, sliding | 3.5 | 0.08 | 1.49 | 2.74 | 3.2 | 0 |

120 s under the limit, and the best output under it was 154.66 MW for an instant. The nameplate is the case's honestly infeasible variant. The plan's alternative, a ramp of 10 % of rating per minute, is not infeasible, because it needs a fuel rate about seven times below the 0.007 per second limit, and the offset-free MPC held 32 MW/min at constant pressure within 1.39 kg/cm$^2$.

### 4.6.6 The Challenge

The challenge starts lined out at 60 MW and 140 kg/cm$^2$ in boiler-following and runs 900 s. Its target is the pressure within $\pm 5$ kg/cm$^2$ of its setpoint and under 145 kg/cm$^2$ at every instant, the level within $\pm 0.25$ m of its setpoint, the output within $\pm 2$ MW and the pressure within $\pm 1$ kg/cm$^2$ of their setpoints over the last 120 s, and a credit of at least 2.5 %, all judged on the truth every physics step. The ramp stops at 120 MW because the nameplate is out of reach. Eight configurations appear in Table 4.5. The shipped settings fail on the pressure band, the final pressure and the credit, and every structure except the suggested one misses at least one part. Only the offset-free MPC with sliding pressure passes, and it does so on all eleven seeds tried (the case's and seeds 1 to 10), while every other configuration fails on all eleven. An attempt is not graded once a model parameter or a disturbance slider has been moved.

The pattern in Table 4.5 is the chapter's argument in one place. Sliding pressure alone earns the credit but lets boiler-following stray 10.0 kg/cm$^2$. Coordinated control holds the band but ends 1.81 kg/cm$^2$ off its setpoint, still settling under PI action. The nominal MPC ends 2.05 kg/cm$^2$ off, the offset of case 6, and the tube MPC 2.74 kg/cm$^2$, inside its tube but with no disturbance model. Turbine-following misses the final output by 2.90 MW. Figure 4.3 draws four of the table's columns against the challenge's targets. Case 7 at 600 s under the tube MPC is shown in Figure 4.4.

### 4.6.7 Verification

The application's 49 checks pass. Halving the step to 0.05 s with the noise off moves the challenge's largest pressure gap under the suggested settings from 1.699 to 1.704 kg/cm$^2$, 0.26 %. Over the shipped challenge $V_t \Delta\rho_f$ equals $\int (141\, a_3 - q_s)\, dt$ to 0.016 kg of $9.41 \times 10^4$ kg moved, a relative residual of $1.7 \times 10^{-7}$. The MPC's linearization agrees with central differences of the balances to $1.4 \times 10^{-11}$. For the tube's bound, the worst sign sequence of the disturbance reaches $h_p$ exactly (2.811 against 2.811 kg/cm$^2$) and 200 random sequences reach at most 0.897 kg/cm$^2$, and an independent bound computed by integrating the error model rather than taking its matrix exponential gives 2.2357 against the simulator's 2.2361 kg/cm$^2$ at case 7's end. In a desktop browser at 1366 by 768 and on a 390 by 844 phone viewport, each 900 s case runs in 7.5 s of wall time at 120 times real time, and the slowest MPC solve measured was 4.8 ms, a wall-clock figure that varies between runs.

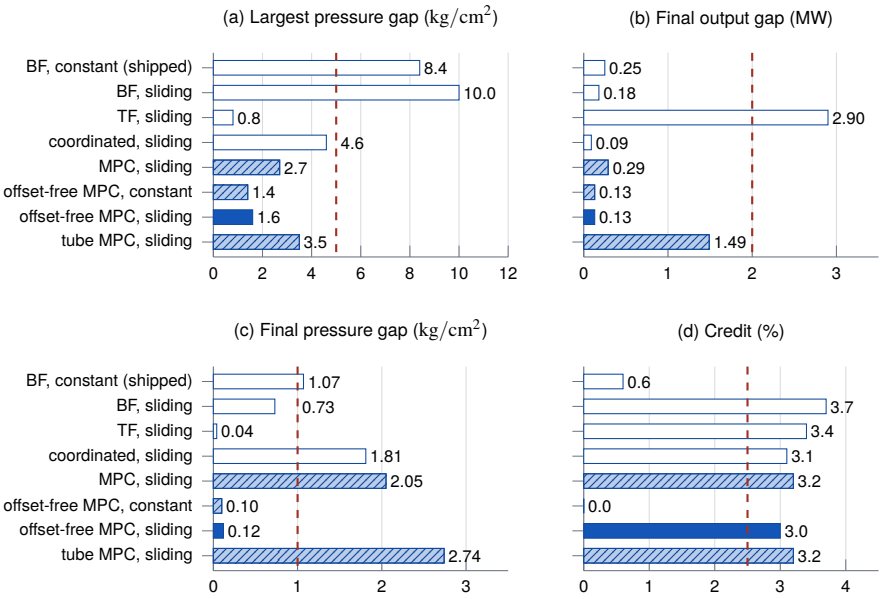



**Figure 4.3** The challenge of Table 4.5 against its targets (dashed lines): (a) the largest pressure gap against the $\pm 5\,\mathrm{kg/cm^2}$ band, (b) the final output gap against 2 MW, (c) the final pressure gap against $1\,\mathrm{kg/cm^2}$ and (d) the heat-rate credit against its 2.5 % floor. White bars are the PI structures, hatched bars the predictive controllers, the dark bar the suggested offset-free MPC with sliding pressure. A value whose label would cross its target line is printed beyond the line. Notice that it is the only row inside all four targets, and that every other row fails at least one.

## 4.7 Technology in Practice

### What Measures It

The unit master of a real plant reads the same three measurements as the simulator: the drum or main steam pressure, the electrical output and the drum level. Pressure detectors serve three functions, indication, alarm and control, and some of them also feed protective features [19]. The drum level and its differential-pressure measurement are the subject of Volume 18, Chapter 4. The chapter's instrument packet holds $p$, $E$ and $L$ with colored noise of 2 s correlation and standard deviations of $0.03\,\mathrm{kg/cm^2}$, 0.1 MW and 3 mm.

> **TECHNOLOGY** **Sensor: The Steam Pressure Transmitter**
>
> A pressure transmitter starts from an elastic element. A bourdon tube, a flattened tube bent into an arc of 270 to 300 degrees, straightens slightly under internal pressure, and a bellows expands against a spring. A strain gauge on a beam moved by the element changes its resistance as its wire stretches, and a bridge circuit turns that change into the pressure signal. Ambient temperature shifts the calibration of the circuitry, humidity can short or corrode it, and the pressure around the detector moves the reading. A failed detector is replaced by a spare element where one is installed, or read at a local mechanical gauge [19]. The chapter's pressure reading carries colored noise of $0.03\,\mathrm{kg/cm^2}$ standard deviation and no calibration shift.

### What Moves It

Three final elements move the unit: the fuel supply to the burners, the turbine's main steam control valve and the feedwater. The simplest turbine-governor models assume a constant steam pressure and an unlimited supply of steam from a boiler that follows the turbine. Grid study guidance calls

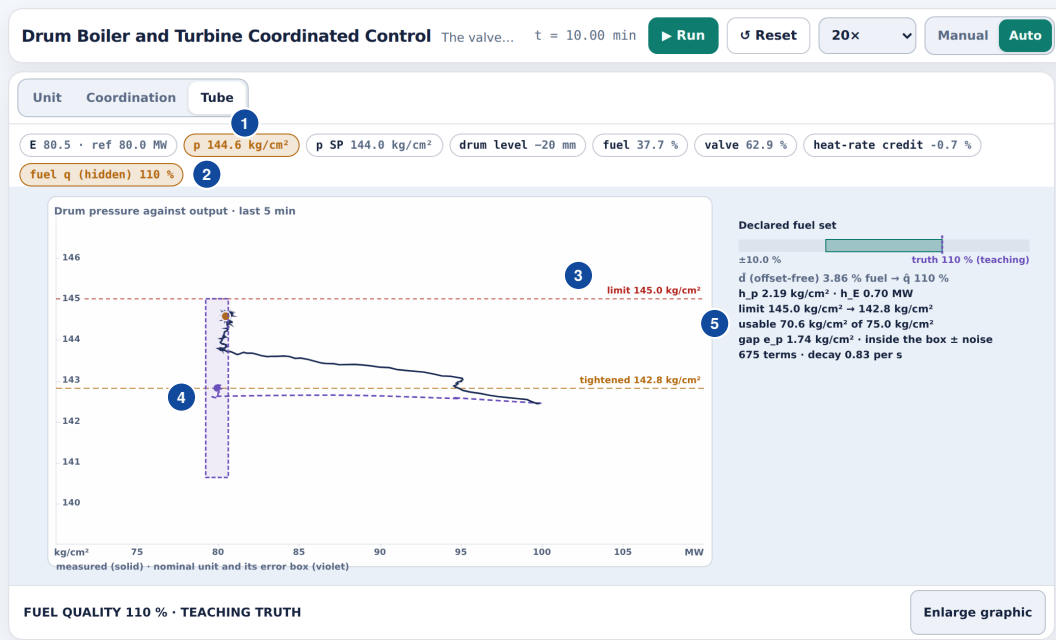



**Figure 4.4** Case 7 under the tube MPC at $t = 600\,\mathrm{s}$ on the Tube view, five minutes after the 20 MW load drop and the 10 % richer fuel, cropped to the header and the Tube view. (1) The measured pressure badge, $144.6\,\mathrm{kg/cm^2}$. (2) The hidden fuel quality, shown as teaching truth. (3) The $145.0\,\mathrm{kg/cm^2}$ limit (upper dashed line) and the tightened $142.8\,\mathrm{kg/cm^2}$ (lower dashed line), with the measured trajectory (solid). (4) The nominal unit (dashed) with its error box. (5) The panel, which reports $h_p = 2.19\,\mathrm{kg/cm^2}$, $h_E = 0.70\,\mathrm{MW}$, a usable range of 70.6 of $75\,\mathrm{kg/cm^2}$ and a gap $e_p = 1.74\,\mathrm{kg/cm^2}$ inside the box, from a sum of 675 terms decaying 0.83 per second, with an offset-free estimate $\hat{d} = 3.86\,\%$ of fuel.

both assumptions extreme simplifications: any significant sudden change of the main control valve drops the throttle pressure, the turbine output drops with it, and more fuel must be burned to restore the steam [20]. Under sliding pressure the control valves are left wide open and the output follows the boiler controls [21]. The chapter's three actuators follow their commands through lags and the report's rate limits, as in Equation (4.3).

> **TECHNOLOGY** **Final Element: The Main Steam Control Valve**
>
> The turbine's main steam control valve sets the steam admitted to the turbine. Turbine-governor models for grid studies carry its maximum opening and closing rates as parameters, estimated with the droop and the valve time constant from tests or recorded disturbances, and measured power against valve position on a large unit is not quite linear. In a recorded frequency event under coordinated control the valve opened at once by droop action, the main steam pressure dropped, and the controls began to close the valve to restore it. A separate fast-valving action suddenly closes the intercept valves after a nearby grid fault [20]. The chapter's valve opens at 0.2 and closes at 2 of its range per second.

### Where the Controller Runs

In a plant the boiler master and the turbine master run in the distributed control system, where a load demand controller selects the operating mode [21]. One recorded response of a large steam unit to a grid frequency event was sampled by the plant's control system at one sample per second,

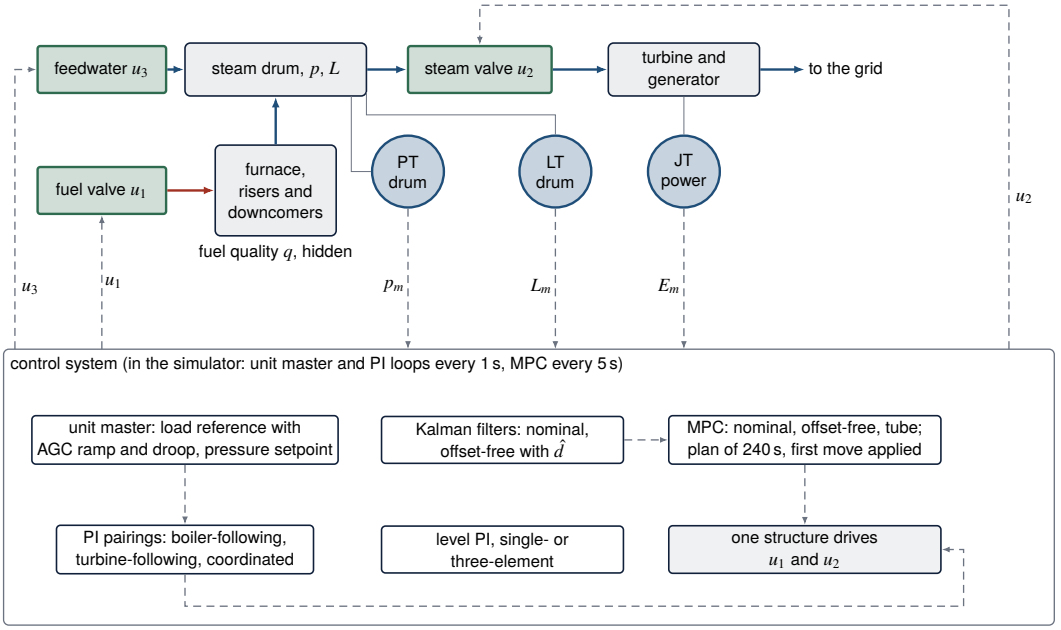



**Figure 4.5** The drum unit's loops on plant hardware, drawn from Sections 4.3 and 4.5. Green boxes are the three final elements, blue circles the drum pressure, drum level and power transmitters, dashed arrows signals. In a plant the boiler master and turbine master run in the distributed control system. Notice that one structure at a time drives the fuel and the steam valve, while the level loop always keeps the feedwater.

and field tests on that unit confirmed a droop of 5 %. Some units add a limit-pressure controller: when the pressure strays beyond a preset limit, the governor valve takes over the pressure and closes quickly, and the output first drops sharply [20]. The simulator's unit master executes every 1 s, the same period, and its predictive controller every 5 s above the PI loops. Figure 4.5 draws the loops on plant hardware.

> **IN PRACTICE** **Checking That the Governor Still Answers**
>
> A steam unit's load demand controller offers turbine follow, boiler follow, coordinated turbine follow and coordinated boiler follow modes, and a technician must make sure the coordinated modes include a frequency trim. Grid reliability guidance lists what silences a governor: turbine-follow mode, in which the output follows the boiler slowly, sliding pressure with the control valves left wide open, and a deadband programmed far too large, such as plus or minus one hertz. Under coordinated control a falling steam pressure partly offsets the governor's response [21]. The chapter's governor acts through a 5 % droop with no deadband.

### What Is Modernizing It

Units built for baseload are being asked to cycle, and the change is mostly in how they are run. One coal station that starts and stops within a day operates at sliding pressure, from 8 to 9 megapascals, for efficiency and flexibility at part load. Its owner credits about 90 % of the later savings to operating procedures, among them trained boiler ramp rates and monitoring equipment that limits ramp rates and tracks temperature changes in headers, heavy fittings and the main steam line [2]. Grid study guidance shows a turbine-governor model that assumed constant pressure failing to match a unit under coordinated control until a simple boiler and a coordinated pressure loop were added, and it recommends validating such models against on-line disturbance recordings [20]. This chapter's

predictive controller carries the pressure, its limit and the fuel's rate limit in one plan, which is where such a ramp procedure would enter as a constraint.

## 4.8 Limitations

The plant is a 1987 fit to a peak-load oil-fired unit at two loads, with its poorest fit at high load for valve moves, and its level is a deviation variable that the simulator has to program. The missing superheater, flue-gas and mill dynamics make the fuel faster than on a coal unit, and the heat-rate credit is the model's loss term rather than a throttling calculation. The actuator lags, the fuel-quality multiplier, the noise and every limit are teaching values. A simulated pass is not a safety case, and a real unit's pressure and level limits come from its design, its safety valves and its grid code.

The MPC is linear, re-linearized at every decision, and plans two handles with the feedwater held. The estimators use one constant gain designed at point #4. The tube is a box on a linear error model with a declared ancillary gain, recomputed at the nominal state, not a polytopic reachable set, and its bound leaves out the actuators' rate limits, the nonlinearity away from the nominal unit and the instrument noise. In case 7 at fuel qualities of 0.9, 1.0 and 1.1 the true pressure gap reaches at most $0.97\,h_p$, while the true output gap reaches $1.81\,h_E$, because the ancillary feedback acts on noisy measurements that the bound leaves out. The disturbances arrive as scheduled steps and do not imitate the variability of a real fuel supply.

> **OPEN CHALLENGE** **Pricing the Damage of a Fast Ramp**
>
> Every start and load swing puts the boiler, steam lines and turbine through large thermal and pressure stresses, worsened by creep-fatigue interaction, and the failures can take several years to appear [2, 22]. A cycling cost study found that faster ramps do more damage, in a relation that is not linear and needs more research [22]. Equation (4.7) prices moves with fixed weights that stand for no damage. Progress would be a damage rate known well enough to enter the cost.

## 4.9 Conclusion

The simulator runs the published third-order model of a 160 MW drum unit as its authors fitted it, with declared actuators and a fuel whose quality no controller can see. A five-point valve step gives 6.6 MW in 40 s by spending $1.5\,\mathrm{kg/cm^2}$ of drum pressure, and the fuel needs 26 s only to stop the fall. On one ramp boiler-following drifts the pressure $3.51\,\mathrm{kg/cm^2}$, turbine-following lags the output 9.67 MW, and coordinated control holds both to 0.81 MW and $1.21\,\mathrm{kg/cm^2}$. A 10 % poorer fuel leaves the nominal MPC $1.3\,\mathrm{kg/cm^2}$ low until the estimator gains a fuel bias, and a 10 % richer fuel at the limit is held only by the tube, at the cost of $4.5\,\mathrm{kg/cm^2}$ of range. The analysis found that point #1 of the published table is not a steady state of the equations. A fuller model would add superheater temperatures, a mill model for coal, a polytopic tube with the rate limits inside it, and a measured fuel analysis with its delay. The unit shows that a fast actuator borrows its effect from an inventory, and that the controller that repays it must know what it cannot measure.

**KEY POINTS**

- Opening a drum unit's steam valve gives megawatts within seconds by drawing down the drum pressure, and only the fuel can restore it.
- On one ramp boiler-following drifts the pressure $3.51\,\mathrm{kg/cm^2}$, turbine-following lags the output 9.67 MW, and coordinated control holds both to 0.81 MW and $1.21\,\mathrm{kg/cm^2}$.
- A 10 % poorer fuel leaves the nominal MPC $1.3\,\mathrm{kg/cm^2}$ low until the estimator gains a fuel bias.
- A 10 % richer fuel at the limit is held only by the tube MPC, at the cost of $4.5\,\mathrm{kg/cm^2}$ of range.

**Where else this applies.** The line pack of a gas pipeline is the same kind of borrowed inventory (Volume 10, Chapter 12).

## References


[1] M. A. Gonzalez-Salazar, T. Kirsten, and L. Prchlik, "Review of the operational flexibility and emissions of gas- and coal-fired power plants in a future with growing renewables," *Renewable and Sustainable Energy Reviews*, vol. 82, pp. 1497–1513, 2018.

[2] J. Cochran, D. Lew, and N. Kumar, "Flexible coal: Evolution from baseload to peaking plant," National Renewable Energy Laboratory, Golden, CO, Brochure NREL/BR-6A20-60575, 2013.

[3] K. J. Åström and K. Eklund, "A simplified non-linear model of a drum boiler-turbine unit," *International Journal of Control*, vol. 16, no. 1, pp. 145–169, 1972.

[4] K. J. Åström and R. D. Bell, "Drum-boiler dynamics," *Automatica*, vol. 36, no. 3, pp. 363–378, 2000.

[5] C. Maffezzoni, "Boiler-turbine dynamics in power-plant control," *Control Engineering Practice*, vol. 5, no. 3, pp. 301–312, 1997.

[6] R. D. Bell and K. J. Åström, "Dynamic models for boiler-turbine-alternator units: Data logs and parameter estimation for a 160 MW unit," Department of Automatic Control, Lund Institute of Technology, Lund, Sweden, Research Report TFRT-3192, 1987. [Online]. Available: https://lup.lub.lu.se/record/8602296

[7] W. Tan, H. J. Marquez, T. Chen, and J. Liu, "Analysis and control of a nonlinear boiler-turbine unit," *Journal of Process Control*, vol. 15, no. 8, pp. 883–891, 2005.

[8] R. Cori and C. Maffezzoni, "Practical-optimal control of a drum boiler power plant," *Automatica*, vol. 20, no. 2, pp. 163–173, 1984.

[9] R. Dimeo and K. Y. Lee, "Boiler-turbine control system design using a genetic algorithm," *IEEE Transactions on Energy Conversion*, vol. 10, no. 4, pp. 752–759, 1995.

[10] R. Garduno-Ramirez and K. Y. Lee, "Multiobjective optimal power plant operation through coordinate control with pressure set point scheduling," *IEEE Transactions on Energy Conversion*, vol. 16, no. 2, pp. 115–122, 2001.

[11] X. Wu, J. Shen, Y. Li, and K. Y. Lee, "Hierarchical optimization of boiler-turbine unit using fuzzy stable model predictive control," *Control Engineering Practice*, vol. 30, pp. 112–123, 2014.

[12] L. Wang, Y. Cai, and B. Ding, "Robust model predictive control with bi-level optimization for boiler-turbine system," *IEEE Access*, vol. 9, pp. 48 244–48 253, 2021.

[13] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[14] K. R. Muske and T. A. Badgwell, "Disturbance modeling for offset-free linear model predictive control," *Journal of Process Control*, vol. 12, no. 5, pp. 617–632, 2002.

[15] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[16] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[17] D. Q. Mayne, M. M. Seron, and S. V. Raković, "Robust model predictive control of constrained linear systems with bounded disturbances," *Automatica*, vol. 41, no. 2, pp. 219–224, 2005.

[18] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[19] *DOE Fundamentals Handbook: Instrumentation and Control, Volume 1*, U.S. Department of Energy, Washington, DC, 1992, dOE-HDBK-1013/1-92.

[20] IEEE Power & Energy Society, Task Force on Turbine-Governor Modeling, "Dynamic models for turbine-governors in power system studies," IEEE Power & Energy Society, Technical Report PES-TR1, 2013. [Online]. Available: https://site.ieee.org/fw-pes/files/2013/01/PES_TR1.pdf

[21] North American Electric Reliability Corporation, "Reliability guideline: Application guide for modeling turbine-governor and active power-frequency controls in interconnection-wide stability studies," North American Electric Reliability Corporation, Atlanta, GA, Tech. Rep., 2019. [Online]. Available: https://www.nerc.com/comm/RSTC_Reliability_Guidelines/Reliability_Guideline-Application_Guide_for_Turbine-Governor_Modeling.pdf

[22] N. Kumar, P. Besuner, S. Lefton, D. Agan, and D. Hilleman, "Power plant cycling costs," National Renewable Energy Laboratory, Golden, CO, Subcontract Report NREL/SR-5500-55433, 2012.

## 4.10 Exercises

**4.1** Concept check · 2 min

In Scenario 5 the demand steps up 20 MW under boiler-following. With either level loop the drum level first rises 38 mm. With the single-element loop it then falls 107 mm below its setpoint, and with the steam-flow feedforward only 73 mm. Why does the single-element loop do worse?

(a) It is tuned with a higher gain, so it overshoots on the way back to the setpoint

(b) It reads the level through a slower filter, so it acts only after the swell has passed

(c) It reads the swell as surplus water and cuts the feedwater while the drum is losing water

(d) It lacks the feedforward that lowers the swell by raising the feedwater before the valve opens

**4.2** Concept check · 1 min

In Scenario 6 the fuel's heating value falls 10 % at 60 s, and the nominal MPC settles 1.25 $\mathrm{kg/cm^2}$ below its pressure setpoint. Which change removes the offset?

(a) Raising the pressure weight, so that the plan values the pressure error more

(b) Lengthening the prediction horizon, so that the plan sees the pressure settle

(c) Removing the fuel's move bound of 0.035, so that the fuel can move faster

(d) Adding a fuel-bias state to the estimator, so that the prediction carries the error

**4.3** Concept check · 2 min

In Scenario 7 the tube MPC keeps the drum pressure under its 145 $\mathrm{kg/cm^2}$ limit through the load drop and the richer fuel, but its plan no longer aims at the 144 $\mathrm{kg/cm^2}$ setpoint. Why?

(a) Its nominal unit must stay under the tightened limit, 145 $\mathrm{kg/cm^2}$ less the bound $h_p$, which lies below 144

(b) Its ancillary feedback is proportional only, so it leaves the same 1.2 $\mathrm{kg/cm^2}$ offset as the nominal MPC

(c) Its slack price of $10^4$ pushes every plan away from the limit by a fixed margin of a few $\mathrm{kg/cm^2}$

(d) The richer fuel drives its fuel command to zero, so the plan has no handle left to hold the pressure

**4.4** Theory and calculation · 6 min

Use Equation (4.1) at point #4 of the table, $E = 66.65\,\mathrm{MW}$ and $p = 108\,\mathrm{kg/cm^2}$, with the nominal fuel $q = 1$.

(a) Find the steady valve position $u_2$ and the steam flow $q_s$. *Checkpoint:* $u_2 = 0.6900$, $q_s = 61.45\,\mathrm{kg/s}$.

(b) Find the steady feedwater $u_3$ and fuel $u_1$, and the fuel flow in kg/s at 14 kg/s per unit of fuel position. *Checkpoint:* $u_3 = 0.4358$, $u_1 = 0.3402$, 4.76 kg/s.

(c) Find the steady output gain of the valve at fixed pressure, and the output a five-point valve step would add if the pressure were held. Say why case 1 shows less than this, 6.55 MW, 40 s after the step. *Checkpoint:* 141.6 MW per unit of valve, 7.08 MW.

**4.5** Theory and calculation · 7 min

Linearize the pressure balance of Equation (4.1) at point #4 ($p = 108\,\mathrm{kg/cm^2}$, $a_2 = 0.6900$, $u_1 = 0.3402$, $u_3 = 0.4358$).

(a) Find $\partial(\frac{\mathrm{d}p}{\mathrm{d}t})/\partial p$ and the time constant of the pressure. *Checkpoint:* $-0.002\,51\,\mathrm{s^{-1}}$, 399 s.

(b) The valve opens from 69 to 74 % with the fuel and feedwater held. Find the initial rate of pressure fall and, treating the pressure as a first-order lag with the time constant of (a), the fall

after 100 s. Compare with case 1. *Checkpoint:* 0.0175 $\mathrm{kg/cm^2}$ per second, 1.54 $\mathrm{kg/cm^2}$ after 100 s.

(c) How many points of fuel cancel the valve's extra steam at the original pressure? Compare with case 1. *Checkpoint:* 1.94 points.

**4.6** Theory and calculation · 6 min

The heat-rate credit of Equation (4.4) compares the steady fuel at the present output and pressure with the same output at 140 $\mathrm{kg/cm^2}$.

(a) Find the steady valve, feedwater and fuel at 66.65 MW and 140 $\mathrm{kg/cm^2}$, and the credit of point #4, where the steady fuel is 0.3402. *Checkpoint:* $u_2 = 0.5708$, $u_3 = 0.4348$, $u_1 = 0.3689$. 7.76 %.

(b) Find the output with the valve fully open at 140 and at 145 $\mathrm{kg/cm^2}$, and the pressure the 160 MW nameplate would need. *Checkpoint:* 148.0 MW, 154.0 MW. 150.04 $\mathrm{kg/cm^2}$.

(c) Which term of the model produces the credit, and what produces it on a real unit?

**4.7** App challenge · 9 min

Load Scenario 2, *Boiler-following on an AGC ramp*, at the default seed and change nothing. Run at the default speed 20 until the run pauses at $t = 360$ s (about 18 s), read the first tile, *Ramp error · heat-rate credit*, and the *Pressure range* tile, and export the CSV file. Do the same with Scenario 3, *Turbine-following on the same ramp*, and Scenario 4, *Coordinated control with load feedforward*, as loaded.

(a) From the last row of each file give the largest output gap, `output_err_max_MW`, and the largest pressure gap, `pressure_dev_max_kgcm2`, to 0.02. *Checkpoint:* 2.69 and 3.51. 9.67 and 0.10. 0.81 and 1.21.

(b) In Scenario 4 set Sliding pressure On, press Reset and run to the pause again. Then set Feedforward lead to 0 s, press Reset and run a third time. Give the largest pressure gap of each run. *Checkpoint:* 2.96 and 4.09 $\mathrm{kg/cm^2}$ (the tile shows $\pm 3.0$ and $\pm 4.1$).

(c) Explain the order of the three structures in (a), and why the lead matters under sliding pressure.

**4.8** App challenge · 8 min

Load Scenario 6, *MPC and the offset: a poorer fuel*, at the default seed. The fuel quality falls 10 % at $t = 60$ s.

(a) Change nothing. Run at speed 60 until the run pauses at $t = 900$ s (about 15 s), read the *Pressure range* tile and export the CSV file. From `pressure_ref_kgcm2` and `truth_pressure_kgcm2` find the mean gap over the rows from 780 to 900 s, to 0.02, and read `est_fuel_bias_pct` and `truth_fuel_bias_pct` in the row at 600 s. *Checkpoint:* Tile 106.7 to 108.0 $\mathrm{kg/cm^2}$, $\pm 1.3$. Mean gap 1.25 $\mathrm{kg/cm^2}$. $\hat{d} = -3.65$ % against –3.76 %.

(b) Press the reload button, set Control structure to Offset-free MPC before Run, and run to the 900 s pause again. Read the tile and, from the file, find the last row after the fuel drop in which the true pressure is more than 0.2 $\mathrm{kg/cm^2}$ from its setpoint. *Checkpoint:* Tile 107.2 to 108.0 $\mathrm{kg/cm^2}$, $\pm 0.8$. Last row at 149 s, 89 s after the drop.

(c) Both runs solve the same quadratic program. Explain why only the second returns to the setpoint.

**4.9** App challenge · 9 min

Load the challenge, *the AGC ramp through a fuel change*, at the default seed. Use speed 60, at which each 900 s run takes about 15 s.

(a) Run it as shipped until it pauses at 900 s and read the *Challenge target* tile. *Checkpoint:* Not met: pressure $\pm 8.4$, final pressure $\pm 1.07\,\mathrm{kg/cm^2}$, credit 0.6 %.

(b) Set Control structure to Coordinated and Sliding pressure On, press Reset and run to the pause again. Which part of the target does it miss? *Checkpoint:* Final pressure, $\pm 1.81\,\mathrm{kg/cm^2}$. Pressure $\pm 4.6$, credit 3.1 %.

(c) Set Control structure to Offset-free MPC, keep sliding pressure, press Reset and run again. Read the tile, and explain why this configuration passes where (a) and (b) fail. *Checkpoint:* Met: pressure $\pm 1.6$, level $\pm 0.09$ m, final output $\pm 0.13$ MW, final pressure $\pm 0.12\,\mathrm{kg/cm^2}$, credit 3.0 %.

**4.10** Application area · 3 min

The challenge's frequency dip is 0.1 Hz on a 50 Hz grid, and the governor has a droop of 5 % on the 160 MW rating.

(a) How much extra output does the governor ask for?

(b) Which handle supplies it within seconds, and what does that cost the drum?

(c) Why does the challenge's ramp stop at 120 MW rather than at the nameplate?

**4.11** Application area · 4 min

The simulator's pressure limit of $145\,\mathrm{kg/cm^2}$ is a teaching value, and the nominal MPC treats it as a soft constraint.

(a) Where does a real drum's pressure limit come from?

(b) In Scenario 7 the nominal MPC reaches $145.35\,\mathrm{kg/cm^2}$ and stays over the limit for 558 s. What role should such a soft constraint play on a real unit?

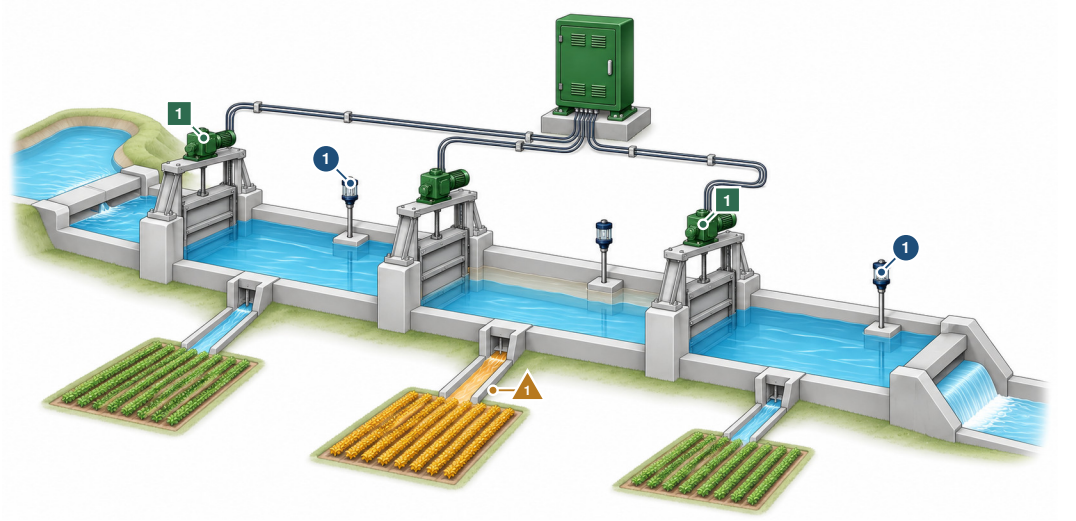
(c) The report gives pressures in $\mathrm{kg/cm^2}$ without saying absolute or gauge. Does the reading matter for the model?

**4.12** Review the agent · 5 min

An AI assistant reviews Scenario 6 and writes: *"The nominal MPC's offset is a tuning problem. Raise the Pressure weight from 1 to 10 and the plan will drive the pressure error to zero, with no need for a disturbance model."* Find the flaw in the claim and say how you would check it in the app.

CHAPTER 5

# Irrigation Canal Simulator for Distant Downstream Control, Resonance and Releasing Water Ahead of Demand



**The controller moves** 1 head gate and two cross-regulators

**The controller reads** 1 water level at each reach end

**What disturbs it** 1 farm orders, booked or unannounced

*Each gate answers to the level at the far end of the reach below it, so the water it releases is seen there only after a wave has crossed the reach. Illustration, not to scale.*

### IN PLAIN TERMS Sending Water Down a Canal on Time

**The main idea.** An irrigation canal carries water from a reservoir to farms through a chain of gates. Water let through a gate needs many minutes to reach the end of the stretch below, so each gate must act early.

**The challenge.** Each stretch also sloshes when its gate is pushed too hard. A controller that plans all the gates at once can share a surprise between the stretches, but it needs the farmers' booked orders.

**Why it matters.** Most of the fresh water that people take goes to crops, so water lost at the end of a canal is a real loss.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the fact in Section 5.1, the history in Section 5.2, the try-it segments in Sections 5.5.1 and 5.5.6, the second fact, and the key points |
| **Study** | 55 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 25 min | the first page, the results (Section 5.5), Technology in Practice (Section 5.6), Limitations (Section 5.7) and the open challenges |

**The control question** Each reach of an irrigation canal stores water, delays every change by a gravity wave's travel time and rings at the wave's round trip. Which gate should answer to which level, how hard can its loop be tuned, and what is a published order, or one plan for every gate, worth against feedback?

**Learning objectives**

- Compute the wave travel time and round trip of a reach, and the lead at which each gate releases a published order.
- Compare local upstream and distant downstream control in water spilled against level deviation, and explain where each pays.
- Explain how a level filter placed between the crossover and the resonance lets a faster gain survive.
- Explain why one plan for three gates shares an unannounced offtake over the reaches, and why it trails gate stroking on published orders.
- Tell what exchanged plans recover over decentralized ones, and what is left to cooperation.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.2 and 1.4.3 (prediction model and quadratic program), 1.4.5 (input-disturbance estimate), 1.4.11 (distributed plans) and 1.6.4 (how an exercise names a run), and Chapter 2 for an integrating plant behind a dead time.

**The app** https://apmonitor.com/apps/irrigation-canal/

**See also** Volume 5, Chapter 2 for fuel that burns after a transport delay and Volume 5, Chapter 4 for a process where every useful move is a prediction. Volume 3, Chapter 6 for feedforward measured ahead of a long delay, and Volume 3, Chapter 2 for a loop structure that decides before any tuning.

## 5.1 Introduction

A canal reach looks like a tank, and a tank is the first process a control course treats. Water enters through a gate at the head, leaves through a turnout and a gate at the tail, and the level integrates the difference. The reach differs from a tank in two ways that decide its control. A change at the head reaches the tail only after a gravity wave has crossed the reach, which on the canal modeled here takes 11 to 19 min, and a reach closed at both ends by flow-controlled gates reflects that wave and resonates at its round trip [1, 2]. A controller designed for the tank is either too slow for the delay or fast enough to feed the resonance.

> ✦ DID YOU KNOW? **Two Hundred Times Slower Than a Pipe**
>
> A flow change travels along a canal as a translatory wave, typically at 3 to 6 m/s, about 200 times slower than a pressure wave in a pipe, so a canal answers a change in demand more slowly than a pipeline [3]. The gravity wave of this chapter's canal runs at about 4.4 m/s (Section 5.3).

The choice that matters most is made before any tuning, as it is for the structure of the level loop of Volume 3, Chapter 2. In local upstream control each gate holds the level just upstream of itself and the head releases a fixed supply, which is easy and keeps every loop local, but whatever the farms do not take spills at the tail. In distant downstream control each gate holds the level at the far end of the reach below it, so the canal releases only what is drawn, but every correction must cross a reach before it can be seen [4]. Published orders offer a way out of the delay. A farmer who books water in advance lets the operator release it early, and doing so gate by gate, one travel time ahead of each order, is the scheduling method known as gate stroking [5, 6].

This chapter is built against the misconception that a canal reach is a tank. The broader class is transport with storage and reflection controlled from its boundaries, which includes rivers regulated by weirs, drainage networks, district heating lines and gas pipelines, and it shares three lessons: the philosophy of control decides whether errors are paid in waste or in delay, the tuning is capped by both a delay and a resonance, and known future demand belongs in a schedule rather than a feedback

loop. Irrigation takes most of the fresh water that people withdraw, which makes operational losses in canals a question of real consequence [7, 8].

This chapter works with a three-reach canal simulator in which the control philosophy, the tuning, the level filter and the use of published orders can be changed while the canal runs, and the outcome is read in water released, delivered and spilled and in level deviations. The hydraulic model, downstream control and gate stroking are established methods. A model predictive controller is offered as a third philosophy, with decentralized and distributed variants, and the chapter compares it with the loops and the variants with one another on the same information about the orders. Section 5.2 traces the automation of canals. Section 5.3 sets out a compact local-inertia model that keeps the wave, the delay and the resonance, with gates that behave as flow sources. Section 5.4 describes the level loops, the order feedforward and the predictive controller with its three architectures. Section 5.5 reports what was measured, starting with the wave timed by hand in Section 5.5.1 and ending with the graded challenge in Section 5.5.7. Section 5.6 describes the gauges, gates and control platforms of real canals, Section 5.7 states where the model stops, and the exercises in Section 5.9 run the same cases in the application.

## 5.2 Automation of Irrigation Canals

Most large canals were built for supply-driven operation. The operator sets a release at the head from the expected demand, gatekeepers adjust the cross-regulators by hand to hold levels upstream of the turnouts, and water that is released but not taken runs to the tail. Hydraulic automatic gates, which hold an upstream level with a float and a counterweight and need no power, made local upstream control automatic, and it remains the most common arrangement in service [9]. The equations that describe the flow, the one-dimensional Saint-Venant equations, were set out for engineering use and solved numerically by the end of the 1960s [10], and methods to compute gate movements that move a canal from one steady state to another without transients followed at the same time [5, 6].

> ⧖ HISTORY **1952: A Little Man at the Check Gate**
>
> On one long supply canal in California, holding the level above a check gate once needed an operator stationed at the gate, and even then the level wandered. Agency staff built a simple device that moved the gate whenever the upstream level left a preset deadband around its target. A canal superintendent called it the best little man on his staff, and the Little-Man went on trial in 1952 as an immediate success. Its successors filtered the level, first through a capillary tube into a second well, which dust and dead insects plugged, then with an electronic circuit, and added a reset that removed the offset of proportional action. That electronic controller held levels downstream of the gates on two canals, with 12 and 3 check structures, from 1974 [3].

Downstream control, in which each gate answers to a level it cannot see, needed remote measurement and a way to live with the delay. A classification of canal control algorithms by controlled variable, manipulated variable, logic and design method, published together with a set of standard test canals, gave the field a common vocabulary in 1998 [4]. Simple controllers based on an integrator-delay model of each reach were developed and tested [2, 11], simplified models that retain the first resonance were derived for controller design [1], and robust tuning of distant downstream PI controllers was worked out for a single reach [12]. The modeling and control of hydrosystems has since been collected in a monograph [13].

Field programs moved the ideas onto canals. Downstream feedback control was analyzed and simulated for branching networks [14], system identification from operating data produced models accurate enough to design controllers without detailed hydraulics [15], and programmable logic controller automation was applied to working canals in modernization projects [8]. Large Australian networks were automated end to end with a combination of these ideas, and the experience has been reviewed as a systems engineering success with continuing challenges [7]. Centralized model

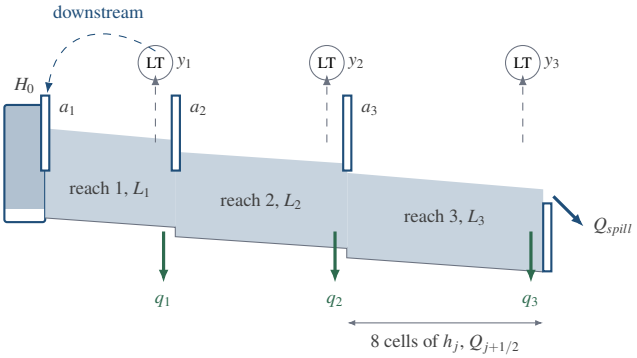



**Figure 5.1** Three-reach irrigation canal. A reservoir at level $H_0$ feeds the canal through the head gate of opening $a_1$, and cross-regulators of openings $a_2$ and $a_3$ divide it into reaches of lengths $L_1$, $L_2$ and $L_3$, each solved on eight cells of depth $h_j$ and face flow $Q_{j+1/2}$. Farms draw $q_1$, $q_2$ and $q_3$ at the reach ends, where the level transmitters read $y_1$, $y_2$ and $y_3$, and the escape weir spills $Q_{spill}$ at the tail. In distant downstream control, as drawn for reach 1, each gate answers to the level at the far end of the reach below it.

**Table 5.1** Variables at the Design Operating Point

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $y_1, y_2, y_3$ | CV, reach end depths | 2.00 m | SP 1.6–2.4 |
| $Q_1, Q_2, Q_3$ | MV, gate flow setpoints | 5.0, 3.5, 2.0 $m^3/s$ | 0–15 |
| $a_1, a_2, a_3$ | states, gate openings | — | 0–1.6 m |
| $h_j, Q_{j+1/2}$ | states, 24 depths, 21 flows | — | $h \geq 0.05$ m |
| $q_1, q_2, q_3$ | DV, farm orders | 1.5, 1.5, 2.0 $m^3/s$ | 0–4 |
| — | DV, published schedule | per scenario | — |
| $n_y$ | DV, level chop ($1\sigma$) | 4 mm | 0–30 |
| $Q_{sup}$ | MV, head supply (upstream) | 5.0 $m^3/s$ | 0–12 |

predictive control, which sets every gate from one optimization over a horizon, was evaluated on ASCE Test Canal 1, one of the standard test canals published with that 1998 classification [16]. Distributed model predictive control, in which the controller of each pool solves its own problem and exchanges information with its neighbors, has been applied to the pools of irrigation canals [17], within a wider taxonomy of decentralized, distributed and hierarchical architectures [18, 19].

## 5.3 Process Model

A reservoir at a fixed level feeds a rectangular canal 6 m wide through a head gate. Two gated cross-regulators split the canal into reaches of 3, 4 and 5 km, each ending in a farm turnout, and the last reach ends in an escape weir. At design the canal carries 5 $m^3/s$ at about 2 m depth: 1.5 to the first farm, 1.5 to the second and 2.0 to the tail. The mean velocity is about 0.42 m/s and the Froude number about 0.1. The arrangement is shown in Figure 5.1.

Table 5.1 classifies the variables. Forty-eight states make up the model: eight depths and seven internal flows in each of the three reaches, and the three gate openings.

**Table 5.2** Canal Parameters, All Editable in Show Model

| Symbol | Meaning | Value |
|---|---|---|
| $B$ | bottom width | 6.0 m |
| $S_0$ | bed slope | $1.0 \times 10^{-4}$ |
| $n$ | Manning roughness | 0.018 s/m$^{1/3}$ |
| $L_1, L_2, L_3$ | reach lengths | 3, 4, 5 km |
| $H_0$ | reservoir level | 3.0 m |
| $C_d$ | gate coefficient | 0.60 |
| $a_{max}$ | maximum opening | 1.6 m |
| — | gate motor speed | 0.24 m/min |
| — | weir crest (depth) | 2.3 m |
| $h_t, h_{min}$ | turnout invert, full-order depth | 1.2, 1.8 m |
| $g$ | gravity (fixed) | 9.81 m/s$^2$ |

### 5.3.1 Local-Inertia Saint-Venant Equations

Each reach is divided into $N_c = 8$ cells, with depths at the cell centers and flows at the faces between them. Mass is balanced on each cell and momentum on each internal face, Equations (5.1) and (5.2).

$$B\,\Delta x\,\frac{\mathrm{d}h_j}{\mathrm{d}t} = Q_{j-\frac{1}{2}} - Q_{j+\frac{1}{2}} - q_r\,\delta_{j,N_c} \tag{5.1}$$

$$\frac{\mathrm{d}Q_{j+\frac{1}{2}}}{\mathrm{d}t} = -gA\left(\frac{z_{j+1}-z_j}{\Delta x} + S_f\right), \quad S_f = \frac{n^2 Q|Q|}{A^2 R^{4/3}} \tag{5.2}$$

where $B$ is the width, $\Delta x = L_r/N_c$ the cell length, $z = z_b + h$ the water surface, $A = Bh$ the area and $R = Bh/(B+2h)$ the hydraulic radius at the face, and $n = 0.018$ the Manning roughness. The advective term of the full momentum equation is dropped. This local-inertia form is accurate at the low Froude numbers of an irrigation canal and is the formulation used for efficient inundation modeling [20]. It retains gravity waves at $\sqrt{gh} \approx 4.4\,\mathrm{m/s}$ and their reflections, so the travel times of the three reaches at 2 m depth are $\tau_r = L_r/\sqrt{2g}$, about 11, 15 and 19 min. The bed falls at $S_0 = 10^{-4}$ and drops a further 0.25 m at each cross-regulator.

### 5.3.2 Gates, Turnouts and the Weir

Each gate is a submerged orifice across the full width, driven by the surface difference across it, Equation (5.3).

$$Q_g = C_d\,B\,a\,\sqrt{2g\,|z_u - z_d|}\;\mathrm{sgn}(z_u - z_d) \tag{5.3}$$

with $C_d = 0.6$. The head gate draws from the reservoir at $H_0 = 3.0\,\mathrm{m}$ above the head bed. Each gate's motor follows its commanded opening with a 10 s lag limited to 0.24 m/min, between zero and 1.6 m. A turnout takes its full order while the depth at the reach end is at least 1.8 m, and less below that, Equation (5.4).

$$q = q_{order}\sqrt{\frac{h - h_t}{h_{min} - h_t}}, \quad h_t < h < h_{min} \tag{5.4}$$

with the turnout invert at $h_t = 1.2\,\mathrm{m}$. The escape weir at the tail is sharp-crested across the full width with its crest at 2.3 m depth, $Q_{spill} = 1.7\,B\,(h - 2.3)^{3/2}$. Freeboard is flagged when a reach end exceeds 2.7 m.

With the values of Table 5.2, every case opens on a steady backwater profile marched upstream from the setpoint depth in every reach, with each gate at the opening that passes its reach's flow.

### 5.3.3 Assumptions

The reaches are prismatic rectangular channels with uniform slope and roughness. Real canals are trapezoidal and carry structures such as siphons and drops that change the wave behavior [13].

Eight cells per reach reproduce the integrator, the delay and the first resonances, and smear the wave front over several hundred meters. Resonant periods are approximate.

Gates are submerged orifices with a constant coefficient and a smooth motor. Gate motors on real canals often move in discrete steps with deadbands, which adds its own limit cycling.

The reservoir level is constant and there is no seepage or evaporation, so water released is delivered, stored or spilled.

Orders are either published exactly or not at all. Real schedules are uncertain in timing and amount.

## 5.4 Implemented Control

The controllers run every 30 s and the hydraulics advance by fourth-order Runge-Kutta steps of 5 s, on the engine of Section 1.6.1.

### 5.4.1 Level Loops in Flow Units

Every level loop is a PI controller in the velocity form of Section 1.6.2, and it computes a gate flow setpoint $Q_k$ every $\Delta t_c = 30\,\mathrm{s}$, as shown in Equation (5.5).

$$\begin{aligned} e_k &= \gamma\left(y_{SP} - y^f_k\right), \qquad P_k = K_c e_k \\ s_k &= s_{k-1} + P_k - P_{k-1} \\ c_k &= Q_{ff,k} + s_k + I_{k-1} \\ Q_k &= \mathrm{clamp}\left(c_k, 0, Q_{max}\right) \\ I_k &= I_{k-1} + \frac{K_c}{\tau_I} e_k \Delta t_c + \frac{\Delta t_c}{T_t}\left(Q_k - c_k\right) \end{aligned} \tag{5.5}$$

where $\gamma$ is the controller direction, $y^f$ the filtered level, $Q_{ff}$ the feedforward bias described below, $Q_{max} = 15\,\mathrm{m^3/s}$ and $T_t = \tau_I$ the tracking time of the back-calculation. The recursion is that of Equations (1.14) to (1.17) without the derivative, with $\gamma$ for the action $a$, $s_k$ for $S_k$, the demanded flow $c_k$ for $v_k$ and the feedforward $Q_{ff}$ as the bias $u_b$, and the integral used at scan $k$ was accumulated through scan $k-1$. The gate opening follows from inverting Equation (5.3) with the water surfaces on either side and the declared gate coefficient and width, $a = Q/(C_{d,0} B_0 \sqrt{2g\,\Delta z})$ with $\Delta z \geq 5\,\mathrm{mm}$, which turns each gate into a flow source at its declared rating and keeps the controller gain in flow units. A gate whose coefficient or width is edited passes a different flow, and the level loop corrects it. Every controller reads the declared gate rating, the declared reach lengths and the turnout meters, which measure what each turnout delivers, so an edit of the canal is model mismatch for the loops as it is for the MPC. The opening is limited to $a_{max}$ after the PI, which does not see that limit, so the integral of a loop whose gate is fully open keeps integrating. In the cases of Section 5.5 only the unstable loops reach either limit. Each level measurement passes a first-order filter $\tau_f$ before the controller. The defaults are $K_c = 15\,\mathrm{m^3/s}$ per meter, $\tau_I = 3600\,\mathrm{s}$ and $\tau_f = 200\,\mathrm{s}$, one tuning shared by the three reaches.

In distant downstream control, gate $r$ holds $y_r$ at the far end of reach $r$ with direct action ($\gamma = +1$), so a falling level raises the flow into the reach, and the bias is the feedforward described below. In local upstream control, the head releases the fixed supply $Q_{sup}$ and gate $r+1$ holds $y_r$ just upstream of itself with reverse action ($\gamma = -1$) and no bias, so a rising level opens it. In manual mode the operator sets the head release, the cross-regulators hold their openings so their flows follow the levels, and each integral is set every scan so that Equation (5.5) reproduces the actual gate flow, which makes the return to automatic bumpless.

### 5.4.2 Order Feedforward

In downstream control each gate can add the orders of every farm below it. With the feedforward set to *now*, gate $r$ adds $\sum_{k \geq r} q_k(t)$, which decouples the reaches but ignores the delay. With it set to

**Table 5.3** Reach Models of the MPC Against the Nominal Values

| Reach | $BL_r$ ($m^2$) | $A_{s,r}$ ($m^2$) | $\tau_r$ (s) | $\theta_r$ (s) | First rise (s) | RMS (mm) |
|---|---|---|---|---|---|---|
| 1 | 18000 | 16217 | 677 | 215 | 580 | 16.1 / 2.5 |
| 2 | 24000 | 21941 | 903 | 295 | 780 | 11.8 / 2.7 |
| 3 | 30000 | 28361 | 1129 | 330 | 980 | 10.0 / 2.0 |

First rise is the time for the reach end to move 1 mm after its own gate steps by 0.3 $m^3/s$. RMS is the error over 3 h in predicting that response, nominal against identified.

*ahead*, gate $r$ adds each order as it will be one travel time later per reach between that gate and that farm, the travel times of the declared reach lengths, Equation (5.6).

$$Q_{ff,r}(t) = \sum_{k=r}^{3} q_k \left( t + k_{lead} \sum_{m=r}^{k} \tau_m \right) \tag{5.6}$$

with $k_{lead} = 1$ by default. The head gate therefore releases for the tail farm three reaches early and the last gate one reach early, which is gate stroking in its simplest form [5, 6]. An unpublished offtake enters through the turnout meters, which read the flow each turnout delivers, only when it happens. A change of feedforward setting is absorbed into each integral, which keeps the gate flows from jumping.

### 5.4.3 Model Predictive Control

A model predictive controller (MPC) is offered as a third control philosophy. Distant downstream control remains the default, both when the application opens and in every scenario other than scenarios 10 and 11, and its three PI loops are the fallback. Every $T_s = 300$ s, or every tenth scan, one quadratic program sets the three gate flow setpoints together over $N = 36$ steps, or 3 h, in 11 move blocks whose lengths in steps are 1 four times, then 2 twice, then 3, 4, 6, 9 and 6. The plan's first move alone goes to the gates and stays until the next decision, the receding horizon of Section 1.4.3 [21, 22]. Each planned flow passes through the same inversion of Equation (5.3), opening limit and 0 to 15 $m^3/s$ clamp as a PI output, so every gate remains a flow source under all three philosophies.

The controller's model is one integrator behind a dead time per reach, the structure long used for canal controller design [2], Equation (5.7). Chapter 2 identifies the same structure as a step response for a balloon's altitude.

$$A_{s,r} \frac{\mathrm{d}y_r}{\mathrm{d}t} = Q_r(t - \theta_r) - Q_{r+1}(t) - q_r(t) + d_r, \quad Q_4 \equiv 0 \tag{5.7}$$

Here $A_{s,r}$ is the reach's effective surface, $\theta_r$ its dead time and $d_r$ an unexplained inflow supplied by the estimator, which also absorbs the tail spill and the model's error. A turnout delivering less than its order enters through the turnout meters the preview reads. The numbers come from step tests run on the simulator at its declared design point, with every parameter at its declared value, orders of 1.5, 1.5 and 2.0 $m^3/s$, levels at 2.0 m and no level chop. The three gate flows were held as flow sources, one gate flow or one order was stepped by $\pm 0.3$ $m^3/s$ after 30 min, and an integrator with a dead time was fitted over the following 3 h in both directions. Each surface is the mean of the fits for every flow into and out of its reach, so balanced flows predict no drift, and each dead time is that of the reach's own gate. The values, frozen in the controller so that an edit of the canal becomes model mismatch, are listed in Table 5.3. An independent refit agrees within 2 $m^2$ and 2 s. Over 3 h after a step the identified model predicts the eight input and output pairs within 2.0 to 3.7 mm RMS, against 5.6 to 16.1 mm for the nominal surfaces $BL_r$ and wave travel times $\tau_r$.

The dead times are the best straight-line fit over the 3 h horizon, not the wave's arrival. A reach end first moves 1 mm after 580, 780 and 980 s, and a fit over 1 h returns 266, 404 and 478 s, so the

**Table 5.4** What Each Structure Is Told About the Farm Orders

| MPC preview | Information | Same as |
|---|---|---|
| Off | the levels only | feedback only |
| Now | the turnout meters now, held over 3 h | feedforward *now* |
| Schedule | the published orders over 3 h, as 5 min means | feedforward *ahead* |

model has no waves, no resonance and no reflection. With $\theta_r = (n_r + f_r)T_s$, $0 \le f_r < 1$, a zero-order hold gives the exact sampled model of Equation (5.8),

$$\begin{aligned} y_{r,k+1} = y_{r,k} + \frac{T_s}{A_{s,r}}\Big[&(1-f_r)Q_{r,k-n_r} + f_r Q_{r,k-n_r-1} \\ &- Q_{r+1,k} - q_{r,k} + d_{r,k}\Big], \qquad d_{r,k+1} = d_{r,k} \end{aligned} \tag{5.8}$$

where $n_r = 0$, 0 and 1 and $f_r = 0.717$, 0.983 and 0.100, so the inflow of a period comes partly from the decision before.

The decisions are the moves of the three gate flows from the flows applied now, constant within each block. Each plan minimizes Equation (5.9), in which $\hat{y}_{r,k}$ is a predicted level, $j$ counts the blocks and $\Delta Q_{r,0}$ is measured from the applied flow.

$$\begin{aligned} \mathscr{J} = &\sum_{k=1}^{N}\sum_{r=1}^{3}\left(\frac{\hat{y}_{r,k} - y_{SP,r}}{0.01\,\mathrm{m}}\right)^2 + \lambda \sum_{j=0}^{10}\sum_{r=1}^{3}\left(\frac{\Delta Q_{r,j}}{0.1\,\mathrm{m^3/s}}\right)^2 \\ &+ \sum_{r=1}^{3}\left(\kappa\,\varepsilon_r^2 + \nu\,\varepsilon_r\right) \\ \text{s.t. } 0 \le Q_{r,j} \le \bar{Q}_r, &\quad |\Delta Q_{r,j}| \le C_d B\sqrt{2g\,\Delta z_r}\,\dot{a}_{max} T_s \\ \underline{y}_{r,k} - \varepsilon_r \le \hat{y}_{r,k} &\le \bar{y}_{r,k} + \varepsilon_r, \quad \varepsilon_r \ge 0 \end{aligned} \tag{5.9}$$

With $\lambda = 3$ a move of 0.1 $\mathrm{m^3/s}$ costs as much as a 1.7 cm level error held for one step. The gate limits are hard. The capacity is $\bar{Q}_r = \min\left(15, C_d B a_{max}\sqrt{2g\,\Delta z_r}\right)$, where $\Delta z_r$ is the water-surface difference across gate $r$ now, as the inversion reads it, $C_d$ and $B$ are the declared values, and $a_{max}$ is the opening limit read as the inversion clips it. The move limit uses the declared motor speed $\dot{a}_{max} = 0.24\,\mathrm{m/min}$. The level bands are soft, with one slack $\varepsilon_r$ per reach priced at $\kappa = 10^7$ per $\mathrm{m^2}$ and $\nu = 10^5$ per m, as in Equation (1.7). Each band, from $\underline{y}_{r,k}$ to $\bar{y}_{r,k}$, is the loops' $\pm 5$ cm about the setpoint, raised where necessary to 2 cm above the full-order depth $h_{min}$ and, at the tail, lowered to 2 cm under the weir crest, so the plan holds water back rather than spill it. When a level already lies outside its band, the band starts at that level and closes back with a 30 min time constant. With the band prices set to zero, every compared case changes its summed IAE by at most 0.56 % and its worst error by at most 0.011 cm, so the bands decide nothing in the comparison below. Each program goes to the dual active-set solver of Goldfarb and Idnani [23], and its result is discarded when the solver does not report it optimal, when a value is not finite or when a flow leaves its bounds.

The orders over the horizon come from a preview with three settings that mirror the feedforward, as listed in Table 5.4. Each reads the same function of the turnout meters and the published schedule that the matching feedforward reads, so each knows exactly what that feedforward knows. The feedforward *ahead* looks one travel time per reach ahead, at most 45 min at $k_{lead} = 1$, and the Schedule preview looks 3 h ahead at the same published orders. An unannounced offtake reaches every structure only when the turnout begins to carry it. Replay tests in which two canals differ only in orders not yet due find identical plans until each preview may see the difference.

A steady-state Kalman filter on $(y_r, d_r)$ runs every 300 s under every philosophy. It predicts with the mean applied gate flows and the mean orders the preview allows over the last period, and corrects with the three level transmitters ahead of the level filter. Its variances per decision are $9 \times 10^{-6}$ m$^2$ for each level, standing for the waves the model leaves out, $2.5 \times 10^{-3} \rho$ (m$^3$/s)$^2$ for each inflow and $2.5 \times 10^{-5}$ m$^2$ for each transmitter, with $\rho = 0.01$. The random-walk inflow is the input-disturbance estimate of Section 1.4.5, and it makes the plan offset-free [24], slowly. The filter's time constants are 162, 219 and 283 min.

The horizon was fixed before tuning. With $\lambda = 1$ and $\rho = 1$, a horizon of 36 steps gave a summed IAE within 1.2 % of 120 steps on four offtake cases and 3.1 % lower on the challenge. One pass over nine values of $\lambda$ from 0.01 to 100 and five of $\rho$ from 0.01 to 100 then scored scenario 10 from 0 to 10 h by its summed IAE relative to distant downstream control with feedforward *ahead*, admitting only settings that move the gates no more than that structure, 3.13 m. Of 45 settings 23 were admitted, and the best, $\lambda = 3$ and $\rho = 0.01$, scored 0.79. Settings with $\rho \geq 10$ and small $\lambda$ ring the reaches by up to 64 cm, most likely because the estimate then follows the waves. Nothing was retuned afterwards for any case.

The three PI loops track the applied flows every scan while the MPC acts, so a rejected plan hands the gates to them without a bump until a later plan solves and resumes control. Three forced rejections with the chop off move the gates by $1.31 \times 10^{-3}$ m$^3$/s at the hand-over, and none of the measured runs had a rejected plan. A switch between distant downstream control and the MPC moves a gate flow by at most 0.019 m$^3$/s. From local upstream control the flows are held until the first plan, which then moves each gate by 3.6 to 6.7 m$^3$/s, its answer to the levels the upstream loops left. The opening limit, the inversion, the weir, the turnout cut-back and the freeboard alarm act under the MPC as under the loops. The controller reads the three transmitters, the applied flows, the head across each gate, the orders its preview allows, the setpoints, the opening limit and the declared values, and replay tests over canals with other widths, lengths, slopes, roughness, gate coefficients, motors, crests and turnouts find identical plans. A plan takes 0.83 to 1.05 ms in Chrome, 0.33 to 0.42 % of one core at 1200 times real time, while the page draws 59.9 to 60 frames per second.

### 5.4.4 Central, Decentralized and Distributed Plans

A second MPC setting, the architecture, decides who plans which gate. Central, the default, is the single program above. Decentralized and Distributed divide it among three agents, one per gate, each with the horizon, blocks, weights, bands, gate limits and preview of the central program. Agent $r$ plans only $Q_r$ and prices only the end of reach $r$, on the rows of reach $r$ in Equation (5.8) with the flow through the gate below replaced by a trajectory $\hat{Q}_{r+1,k}$ that the agent is told, as shown in Equation (5.10).

$$\begin{aligned} \mathscr{J}_r = &\sum_{k=1}^{N} \left( \frac{\hat{y}_{r,k} - y_{SP,r}}{0.01\,\mathrm{m}} \right)^2 + \lambda \sum_{j=0}^{10} \left( \frac{\Delta Q_{r,j}}{0.1\,\mathrm{m^3/s}} \right)^2 \\ &+ \kappa\, \varepsilon_r^2 + \nu\, \varepsilon_r \\ \hat{y}_{r,k+1} = \hat{y}_{r,k} + &\frac{T_s}{A_{s,r}} \Big[ (1 - f_r) Q_{r,k-n_r} + f_r Q_{r,k-n_r-1} \\ &- \hat{Q}_{r+1,k} - q_{r,k} + d_{r,k} \Big], \qquad \hat{Q}_{4,k} = 0 \end{aligned} \tag{5.10}$$

Agent $r$ minimizes $\mathscr{J}_r$ over its own 11 moves, subject to the limits of gate $r$ and the band of reach $r$ from Equation (5.9). When every $\hat{Q}_{r+1}$ equals the plan of the gate below, the local objectives add up to the central one, $\mathscr{J} = \mathscr{J}_1 + \mathscr{J}_2 + \mathscr{J}_3$, which the central program minimizes over all 33 moves at once.

The architectures differ only in $\hat{Q}_{r+1}$, as listed in Table 5.5. A decentralized agent is told the flow that the gate below passes now, held over the horizon, and nothing is exchanged. A distributed agent is told the latest plan of the gate below. The three agents solve at once, exchange their plans and solve

**Table 5.5** What Each Gate's Plan Is Told About the Gate Below

| Architecture | $\hat{Q}_{r+1}$ over the horizon | Programs |
|---|---|---|
| Central | none needed, one plan sets all three flows | 1 of 33 moves |
| Decentralized | the flow it passes now, held | 3 of 11 moves |
| Distributed | its latest plan, exchanged until settled | 8.86 on average |

Programs per decision of scenario 11. The tail agent is told nothing, since no gate lies below it.

again, round $p$ taking the plans of round $p-1$, until no plan an agent was told moves by more than $\delta$, as shown in Equation (5.11),

$$\hat{Q}_{r+1}^{(p)} = Q_{r+1}^{(p-1)}, \qquad \max_{r,k} \left| Q_{r+1,k}^{(p)} - \hat{Q}_{r+1,k}^{(p)} \right| \le \delta \tag{5.11}$$

with $\delta = 1 \times 10^{-3}\,\mathrm{m^3/s}$ and at most 10 rounds. The first round starts from the plans accepted at the previous decision, shifted one period, or from the held flows after a change of architecture, a standby or a refused plan, so one round without earlier plans is the decentralized decision itself. This is the iterative, non-cooperative form of distributed MPC [18, 19]. The agents exchange whole plans and no price, unlike the dual decomposition of Section 1.4.11, which Chapter 6 applies to a heat network.

Because each gate is a flow source in Equation (5.10), the tail agent's problem depends on no other plan, gate 2's only on the tail's, and the head's only on gate 2's. The exchange therefore settles exactly in three rounds, with the tail's plan final after the first, gate 2's after the second and the head's after the third. In scenario 11, 137 of the 144 decisions took three rounds and seven stopped after two, their warm start already within $\delta$. A zero tolerance gives three rounds at every decision and the same run, and the cap of 10 is never reached.

**Pause.** Try Exercise 5.3 before reading on.

The settled plans form a Nash point: each is its agent's best response to the plan of the gate below. No agent prices the reach above its own gate, although its flow leaves that reach at once as the term $\hat{Q}_{r,k}$ in the prediction of agent $r-1$. The central program prices that coupling and the agents do not, which is why a non-cooperative scheme need not reach the central optimum [25]. Computed independently for every decision of scenario 11, by back-substitution from the tail with a separate interior-point solver, the Nash point lies within $9.4 \times 10^{-4}\,\mathrm{m^3/s}$ of the accepted distributed plans over the whole horizon and within $3.4 \times 10^{-4}\,\mathrm{m^3/s}$ in the first move, inside $\delta$. It differs from the central plan by up to 0.128, 0.175 and $0.146\,\mathrm{m^3/s}$ at gates 1, 2 and 3, and judged by the central objective it costs 1.13 times the optimum on average, from 1.004 to 1.32 times. Agents that also price the reach above their gate, minimizing $\mathscr{J}_r + \mathscr{J}_{r-1}$ in turns from the tail, reach the central optimum, within $1.2 \times 10^{-3}\,\mathrm{m^3/s}$ after 20 turns each and $7 \times 10^{-5}\,\mathrm{m^3/s}$ after 40 at the three decisions examined. The whole gap between the Nash point and the central plan is therefore the cross-reach term that no agent prices. At 6.33 h the agents' own costs are 52, 48 and 26 at the Nash point against 27, 33 and 40 under the central plan, which buys reaches 1 and 2 with reach 3.

All three architectures use the Kalman filter described above. Its steady-state gain is exactly block diagonal, the twelve entries that would correct one reach's level or inflow from another reach's transmitter being zero, so the shared filter is already three filters, one per reach, and agent $r$ starts from its own reach's estimate.

A local plan reaches the gates only after a full acceptance check. The solver must report it optimal, the first move must be finite and equal the plan's first step, all 36 planned flows must be finite and inside the gate's limits, and the predictions, the cost and the slack must be finite, the slack non-negative. One refused plan hands all three gates to the tracking PI loops and marks the historian, and the next accepted plans are blended in over ten scans. Seven faults injected into each agent under both

**Table 5.6** Measured Results

| Case and condition | Result |
|---|---|
| 1, head +0.5 m³/s | reach ends up 2 mm after 10, 29, 92 min |
| 2, upstream, farm 2 closes | 9900 m³ spilled in 6 h, reach 3 +56 cm |
| 7 on upstream, tail +1 m³/s | tail farm 5300 m³ short in 6 h |
| 3, downstream, farm 2 closes | no spill, upper reaches +10 cm |
| 3, same, release over 6 h | 28 300 m³ less than case 2 |
| 4, $K_c = 3$, tail +1 m³/s | tail −14 cm, reach 1 −23 cm |
| 4, $K_c = 15$ | worst 6.7 cm |
| 5, $K_c = 20$, no filter | ring 1–3 cm, gate travel 10.6 m in 5 h |
| 5, $K_c = 25$, no filter | grows to 34 cm, gates shut to open |
| 6, $K_c = 20$, $\tau_f = 400$ s | ring <1 cm, gate travel 1.2 m |
| 7, feedback only | reach ends 7, 6, 6 cm |
| 7, feedforward now | reaches 1 and 2 about 10 cm |
| 8, tail order +1.5 m³/s, feedback | 10, 9, 9 cm |
| 8, feedforward now | 16, 15, 8 cm |
| 8, feedforward ahead | 2.3, 1.9, 2.0 cm |
| 8, lead 0.5 / 2 travel times | about 6 / 10 cm |
| 9, three overlapping orders, ahead | about 2 cm |
| 3, MPC, preview now | no spill, 4.8, 5.8, 4.1 cm |
| 7, MPC, preview now | 3.2, 4.2, 5.6 cm |
| 8, MPC, preview schedule | 1.6, 2.0, 2.9 cm |
| 9, MPC, preview schedule | 2.0, 2.3, 2.9 cm |
| 10, downstream, ahead | 5.4, 5.7, 3.0 cm |
| 10, MPC, preview schedule | 2.0, 2.3, 2.8 cm |

architectures, 42 runs with the chop off, were each refused with its reason. The hand-over moved a gate by $1.58 \times 10^{-2}$ m³/s under Decentralized, against the loops’ own next move of $1.70 \times 10^{-2}$ m³/s, and by $5.7 \times 10^{-4}$ m³/s under Distributed, and a take-back of 0.218 m³/s arrived as ten steps of at most 0.0218 m³/s. Central keeps its acceptance as described above.

A decision costs one program under Central, three under Decentralized and 8.86 on average under Distributed in scenario 11, at 2.95 rounds. In Chrome on the built application at 1200 times real time, over three sessions on a shared host, the whole controller takes 0.39 to 0.59 % of one core under Central, 0.27 to 0.31 % under Decentralized and 0.35 to 0.59 % under Distributed, at 60.1 frames per second, and the page’s run equals a headless run at every one of 602 logged instants.

## 5.5 Results

Table 5.6 gathers what cases 1 to 10 produced on the engine at their shipped seed. Each case carries the briefing and completion rule of Section 1.6.4, and the challenge is judged on the case as given, so a changed model parameter or disturbance voids the attempt. Every case except 1 and 4 carries the 4 mm level chop, and over five further seeds no comparison below changes order, except the six named in Section 5.5.5, while Section 5.5.6 reports its own counts over ten seeds. Level deviations are peaks at the 5 s physics step. The 30 s record that the application displays matches them to 0.02 cm, except for reach 1 in case 8, where it reads 2.2 cm against 2.3 cm.

### 5.5.1 Timing the Wave

In manual, a 0.5 m³/s step in the head release raises the end of reach 1 by 2 mm after about 10 min, a little under its 11 min wave travel time, the end of reach 2 after about 29 min and the end of reach 3 after about 92 min, counted from the instant the head gate is commanded. The delay grows faster than the distance because each fixed gate passes only the part of the change that its rising upstream level can push through, so each reach has to fill a little before the next one hears about it. Two hours after the step, gate 3 passes only 0.05 m³/s of the extra release and reach 3 has risen 0.4 cm.

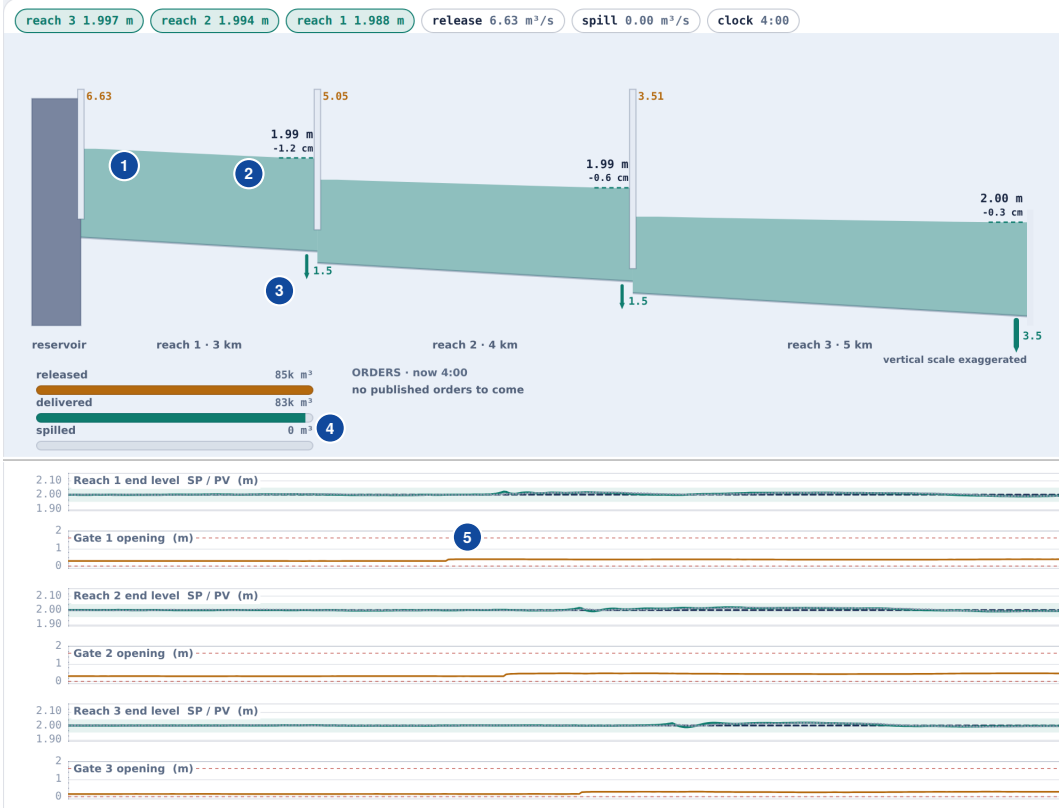



**Figure 5.2** The canal running scenario 8, distant downstream control with a 1.5 $m^3/s$ order for the tail farm published to start at two hours and released one travel time ahead at each gate, four hours into the run. The figure joins two parts of the application's screen, the long section above and the trends below, and leaves out the control and Learn panels. The long section is drawn with an exaggerated vertical scale and shows the reservoir, the three reaches with their water surfaces, the gates with their flows (1), the level at each reach end with its deviation from setpoint (2), the farm takes (3) and the escape weir, over the water ledger of released, delivered and spilled volumes (4). The trends show the three reach end levels held within about 2 cm of setpoint (a worst level error of 2.2 cm in the 30 s record, Section 5.5) while the three gate openings step up one after another (5), the head gate first, as the order is released ahead of its start (by the record at 1.26 h, gate 2 at 1.45 and gate 3 at 1.70 h). At the instant shown the gates pass 6.63, 5.05 and 3.51 $m^3/s$ and nothing has spilled.

> **TRY IT IN THE APP** **1 · Open the Head Gate and Follow the Wave**
>
> Load the scenario, press Run and let half an hour pass, then raise Head gate release from 5.0 to 5.5 $m^3/s$, as the instructions ask. Watch the three level trends. The end of reach 1 first rises 2 mm after about 10 min, reach 2 after about 29 min and reach 3 only after about 92 min.

### 5.5.2 Two Philosophies

Local upstream control holds the upper reaches close to setpoint, and the cost appears at the tail. When the reach 2 farm closes its turnout, reach 1 stays within 0.2 cm, reach 2 peaks 7.1 cm high 35 min later and is back within 1 cm by 3.3 h, while reach 3 rises about 56 cm and about 9900 $m^3$ spills in the first six hours. When the tail farm takes 1 $m^3/s$ more without ordering it instead (case 7 run on local upstream control), the tail farm goes about 5300 $m^3$ short. Distant downstream control

spills nothing when the reach 2 farm closes, lets the upper reaches ride about 10 cm high for a while, and releases about 28 300 $m^3$ less in six hours. The trade is waste against delay, and on a canal the waste is water.

🕮 **Pause.** Try Exercise 5.1 before reading on.

### 5.5.3 Tuning Against the Delay and the Resonance

A gain of 3 $m^3$/s per meter is too timid for the delay. After an unannounced 1 $m^3$/s tail offtake the tail falls about 14 cm, reach 1 falls about 23 cm, below the 1.8 m at which its turnout delivers the full order, so the reach 1 farm goes about 150 $m^3$ short, and the levels are still rolling by up to 19 cm five hours later. At the default gain of 15 the worst excursion is about 6.7 cm.

Raising the gain toward what the delay alone would allow finds the resonance. Driven at the head by a sinusoidal flow while gate 2 passes a constant flow, reach 1 answers 3.7 to 6.8 times more than a pure integrator of the same surface at periods between 10 and 26 min, with peaks near 12 and 22 min, the half and the whole of the 22.6 min round trip of a wave [1]. At a gain of 20 with no filter the levels ring by 1 to 3 cm from two to five hours after the offtake, and the gates travel about 10.6 m in the five hours after it. The ring decays by about a quarter each hour without the level chop, but the 4 mm chop keeps reach 1 ringing by about 1 to 2 cm through a 24 h run. Started at 25, the loop is unstable, the reach 1 ring grows from 15 cm in the hour after the offtake to 34 cm four hours later, and gates 2 and 3 swing between shut and fully open. A 400 s level filter, placed between the loop crossover and the resonance, reduces the ring below 1 cm and the gate travel to about 1.2 m. Most of the unfiltered travel is the level chop passed through the gain, and without the chop the gates travel 3.9 m unfiltered and 0.8 m filtered. At 25 with the filter the loop no longer runs away. Started at 25, the reach 1 ring, about 12 cm in the hour after the offtake, decays to about 1.5 cm four hours later, and raised to 25 at three hours, as scenario 6 asks, every reach end stays within about 1 cm. Without the chop, the gain at which the ring stops decaying rises from about 22 to about 30 with the filter. A 600 s filter rings more than a 400 s one, 1.7 against 0.8 cm in reach 1 from two to five hours after the offtake, most likely because its phase lag at the crossover begins to cost more margin than it saves at the resonance.

### 5.5.4 Feedforward Early and on Time

An unannounced offtake can only be answered after it is seen. Volume 3, Chapter 6 measures its disturbance where it enters, ahead of the delay. A turnout meter reads an offtake where the water leaves, behind every gate, so it gives the gates no lead. Feedback alone holds the three reach ends to about 7, 6 and 6 cm, and adding feedforward from the measured offtakes, with every gate moving at once, improves reach 3 slightly and makes reaches 1 and 2 worse, about 10 cm each. Every gate moves together, so each upper reach gives up a travel time of water to the reach below before its own extra supply arrives.

🕮 **Pause.** Try Exercise 5.2 before reading on.

A published order changes the picture, as shown in Figure 5.2. For a 1.5 $m^3$/s order at the tail, feedback alone moves the reach ends by about 10, 9 and 9 cm, and releasing the order at every gate when it starts makes the upper two worse, 16 and 15 cm, with 8 cm at the tail. Released one travel time ahead per reach, the same order moves the levels by only 2.3, 1.9 and 2.0 cm. Half the lead or twice the lead gives about 6 and 10 cm, so the lead that works is the wave travel time and not a tuning parameter to be searched. Three overlapping orders scheduled the same way hold all three reach ends to about 2 cm.

### 5.5.5 One Plan or Three Loops

Scenario 10 publishes two order changes, the tail farm up 1 $m^3$/s at 2 h and, at 5 h, the reach 1 farm down 1 $m^3$/s while the reach 2 farm goes up by as much, and at 7 h the tail farm takes 0.5 $m^3$/s more without asking. The MPC with the Schedule preview, as the scenario loads it, is compared with distant downstream control and the feedforward *ahead*, which receives the same published orders, over

**Table 5.7** Scenario 10 in Two Windows. IAE in m s, Worst Level Error in cm, Gate Opening Travel in m Over 12 h

| | 0–12 h | | 0–7 h | | 7–10 h | | |
|---|---|---|---|---|---|---|---|
| Structure | IAE | Worst | IAE | Worst | IAE | Worst | Travel |
| MPC, schedule | 581 | 2.84 | 269 | 2.33 | 275 | 2.84 | 1.20 |
| Each reach alone[a] | 1065 | 5.37 | 649 | 5.37 | 385 | 3.90 | 1.51 |
| Downstream, ahead | 720 | 5.66 | 267 | 1.56 | 421 | 5.66 | 3.51 |
| Same, $\tau_I = 7200$ s[b] | 620 | 5.65 | 244 | 1.54 | 337 | 5.65 | 3.47 |
| MPC, off | 5084 | 9.80 | 2692 | 9.80 | 1756 | 9.76 | 1.00 |
| Downstream, feedback | 1664 | 6.93 | 1182 | 6.93 | 432 | 4.12 | 2.85 |

[a] A diagnosis copy of the MPC, not a setting of the application, in which each reach's model replaces the flow through the gate below by the orders below it, so no plan for one gate accounts for the others. It is not the Decentralized architecture of Section 5.4.4, which holds the gate below at its present flow. [b] The best of 225 fixed loop settings by 12 h IAE among the 197 that neither spill nor leave an order short, chosen after the run ($K_c$ from 5 to 25, $\tau_I$ of 1800, 3600 or 7200 s, $\tau_f$ of 0, 200 or 400 s, feedforward off, *now* or *ahead* at $k_{lead}$ of 0.75, 1 or 1.25). It ties at 620 with $K_c = 10$, $\tau_I = 7200$ s, $\tau_f = 0$ and $k_{lead} = 0.75$, which reaches 5.13 cm with 12.70 m of travel.

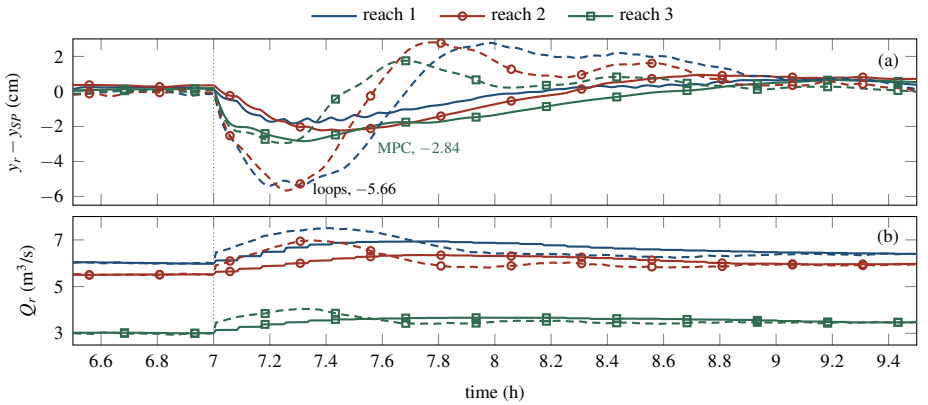



**Figure 5.3** Scenario 10 from 6.5 to 9.5 h under the MPC with the Schedule preview (solid) and distant downstream control with the feedforward *ahead* (dashed). (a) Level error at each reach end and (b) gate flow, reach 1 without markers, reach 2 with circles and reach 3 with squares, with a dotted mark at the unannounced 0.5 m$^3$/s tail offtake at 7 h. The loops add the offtake at every gate from the next scan and reaches 1 and 2 fall 5.41 and 5.66 cm. The plan raises the three gates together over about 40 min and reach 3 falls furthest, 2.84 cm.

12 h at the same seed and chop, with the loops at their default tuning and the MPC at the scenario's settings. In this subsection the worst error is read from the 30 s record, as the scorecard reads it, and IAE is the scorecard's integral of absolute level error summed over the three reaches. The MPC rows of Table 5.6 agree with the physics step to 0.01 cm. The MPC made 144 plans in each 12 h run and none was rejected. Table 5.7 splits the run at the surprise.

Before the surprise the two structures tie on IAE, 269 against 267, and the plan's worst error is the larger, 2.33 against 1.56 cm. The whole lead of the MPC over 12 h, 581 against 720, comes after 7 h, where it holds 275 against 421 and 2.84 against 5.66 cm. The turnout meters tell both structures of the offtake at once. Every loop then adds the whole 0.5 m$^3$/s to its own gate, so each upper reach gives up water before its own extra supply arrives, the mechanism of case 7, while the plan ramps the head gate from 5.98 to 6.93 m$^3$/s by 7.67 h and spreads the shortfall over the reaches, as shown in Figure 5.3. The plan also moves the gate openings 1.20 against 3.51 m. Over ten seeds the MPC has the lower IAE and the lower worst error over 12 h at all ten, 585 against 728 on average and 2.8 against 5.6 cm. Before 7 h its worst error is the larger at all ten, 2.3 against 1.8 cm on average, although its IAE is the lower at eight of ten, so the IAE tie at the default seed is not general. After 7 h

**Table 5.8** Every Offtake Case Over 12 h: IAE in m s / Worst Level Error in cm, From the 30 s Record

| Case (MPC preview) | Loops, same information | MPC | Each reach alone[a] | Best fixed setting[b] | MPC, off | Loops, feedback |
|---|---|---|---|---|---|---|
| 3, farm 2 closes, unannounced (now) | 1194 / 11.46 | 406 / 5.84 | 388 / 8.10 | 459 / 9.36 | 3099 / 12.91 | 1141 / 9.73 |
| 7, tail +1 $m^3$/s, unannounced (now) | 797 / 10.32 | 543 / 5.59 | 652 / 5.67 | 698 / 8.64 | 2500 / 9.87 | 1064 / 7.32 |
| 8, tail order +1.5 $m^3$/s (schedule) | 383 / 2.15 | 396 / 2.88 | 824 / 8.37 | 278 / 5.11 | 3644 / 15.13 | 1499 / 10.01 |
| 9, three orders (schedule) | 412 / 2.15 | 501 / 2.88 | 985 / 8.37 | 279 / 5.13 | 5475 / 15.13 | 1713 / 10.01 |
| 10, one plan or three loops (schedule) | 720 / 5.66 | 581 / 2.84 | 1065 / 5.37 | 620 / 5.65 | 5084 / 9.80 | 1664 / 6.93 |
| 12, challenge, MPC as loaded (schedule) | 741 / 6.07 | 668 / 3.28 | 1027 / 4.27 | 579 / 5.39 | 4198 / 11.49 | 2862 / 7.32 |

The loops with the same information use the feedforward *now* where the MPC preview is Now and *ahead* where it is Schedule, at the default tuning, which is also the challenge's suggested configuration. [a] The diagnosis copy of Table 5.7, not a setting of the application. [b] Chosen per case after the run from the same 225 settings, among those that neither spill nor leave an order short, a bound rather than a design. In cases 9 and 12 it moves the gate openings 6.74 and 12.87 m, against 1.16 and 1.46 m for the MPC.

both are lower at all ten.

The lead after 7 h is partly the comparators'. The best fixed loop setting, the default tuning with $\tau_I$ doubled, holds 337 after the surprise and beats the MPC before it, 244 against 269 and 1.54 against 2.33 cm, although its IAE there is the higher at one of five further seeds. Over 12 h it ties with a second setting at 620 against the MPC's 581. None of the 225 settings beats the MPC on both IAE and worst error, and none that spills nothing moves the gates as little. What only the joint plan delivers is the shared worst error after the surprise, 2.84 cm against 3.90 cm for the same model planned reach by reach and 5.65 cm for the best tuning, and that order holds at all five further seeds. The 3.90 cm belongs to the diagnosis copy, whose gates pass the orders below them on. The Decentralized architecture that the application offers, whose gates instead hold the flow of the gate below, lets reach 1 fall 7.01 cm after 7 h in the same scenario, and the Distributed architecture lets reach 2 fall 3.36 cm (Section 5.5.6).

The other unannounced cases, in Table 5.8, show the same split. On the reach 2 farm's closure the MPC with the Now preview scores 406 against 1194 for the loops with the feedforward *now*, but most of that margin is the comparators'. The same model planned reach by reach scores 388, and the order of those two changes over five further seeds, while the best fixed setting scores 459. The joint plan's own contribution is the shared worst error, 5.84 against 8.10 cm per reach and 9.36 cm for the best setting, lower at all five further seeds. On the unannounced 1 $m^3$/s tail offtake the joint plan keeps part of the total, 543 against 652 per reach and 698 for the best setting, while the worst errors tie, 5.59 against 5.67 cm, an order that changes at one of five further seeds. With the same Now information in scenario 10 the MPC scores 937 against 1361 and 5.63 against 10.53 cm.

On published orders alone the MPC loses to gate stroking. Case 8 gives 396 against 383 and 2.88 against 2.15 cm, and case 9 gives 501 against 412 and 2.88 against 2.15 cm. The worst error is the larger at all five further seeds in both cases, and so is the IAE of case 9, while the IAE order of case 8 changes at three of five. The cause is the model. Its reaches are integrators behind dead times of 215 to 330 s with no waves, while the wave first arrives 580 to 980 s after a gate moves, so the plan meets a stepped order with a ramp. For the 1.5 $m^3$/s tail order of case 8 the plan's gate flows pass half their change 19.5, 14.5 and 4.5 min before the order starts, against 44.5, 33.0 and 18.0 min for gate stroking, which steps each gate once at the wave's lead. Reach 3 therefore fills ahead of the turnout, 2.88 cm high as the order starts and 2.76 cm low five minutes later, and reaches 1 and 2 fall 1.6 and 2.0 cm about 23 min after it. Nothing tried removes the peak. A move price $\lambda$ of 0.1, 0.3, 1 or 10 raises case 8's worst error to 4.98, 4.57, 3.91 or 4.47 cm, a decision every 150 s to 6.15 cm, every 60 s makes the plan unstable, and model dead times halfway to the wave travel times or equal to them

raise it to 18.3 and 76.7 cm. None of these was adopted, and the MPC carries the loss as measured.

Told nothing about the orders, the MPC loses every case to the feedback-only loops, by a ratio of IAE from 2.35 to 3.20 on cases 3, 7, 8, 9 and 10 and 1.47 on the challenge. The preview Off has only its estimate of the unexplained inflow, and with $\rho = 0.01$ that estimate is slow. It is offset-free. Over 48 h every reach end settles within 0.03 cm of its setpoint and each estimate within 0.007 $m^3/s$ of minus its order, but in scenario 10 the worst error falls below 1 cm only after about 20 h. The Off loss is a tuning trade, made on a case with published orders.

> ✦ DID YOU KNOW? **A Planner Told Nothing Does Worse Than Feedback**
>
> From the simulator: in scenario 10 the MPC with its preview Off accumulates an IAE of 5084 m s and a worst level error of 9.80 cm over 12 h, against 1664 and 6.93 cm for three feedback-only loops (Table 5.7). A plan is only as good as what it is told about the orders.

Reach 3 has the larger IAE under the MPC than under the loops with the same information in every case, 132 against 51 on the reach 2 farm's closure, 236 against 162 on the tail offtake, 175 and 195 against 91 in cases 8 and 9, 225 against 140 in scenario 10 and 265 against 147 on the challenge. The tail is the joint plan's buffer. The plan prices every reach end alike, so a volume error shared by the canal spreads roughly in proportion to the reach surfaces, 1 to 1.35 to 1.75, and on the tail offtake the three reach ends fall 3.24, 4.16 and 5.59 cm. Planned reach by reach, reach 3 returns toward the loops, 34 on the closure, 147 on the tail offtake and 169 in scenario 10.

Water is not the price of any of this. Every structure that releases on demand releases the same water within 0.1 %, 264628 against 264 617 $m^3$ in scenario 10, and neither the MPC nor any distant downstream row spills or leaves an order short in any case. Local upstream control with its fixed supply spills or leaves an order short in every case, from 24 664 $m^3$ short on the tail offtake to 41 810 $m^3$ spilled on the closure. Asked to hold the tail at 2.4 m, above the 2.3 m weir crest, in case 3, the plan's tail band keeps the level at 2.288 m and spills 362 $m^3$ in 12 h, against 6614 $m^3$ in the first six hours for the feedback-only loops, which hold 2.40 m.

### 5.5.6 One Plan per Gate

Scenario 11 publishes one order change, the tail farm up 1 $m^3/s$ at 2 h, and at 6 h the same farm takes 0.5 $m^3/s$ more without asking. It loads the MPC with the Schedule preview and the Distributed architecture. Each architecture was run for 12 h at the scenario's settings, with the same seed, chop, orders and gates as the loops at their default tuning, the feedforward *ahead* receiving the same published orders. Table 5.9 splits the run at the surprise, with IAE and the worst error read from the 30 s record as in Section 5.5.5, and none of the 144 decisions of any MPC run was refused. The level errors under the three architectures are shown in Figure 5.4.

On the published order Distributed holds every reach end within 2.18 cm and Decentralized within 4.99 cm, against 1.93 cm for Central. The difference is how early each gate releases. Told only the flow of the gate below now, a decentralized gate follows each move of that gate after it is made, and its own water reaches its reach end a dead time after the reach has given water up. Between 1 and 2 h the decentralized head gate releases up to 0.11 $m^3/s$ less than gate 2 and the reach 1 farm draw, and reaches 1, 2 and 3 fall 4.99, 4.12 and 2.18 cm. Told the plan of the gate below, each distributed gate above the tail releases 0.03 to 0.16 $m^3/s$ more than the gate below and its own farm draw in the same hour, which is gate stroking done by messages, and the reaches fall 1.46, 1.66 and 2.18 cm. After the surprise the exchange carries the tail gate's answer up the canal within the decision, 3.10 against 6.68 cm. What remains is cooperation. At 6.6 h the distributed tail gate draws 3.769 $m^3/s$, what its own reach needs, and reaches 1 and 2 fall 3.10 and 3.09 cm, while the central plan draws 3.631 $m^3/s$, lets reach 3 take the largest error, 2.90 cm, and spares reaches 1 and 2, 2.01 and 2.28 cm. The worst errors of every structure in both windows are redrawn in Figure 5.5.

**Table 5.9** Scenario 11 in Two Windows. IAE in m s, Worst Level Error in cm, Gate Opening Travel in m Over 12 h

| | 0–6 h | | 6–10 h | | 0–12 h | |
|---|---|---|---|---|---|---|
| Structure | IAE | Worst | IAE | Worst | IAE | Travel |
| MPC, central | 183 | 1.93 | 281 | 2.90 | 490 | 0.94 |
| MPC, distributed | 243 | 2.18 | 281 | 3.10 | 550 | 1.26 |
| MPC, decentralized[a] | 563 | 4.99 | 473 | 6.68 | 1067 | 1.78 |
| Distributed, one warm round[b] | 271 | 2.18 | 341 | 4.88 | 641 | 1.40 |
| Central, 1 h horizon | 327 | 2.65 | 324 | 3.04 | 677 | 1.09 |
| Distributed, 1 h | 257 | 2.08 | 284 | 3.08 | 568 | 1.28 |
| Decentralized, 1 h | 526 | 5.85 | 475 | 6.60 | 1030 | 1.72 |
| Downstream, ahead | 230 | 1.52 | 390 | 5.31 | 647 | 3.01 |
| Downstream, now | 718 | 10.53 | 398 | 5.67 | 1143 | 3.93 |
| Downstream, feedback | 944 | 6.93 | 480 | 3.47 | 1458 | 2.58 |
| Local upstream[c] | 3529 | 41.01 | 7643 | 58.48 | 15452 | 1.38 |

[a] Also the Distributed architecture with one round and no earlier plans, identical in every logged column. [b] One round per decision, starting from the plans of the decision before. [c] The tail farm goes 29 989 $m^3$ short.

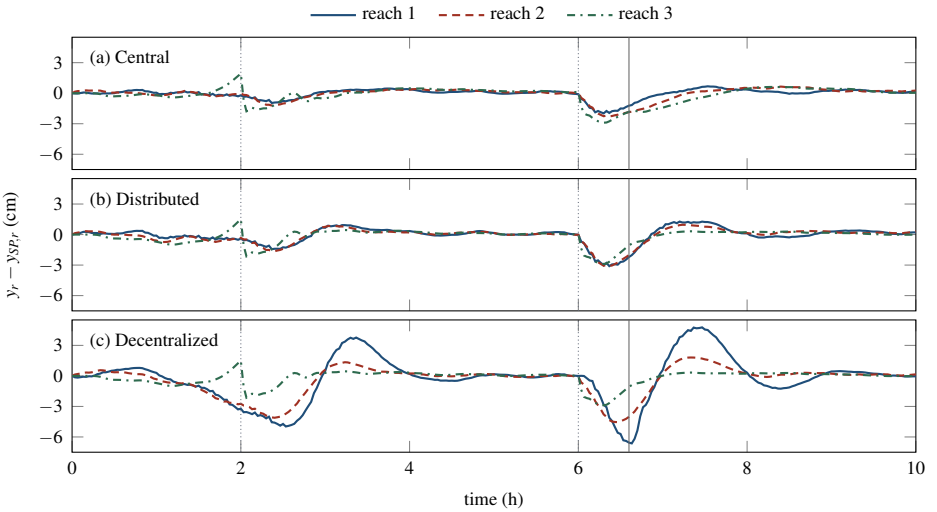



**Figure 5.4** Scenario 11 from 0 to 10 h under (a) Central, (b) Distributed and (c) Decentralized, level error at each reach end, reach 1 solid, reach 2 dashed and reach 3 dash-dotted. Dotted lines mark the published tail order at 2 h and the unannounced offtake at 6 h. At the thin solid line, 6.6 h (23 760 s), reach 1 stands at 1.988, 1.978 and 1.934 m, and each plan predicts a worst deviation over its next 3 h of 2.2, 2.1 and 4.9 cm. Drawn from the engine's 30 s record every 2 min, since three captures of the application would be illegible at this size.

Over ten seeds Distributed beats Decentralized at all ten in both windows, and Central beats Distributed at nine of ten before the surprise and eight of ten after it. An independent set of ten seeds gives 10, 10, 8 and 8 of ten. The loops with the feedforward *ahead*, given the same published orders, keep the smallest peak before the surprise at every seed of both sets, 1.65 cm on average in the first set against 2.03 for Central and 2.38 cm for Distributed, and lose to Central and Distributed after it at every seed, 5.13 against 2.95 and 3.10 cm. On the published-order peak no architecture of the MPC beats gate stroking, as the central plan does not in Section 5.5.5.

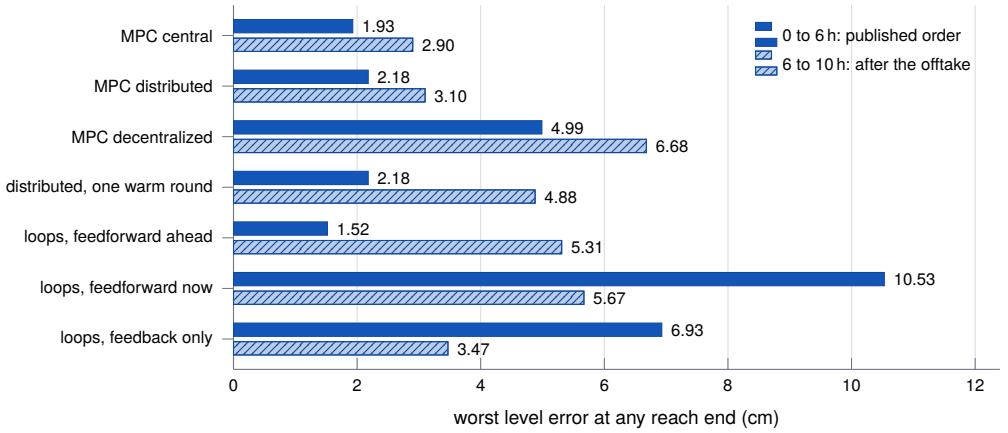



**Figure 5.5** Scenario 11 redrawn from Table 5.9: the worst level error at any reach end on the published order (dark bars) and after the unannounced offtake (hatched bars). The 1 h horizon variants and local upstream control, 41.01 and 58.48 cm, are left out. Notice that gate stroking by the loops holds the smallest error on the published order, that Central and Distributed hold the smallest after the surprise, and that the decentralized agents, which exchange nothing, trail both in each window.

**TRY IT IN THE APP** **11 · Central, Decentralized or Distributed MPC**

Load the scenario and run it nine hours with MPC architecture on Distributed. Every reach end stays within 2.18 cm on the published order and within 3.10 cm after the unannounced offtake at 6 h. Reset, set MPC architecture to Decentralized before Run and repeat: 4.99 cm and 6.68 cm.

The counterfactuals attribute the gain. Removing the exchange is the Decentralized architecture, and one round per decision without earlier plans reproduces Decentralized in all 78 logged columns other than the architecture's own and the wall clock, in scenarios 10, 11 and the challenge. One warm round per decision keeps the published-order result, 2.18 cm, but gives up 4.88 cm after the surprise, since news of the offtake then climbs one gate per decision, and capping the exchange at two rounds leaves 137 decisions unsettled and 4.11 cm. The agents plan over the central horizon, so the central controller at the local horizon is Central itself. With all three shortened to 1 h, Distributed still beats Decentralized in both windows, 2.08 against 5.85 and 3.08 against 6.60 cm, so the horizon does not cause the exchange's gain. The shorter horizon reverses the order of Central and Distributed before the surprise, 2.65 against 2.08 cm, as Central loses 0.72 cm and Distributed gains 0.10 cm, while after it Central still leads, 3.04 against 3.08 cm.

The other cases tell the same story. On the reach 2 farm's closure, the unannounced tail offtake and the published orders of cases 8 and 9, with the previews of Table 5.8, the worst error over 12 h is 5.84, 5.59, 2.88 and 2.88 cm under Central, 8.14, 5.68, 3.38 and 3.38 cm under Distributed and 11.00, 10.71, 8.42 and 8.37 cm under Decentralized. In scenario 10 after the surprise the diagnosis copy of Table 5.7, whose gates pass the orders below them on, reaches 3.90 cm, while the shipped Decentralized reaches 7.01 cm. With the loaded challenge switched to the MPC, Central and Distributed meet it, 3.28 and 3.35 cm all day and 1.37 and 1.46 cm over the last four hours, and Decentralized misses at 8.56 and 4.73 cm, as listed in Table 5.11.

The safety comparison covers the twelve scenarios and thirteen stress cases under eight structures, with the worst rows in Table 5.10. No architecture overtops the bank, and neither local architecture spills where Central does not, 345 against 362 $m^3$ with the tail asked above the crest. Under the MPC, orders go short only in the stress cases. When every farm asks for 4 $m^3/s$ at once, unannounced at 1 h, Central leaves 293 $m^3$ short, Distributed 1131, Decentralized 1964 and the loops with the feedforward

**Table 5.10** Worst Cases Under Every Structure

| Case | Cen. | Dec. | Dist. | FB | Ahead | Upstr. |
|---|---|---|---|---|---|---|
| All shut, overtopping (s) | 0 | 0 | 0 | 0 | 0 | 33850 |
| Tail asked 2.4 m, spill | 362 | 345 | 345 | 13589 | 13590 | 53564 |
| All at 4, 1 h, short | 293 | 1964 | 1131 | 3583 | 1534 | 233466 |
| All at 4, 6 h, short | 301 | 2968 | 2231 | 5780 | 2289 | 143376 |
| All at 4, 1 h, slow motor, short | 693 | 2170 | 1638 | 3723 | 2137 | 233466 |
| All at 4, published, short | 0 | 112 | 0 | 3517 | 0 | 208266 |
| $h_{min} = 2.0$ m, short | 387 | 701 | 381 | 1508 | 365 | 32026 |
| $B = 4.5$ m, $C_d = 0.5$, short | 9 | 0 | 0 | 1 | 1 | 106743 |
| All at 4, 6 h, lowest (m) | 1.692 | 1.553 | 1.585 | 1.551 | 1.568 | 1.183 |

Volumes in $m^3$ over 12 h. Central, Decentralized and Distributed MPC, distant downstream control with feedback only and with the feedforward *ahead*, and local upstream control. All at 4: every farm asks for 4 $m^3$/s, unannounced at 1 h on case 3, at 6 h on scenario 11, or published to start at 2 h.

**Table 5.11** Challenge Outcomes Over Twelve Hours

| Configuration | Worst, day | Worst, 8–12 h | Short |
|---|---|---|---|
| as loaded, local upstream | 68.0 cm | 68.0 cm | 26 400 $m^3$ |
| downstream, feedback | 7.3 cm | 6.7 cm | 0 |
| downstream, now | 12.2 cm | 4.1 cm | 0 |
| downstream, ahead | 6.1 cm | 1.2 cm | 0 |
| MPC, schedule | 3.3 cm | 1.4 cm | 0 |
| MPC, schedule, distributed | 3.3 cm | 1.5 cm | 0 |
| MPC, schedule, decentralized | 8.6 cm | 4.7 cm | 0 |
| MPC, now | 6.1 cm | 3.7 cm | 0 |
| MPC, off | 11.5 cm | 4.6 cm | 0 |

*ahead* 1534 $m^3$, so Decentralized is then less safe than the loops it would replace. Published two hours ahead, the same demand leaves Decentralized 112 $m^3$ short and every other structure except feedback alone and local upstream control nothing. With the full-order depth raised to the setpoint Decentralized leaves 701 $m^3$ short against 387 for Central and 381 for Distributed. On a canal with narrower gates of lower coefficient, which every controller now sees as mismatch, Central is the one MPC that leaves an order short, 9 $m^3$.

### 5.5.7 The Graded Challenge

The graded case covers twelve hours with five published order changes and a tail farm that takes 0.5 $m^3$/s more without asking at six hours, starting on local upstream control with a 5 $m^3$/s supply. It is met with less than 1 $m^3$ spilled, less than 1 $m^3$ of any order left short, every reach end within 8 cm all day and within 3 cm over the last four hours. Nine configurations were run over the full twelve hours, with the outcomes in Table 5.11.

As loaded, local upstream control with a fixed supply of 5 $m^3$/s leaves the tail farm 26 368 $m^3$ short, about 26 400 $m^3$, and reach 3 ends 68.0 cm low, so the case fails. Distant downstream control delivers every order but misses 3 cm late in the day, releasing at the moment of each order is worse over the whole day than feedback alone, and releasing ahead meets every condition with the last four hours held to 1.2 cm. That suggested configuration, distant downstream control at the default tuning with the feedforward released *ahead* at $k_{lead} = 1$, spills nothing, leaves no order short and stays inside the two bands by 1.9 and 1.8 cm. The unannounced offtake at six hours is the largest remaining excursion, 6.1 cm in reach 1 against 1.8 cm before it, because it is the only change left to feedback alone. Over five further seeds the suggested configuration holds the last four hours to 1.2 to 1.6 cm and feedback alone misses at 6.6 to 6.9 cm.

Changing only the control philosophy of the loaded case to the MPC, with its Schedule preview, also meets the target. Nothing spills, no order goes short, and the worst deviations are 3.3 cm all day

and 1.4 cm over the last four hours, with the gate openings moving 1.46 m against 3.78 m for the suggested configuration. Over the last four hours that is a loss at the shipped seed, 1.4 against 1.2 cm, and over five further seeds the order changes, 1.22 to 1.48 cm against 1.15 to 1.59 cm, with the MPC lower at three. The MPC reads no loop tuning, so the suggested settings with the MPC give the same run. With the Now preview it misses late in the day, 3.7 cm, and with Off it misses at 11.5 cm. The best fixed loop setting chosen after the run, $K_c = 10$, $\tau_I = 7200$ s, no filter and $k_{lead} = 0.75$, has the lower IAE, 579 against 668, with a worst error of 5.39 cm and 12.87 m of gate travel. The challenge, its target and its suggested configuration are unchanged.

## 5.6 Technology in Practice

### What Measures It

An automated canal measures water levels, flows and gate positions at its control structures, and its control equipment must detect an abnormal sensor reading before the controller acts on it [3]. Distant downstream control needs the level at the far end of each pool, away from the gate that answers to it. The three level transmitters of Figure 5.1 stand for these gauges, and the turnout meters for the flow measured at each farm offtake.

> **TECHNOLOGY Sensor: The Pool Level Gauge**
>
> An automated canal reads the level in each pool with a float, pulley and counterweight that turn an absolute shaft encoder, with a bubbler that feeds gas slowly through a tube to an orifice below the lowest water and reads the back pressure, or with a submerged pressure transmitter. Float and pressure gauges reach 0.1 percent of range or better, and a float system spanning 15 m errs by about 0.01 m in its mechanics. The float sits in a stilling well whose small inlet damps wind waves and surges [26, 27]. A silted intake gives erroneous readings, a clogged bubbler tube or a fallen object on the float upsets the gauge, and frost damages a submerged transmitter. The chapter's transmitters carry a 4 mm chop and no such fault.

### What Moves It

In the classic electronic downstream controller a float-driven potentiometer reads the level in the pool below, a second potentiometer driven by the hoist reads the gate position, and raise and lower relays run the gate motor until the gate reaches the position the algorithm has computed [3]. A set-operate-time actuator moves the gate in increments, and a 30 mm step at 7.7 mm/s takes 4 s. Check structures also carry bypass weirs sized for the canal's largest flow in case the gates close while the canal runs full [26]. The chapter's gates are flow sources: each opening follows from inverting Equation (5.3), and only the tail has a weir.

> **TECHNOLOGY Final Element: The Motorized Check Gate**
>
> A check structure holds the level above it with radial gates on larger canals and vertical lift gates on small ones, raised by electric, hydraulic or manual hoists with constant or variable speed motors [3]. Automated hoists are designed to set a gate within about 6 to 15 mm, typically move 0.3 to 0.6 m/min and carry a high duty cycle [26]. The flow under a radial gate follows from both water levels and the gate position, and silt or subsidence shifts that rating until the gate is recalibrated [27]. On a power failure a gate closes or stays put, as designed [3]. The chapter's gates move at 0.24 m/min and invert a fixed declared rating.

### Where the Controller Runs

Local automatic control runs on microprocessor equipment at each check structure. Supervisory control places a remote terminal unit at each site, which reports levels, gate positions and equipment status to a master station and moves its gate on the instructions it receives over a two-way link [3]. Typical scan times of such water systems range from 2 s to 10 min [26]. Gate stroking needs the central system, because every gate must move at a precise time and by the correct amount, and the schedule can be stored in each remote unit for execution [3]. The simulator's loops scan every 30 s

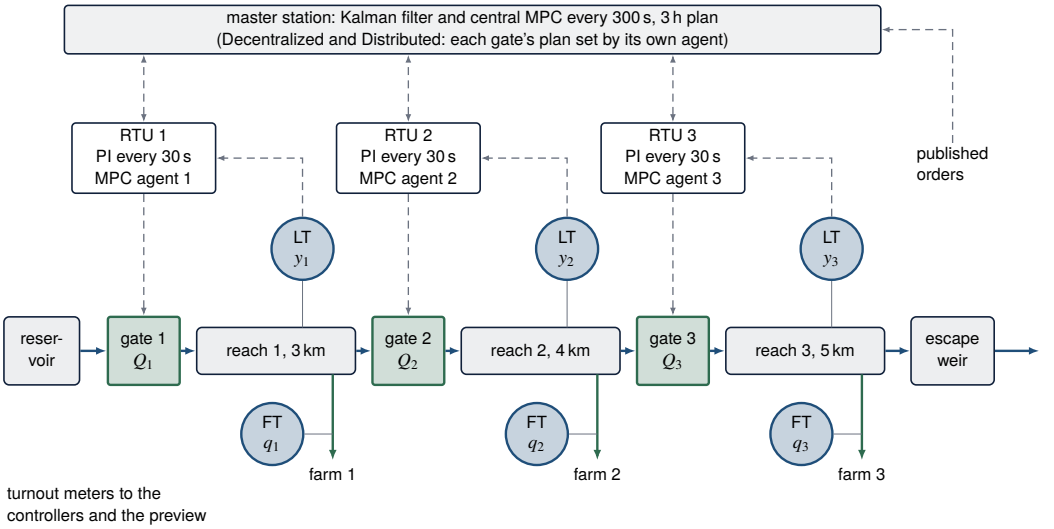



**Figure 5.6** The canal's loops on field hardware, drawn from Sections 5.3 and 5.4. Green boxes are the three gates, blue circles the level transmitters at the reach ends and the turnout meters, dashed arrows signals. In a field system a remote terminal unit at each gate reports to a master station. Notice that each gate answers to the level at the far end of the reach below it, so its water is seen only after a wave has crossed that reach.

and its MPC decides every 300 s, inside that range. Its Central architecture plays the master station, and its Decentralized and Distributed agents play controllers at the gates, as drawn in Figure 5.6.

> **IN PRACTICE** **Filter and Deadband Before Gain**
>
> Canal practice filters the level before the controller to remove wave action and cut the number of gate moves, and the filter keeps the controller from amplifying critical frequencies. Oscillation follows when the pool's natural frequency matches the frequency of control actions, or when each correction is too large for the error. A deadband around the target stops the gate while the level stays inside it [3]. For a proportional-plus-reset controller a filter time constant of 100 s is usually enough [26]. The chapter's loops use 200 s, 400 s against the resonance, and no deadband.

### What Is Modernizing It

Water users have long ordered water 24 to 48 hours ahead and received it on a schedule [3]. On networks fitted with sensors, actuators and a supervisory network, system identification and control have given near on-demand supply, with a distribution efficiency better than 90 percent against 70 percent under manual operation and better than 90 percent of orders met on demand. In simulations of larger systems, decentralized control with feedforward could not be told apart from centralized control in most scenarios, and the centralized control could not be implemented there [28]. This chapter puts the same question to one canal with its three architectures and its Schedule preview, and finds that exchanged plans keep most of what the single plan delivers.

## 5.7 Limitations

The channel is prismatic and rectangular, which puts the resonant periods near their textbook values. Real canals with varying cross-sections and structures resonate at less regular periods [13].

The local-inertia form omits advection. It is accurate at Froude numbers near 0.1 and would not be at the steep, fast flows of mountain canals.

The coarse grid smears fronts and approximates the resonance. A controller designed on this model and deployed on a real reach would need identification of the actual reach, as field work has done [15].

Gate motors are smooth and fast relative to real stepping motors with deadbands.

The MPC's model leaves out the gravity waves and the resonance, the change of the friction slope with flow, the gate motor's lag, the turnout cut-back except at the delivery its meter reads now, and the weir except through the estimated inflow. Its dead times are the best fit over its 3 h horizon rather than the wave's arrival, and the plan is sensitive to them, since case 8's worst error grows from 2.88 to 18.3 cm with dead times halfway to the wave travel times. The model is frozen at the design point, so an edit of the canal is model mismatch, and with the preview Off its estimate needs about 20 h to bring scenario 10 within 1 cm. The preview takes the published orders as exact.

The decentralized and distributed architectures share the central model, estimate, weights and horizon, so they differ only in who plans which gate and what each agent is told. The exchange of plans is instantaneous and never loses a message, while a field network between cross-regulators has delays and dropouts. Only the non-cooperative exchange is offered, and a decision that reaches the cap of ten rounds would be accepted unsettled without a mark in the historian, which the shipped tolerance never reaches because the exchange settles in three.

> **OPEN CHALLENGE** **Plans Shared Over a Patchy Network**
>
> On one automated network each site samples regularly, but the center hears only of events, such as a level that has moved 2 cm up or down. A local controller does not suffer from this. A controller for the whole network must cope with information that arrives at varying rates and delays, and its designers judged that substantial further research was needed [28]. This chapter's agents exchange plans at once and never lose one. Progress would be plans that stay safe and close to the central optimum when messages arrive late or not at all.

The farms either publish their orders exactly or not at all, and the canal has no branching, seepage or evaporation.

> **OPEN CHALLENGE** **Which Gate Should Listen to Which Level**
>
> Which inputs a decentralized canal controller should pair with which outputs has had too little attention, and the possible structures are combinatorially many, so heuristics are needed. How a disturbance travels along a string of pools is open too, in particular how much feedforward speeds the response before it amplifies the disturbance [28]. This chapter compares one plan, three silent plans and three exchanged plans on three pools. Progress would be rules that pick the structure for long chains of pools.

## 5.8 Conclusion

A canal reach is a tank with a delay in front of it and a resonance inside it, and the simulator of this chapter lets each of the three be measured. A 0.5 $m^3/s$ step at the head reaches the three reach ends after about 10, 29 and 92 min. Local upstream control spills about 9900 $m^3$ in six hours when a farm closes, while distant downstream control spills nothing and lets the upper reaches ride about 10 cm high. A gain tuned to the delay rings at the reach resonance and runs away at a slightly higher gain, and a 400 s level filter tames it. Feedforward that releases a published order at every gate at once makes the upper reaches worse, 16 against 10 cm, while releasing it one wave travel time ahead per reach moves the levels only about 2 cm, and over a twelve-hour challenge the same schedule holds the last four hours to 1.2 cm. A model predictive controller, offered as an option, that sets the three gates in one plan spreads an unannounced offtake over the reaches, 2.84 cm against 5.66 cm for the loops and 3.90 cm for the same model planned reach by reach with the orders below each gate passed on, and trails gate stroking on published orders because its model has no waves. Divided into one plan per gate, the same controller keeps most of that result only when the gates exchange their plans,

3.10 cm against 6.68 cm without the exchange and 2.90 cm for the single plan after the offtake of scenario 11, and the remainder is the cooperation that no gate's own objective prices.

The model uses prismatic reaches, a coarse grid and perfect published orders. A fuller model would add trapezoidal sections, gate deadbands, uncertain orders and a branching network. The canal remains a compact case of controlling water that arrives late, stays, and echoes.

**KEY POINTS**

- A canal reach is a tank with a delay in front of it and a resonance inside it.
- When a farm closes, local upstream control spills about 9900 $m^3$ in six hours and distant downstream control spills nothing.
- Released one wave travel time ahead per reach, a published order moves the levels about 2 cm, and the MPC, whose model has no waves, trails this gate stroking.
- One plan for three gates holds an unannounced offtake to 2.84 cm, against 5.66 cm for the loops. Split per gate, the plans keep most of that only when exchanged, 3.10 cm against 6.68 cm.

**Where else this applies.** Transport with storage and reflection also runs district heating lines (Chapter 6) and gas pipelines (Volume 10, Chapter 12).

## References

[1] X. Litrico and V. Fromion, "Simplified Modeling of Irrigation Canals for Controller Design," *Journal of Irrigation and Drainage Engineering*, vol. 130, no. 5, pp. 373–383, 2004.

[2] J. Schuurmans, A. J. Clemmens, S. Dijkstra, A. Hof, and R. Brouwer, "Modeling of Irrigation and Drainage Canals for Controller Design," *Journal of Irrigation and Drainage Engineering*, vol. 125, no. 6, pp. 338–344, 1999.

[3] U.S. Bureau of Reclamation, *Canal Systems Automation Manual, Volume 1*, 1st ed. Denver, CO: U.S. Department of the Interior, Bureau of Reclamation, 1991. [Online]. Available: https://www.usbr.gov/tsc/techreferences/mands/mands-pdfs/CanalSysAuto1.pdf

[4] P.-O. Malaterre, D. C. Rogers, and J. Schuurmans, "Classification of Canal Control Algorithms," *Journal of Irrigation and Drainage Engineering*, vol. 124, no. 1, pp. 3–10, 1998.

[5] E. B. Wylie, "Control of Transient Free - Surface Flow," *Journal of the Hydraulics Division*, vol. 95, no. 1, pp. 347–362, 1969.

[6] W. E. Bodley and E. B. Wylie, "Control of Transients in Series Channel with Gates," *Journal of the Hydraulics Division*, vol. 104, no. 10, pp. 1395–1407, 1978.

[7] I. Mareels, E. Weyer, S. K. Ooi, M. Cantoni, Y. Li, and G. Nair, "Systems engineering for irrigation systems: Successes and challenges," *Annual Reviews in Control*, vol. 29, no. 2, pp. 191–204, 2005.

[8] C. M. Burt and X. Piao, "Advances in PLC-based irrigation canal automation," *Irrigation and Drainage*, vol. 53, no. 1, pp. 29–37, 2004.

[9] D. C. Rogers and J. Goussard, "Canal Control Algorithms Currently in Use," *Journal of Irrigation and Drainage Engineering*, vol. 124, no. 1, pp. 11–15, 1998.

[10] T. S. Strelkoff, "One-Dimensional Equations of Open-Channel Flow," *Journal of the Hydraulics Division*, vol. 95, no. 3, pp. 861–876, 1969.

[11] J. Schuurmans, A. Hof, S. Dijkstra, O. H. Bosgra, and R. Brouwer, "Simple Water Level Controller for Irrigation and Drainage Canals," *Journal of Irrigation and Drainage Engineering*, vol. 125, no. 4, pp. 189–195, 1999.

[12] X. Litrico and V. Fromion, "Tuning of Robust Distant Downstream PI Controllers for an Irrigation Canal Pool. I: Theory," *Journal of Irrigation and Drainage Engineering*, vol. 132, no. 4, pp. 359–368, 2006.

[13] ——, *Modeling and Control of Hydrosystems*. Springer London, 2009.

[14] B. T. Wahlin and A. J. Clemmens, "Automatic Downstream Water-Level Feedback Control of Branching Canal Networks: Theory," *Journal of Irrigation and Drainage Engineering*, vol. 132, no. 3, pp. 198–207, 2006.

[15] E. Weyer, "System identification of an open water channel," *Control Engineering Practice*, vol. 9, no. 12, pp. 1289–1299, 2001.

[16] B. T. Wahlin, "Performance of Model Predictive Control on ASCE Test Canal 1," *Journal of Irrigation and Drainage Engineering*, vol. 130, no. 3, pp. 227–238, 2004.

[17] R. R. Negenborn, P.-J. van Overloop, T. Keviczky, and B. De Schutter, "Distributed model predictive control of irrigation canals," *Networks & Heterogeneous Media*, vol. 4, no. 2, pp. 359–380, 2009.

[18] R. Scattolini, "Architectures for distributed and hierarchical Model Predictive Control – a review," *Journal of Process Control*, vol. 19, no. 5, pp. 723–731, 2009.

[19] P. D. Christofides, R. Scattolini, D. Muñoz de la Peña, and J. Liu, "Distributed model predictive control: A tutorial review and future research directions," *Computers & Chemical Engineering*, vol. 51, pp. 21–41, 2013.

[20] P. D. Bates, M. S. Horritt, and T. J. Fewtrell, "A simple inertial formulation of the shallow water equations for efficient two-dimensional flood inundation modelling," *Journal of Hydrology*, vol. 387, no. 1-2, pp. 33–45, 2010.

[21] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[22] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[23] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[24] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[25] J. B. Rawlings and B. T. Stewart, "Coordinating multiple optimization-based controllers: New opportunities and challenges," *Journal of Process Control*, vol. 18, no. 9, pp. 839–845, 2008.

[26] U.S. Bureau of Reclamation, *Canal Systems Automation Manual, Volume 2*, 1st ed. Denver, CO: U.S. Department of the Interior, Bureau of Reclamation, 1995. [Online]. Available: https://www.usbr.gov/tsc/techreferences/mands/mands-pdfs/CanalSysAuto2.pdf

[27] ——, *Water Measurement Manual*, third, revised reprint ed. U.S. Department of the Interior, Bureau of Reclamation, Water Resources Research Laboratory, 2001. [Online]. Available: https://www.usbr.gov/tsc/techreferences/mands/wmm/

[28] I. Mareels, E. Weyer, S. K. Ooi, M. Cantoni, Y. Li, and G. Nair, "Systems engineering for irrigation systems: Successes and challenges," in *IFAC Proceedings Volumes (16th IFAC World Congress)*, vol. 38, no. 1, 2005, pp. 1–16.

## 5.9 Exercises

**5.1** Concept check · 2 min

Under local upstream control the head releases a fixed supply and each cross-regulator holds the level just upstream of itself. At one hour the reach 2 farm closes its turnout. Where does the water it no longer takes go?

(a) It stays in reach 2, whose level rises until the farm opens its turnout again

(b) It is passed from gate to gate down to the tail and spills over the escape weir

(c) The head gate cuts its release, so the extra water never enters the canal

(d) It backs up into reach 1, because gate 2 cannot pass more than its design flow

**5.2** Concept check · 2 min

In case 7 the tail farm takes 1 $m^3/s$ more without ordering it. With the order feedforward on Now, every gate adds the metered offtake to its flow at the next scan. Feedback alone lets the reach ends fall about 7, 6 and 6 cm. With the feedforward, reach 3 improves slightly and reaches 1 and 2 fall about 10 cm each. Why do the upper reaches get worse?

(a) The turnout meters report the offtake late, so the feedforward acts after the feedback has

(b) The feedforward adds to the integral term, so each loop overshoots after the offtake

(c) Each gate draws its extra flow out of the reach above at once, and that reach's own extra supply arrives a travel time later

(d) The level filter slows the feedback, so the feedforward acts alone through the first hour

**5.3** Concept check · 2 min

In the Distributed architecture the three gate agents exchange plans until no plan an agent was told moves by more than $\delta$. On this canal the exchange settles in three rounds. Why three?

(a) The number of rounds is capped at three to fit inside the five-minute decision period

(b) The Kalman filter's gain is block diagonal, so each agent needs only its own reach's estimate

(c) Each agent also prices the reach above its gate, so the three objectives agree after three swaps

(d) Each agent's problem depends only on the plan of the gate below, so the plans become final one gate at a time from the tail

**5.4** Theory and calculation · 6 min

The three reaches are 3, 4 and 5 km long and run at 2 m depth.

(a) Compute the gravity-wave speed $\sqrt{gh}$ and the wave travel time of each reach in minutes. *Checkpoint:* 4.43 m/s, then 11.3, 15.1 and 18.8 min.

(b) A 1.5 $m^3$/s tail order is published to start at 2 h, as in scenario 8. With $k_{lead} = 1$, at what time does each gate begin to release it according to Equation (5.6)? Give the times in hours. *Checkpoint:* Head gate at 1.25 h, gate 2 at 1.44 h, gate 3 at 1.69 h.

(c) Reach 1 resonates at the round trip of a wave between its two gates. Compute that period and compare it with the periods at which reach 1 answers most strongly. *Checkpoint:* 22.6 min. The peaks near 12 and 22 min are its half and its whole.

**5.5** Theory and calculation · 6 min

The MPC models reach $r$ as an integrator of surface $A_{s,r}$ behind a dead time $\theta_r$, sampled every $T_s = 300$ s (Equation (5.8) and Table 5.3).

(a) Write each dead time as $\theta_r = (n_r + f_r)T_s$ with $0 \le f_r < 1$. *Checkpoint:* $n_r = 0, 0, 1$ and $f_r = 0.717$, 0.983, 0.100.

(b) All flows are balanced when the head gate flow steps up by 0.3 $m^3$/s at decision $k$ and holds. How much does the model raise the end of reach 1 over the period from $k$ to $k+1$, and over each period after that? Compare with the first rise of the real reach in Table 5.3. *Checkpoint:* 1.57 mm, then 5.55 mm per period.

(c) The plan prices the three reach ends alike, so a volume error shared by the canal tends to spread into equal level errors. What share of the volume does each reach take, and how far do all three fall if an unannounced 0.5 $m^3$/s offtake goes unanswered for 30 min? *Checkpoint:* 24.4 %, 33.0 % and 42.6 %, and about 1.35 cm.

**5.6** Theory and calculation · 6 min

Three boundary laws decide who goes short and where water leaves the canal: the turnout cut-back of Equation (5.4) with $h_t = 1.2$ m and $h_{min} = 1.8$ m, and the escape weir $Q_{spill} = 1.7\,B\,(h - 2.3)^{3/2}$ with $B = 6$ m.

(a) At the timid gain of case 4, reach 1 falls about 23 cm from its 2.00 m setpoint. What does the reach 1 turnout deliver at 1.77 m against its 1.5 $m^3$/s order? *Checkpoint:* 1.462 $m^3$/s, 0.038 $m^3$/s or about 2.5 % short.

(b) Feedback-only loops asked to hold the tail at 2.40 m in case 3 hold it there. What does the weir spill at that level, and how much in six hours? The chapter measures 6614 $m^3$ in the first six hours. Why is the measurement lower? *Checkpoint:* 0.323 $m^3$/s and about 6970 $m^3$.

(c) Under local upstream control with every farm shut, the head keeps releasing 5 $m^3$/s. At what tail depth would the weir pass all of it? Compare with the 2.7 m freeboard flag. *Checkpoint:* About 2.92 m, above the flag.

**5.7** App challenge · 6 min

Load scenario 1, *Open the head gate and follow the wave (MANUAL)*, default seed. It starts in Manual with the head release at 5.0 $m^3$/s and no level chop. Press Run, pause at about 0.5 h (the exact instant does not matter), raise Head gate release to 5.5 $m^3$/s, and run at speed 300 until at least two hours after the step, about 2.6 h. Export the CSV, whose time column is in hours.

(a) Take the step as the last row before `gate1_flow` rises. For each reach end, find the first row in which `y1_m`, `y2_m` or `y3_m` stands 2 mm above its value at the step, and give the delay in minutes. *Checkpoint:* About 10, 29.5 and 92.5 min.

(b) Read how much `gate3_flow` and `y3_m` have risen two hours after the step, and compare the delays of (a) with the wave travel times of reaches 1, 2 and 3 taken in turn. Explain the difference. *Checkpoint:* About 0.05 $m^3$/s and 0.4 cm.

Give the delays to 1.5 min.

**5.8** App challenge · 8 min

Load scenario 8, *A published order, released one travel time ahead*, default seed. It runs distant downstream control with Order feedforward on Ahead and Lead, in travel times at 1, with a 1.5 $m^3$/s tail order published to start at 2 h. Run each part to 4 h at speed 1200 and read the Worst level error tile, or the per-reach maxima in the last rows of the CSV columns `metric_maxdev1_m` to `metric_maxdev3_m`.

(a) Run the scenario as loaded. *Checkpoint:* 2.2 cm.

(b) Reset, set Order feedforward to Now before Run, and repeat. *Checkpoint:* 16.2 cm.

(c) Press the reload button to restore the scenario, set Lead to 0.5 before Run, and repeat. Then reload again and repeat with Lead at 2. *Checkpoint:* 6.5 cm and 10.6 cm.

(d) Say why the lead of one travel time wins, and why part (c) needs a reload rather than a Reset.

Give the tile readings to 0.2 cm.

**5.9** App challenge · 9 min

Load scenario 11, *Central, decentralized or distributed MPC*, default seed. It runs the MPC with the Schedule preview and MPC architecture on Distributed (advanced view), with a 1 $m^3$/s tail order published for 2 h and an unannounced 0.5 $m^3$/s tail offtake at 6 h. Run each part to 10 h at speed 1200 and export the CSV, whose time column is in hours. Read the largest of `metric_maxdev1_m` to `metric_maxdev3_m` in the row at 6 h (the worst before the surprise), and the three columns in the last row.

(a) Run the scenario as loaded. *Checkpoint:* 2.18 cm before, then 3.10, 3.09 and 2.89 cm at the end.

(b) Reset, set MPC architecture to Decentralized before Run, and repeat. *Checkpoint:* 4.99 cm before, then 6.68, 4.55 and 2.89 cm at the end.

(c) Reset, set MPC architecture to Central before Run, and repeat. *Checkpoint:* 1.93 cm before, then 2.01, 2.28 and 2.90 cm at the end.

(d) Explain what the exchange of plans buys on the published order, and what only the central plan buys after the surprise.

Give the values to 0.05 cm.

**5.10** Application area · 3 min

A district manager prefers local upstream control because the trends of the upper reaches stay flat and every gate acts on a level it can see. Using the results of cases 2 and 3, state what each philosophy costs on this canal and who pays it.

**5.11** Application area · 4 min

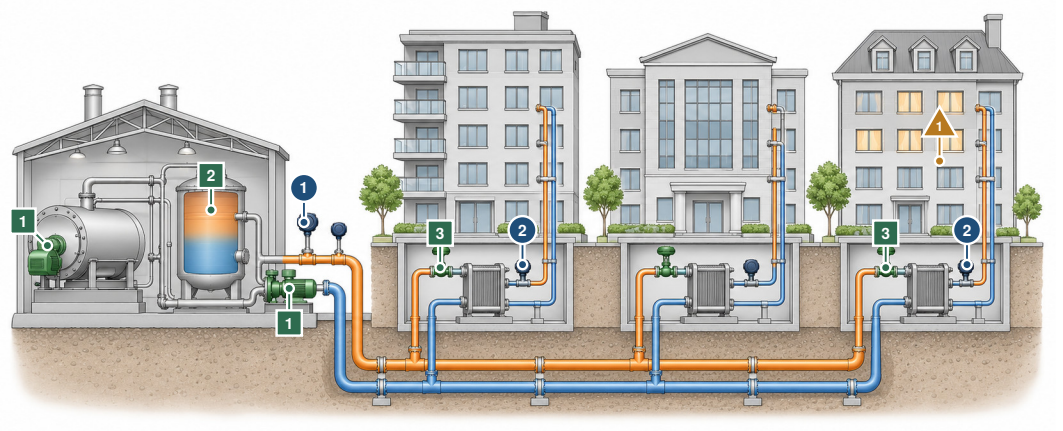
A canal authority plans to run the Distributed architecture in the field, with one controller at each cross-regulator and a radio link between neighbors. Which results of this chapter rest on assumptions the field will not meet, and what would you test before commissioning?

**5.12** Review the agent · 5 min

An AI assistant reviews case 8 under the MPC, which reaches a worst level error of 2.88 cm against 2.15 cm for gate stroking, and writes: “The plan is too sluggish. The move price $\lambda = 3$ penalizes gate moves, so the plan ramps the gates instead of stepping them. Lower $\lambda$ to 0.3 and the MPC will step the gates like gate stroking and beat it.” Find the flaw, and say how you would check the claim in the application.

CHAPTER 6

# Pricing Heat in a District Heating Network Whose Delay Follows the Flow




**The controller moves** 1 heat from the plant and the pump's push 2 storage tank, charged or drawn down 3 each building's valve, setting its draw

**The controller reads** 1 supply temperature and pressure at the plant 2 each building's heat meter

**What disturbs it** 1 cold outdoors and each building's changing load

*Hot water leaving the plant takes hours to reach the far building, and how many hours depends on how much the valves draw. Illustration, not to scale.*

**IN PLAIN TERMS** **Sending Heat Ahead Through a Town's Pipes**

**The main idea.** A heat network pumps hot water from one plant through buried pipes to many buildings. The water takes hours to arrive, and how many hours depends on how much water the buildings draw.

**The challenge.** Heat sent early can be kept in the pipes, a tank and the buildings, but it pays only if cheap heat was spare when it was sent. A posted price for heat can share the work between the plant and the buildings, yet it stops working when the plant runs short.

**Why it matters.** Fuel bought at the wrong hour costs money, and a building left short of heat goes cold.

**THREE WAYS TO READ THIS CHAPTER**

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 6.2, the facts in Sections 6.4 and 6.6, the try-it segments in Section 6.6, and the key points |
| **Study** | 50 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 20 min | the first page, the results (Section 6.6), Technology in Practice (Section 6.7), Limitations (Section 6.8) and the open challenges |

**The control question** In a heat network the flow the consumers draw sets the transport delay, so heat sent ahead of a morning peak arrives when the valves allow. When does sending heat ahead pay, and can a posted price of heat coordinate a plant and six consumers as well as one central plan, even when the plant runs short?

**Learning objectives**

- Compute a consumer's transport delay as the sum of its segments' residences, each at its own flow, and explain why it is not the path's volume over the plant's flow.
- Explain when charging the pipes, the tank and the buildings ahead of a peak saves fuel, and why a boost that arrives on time can still cost more.
- Describe how dual decomposition splits the planning model into a plant problem and six consumer problems coordinated by a price per block.
- Judge from the imbalance and the duality gap whether a price has settled, and explain why a fixed price step fails when a capacity binds or the plant is short.
- Value the heat left in the stores at the end of a run and compare strategies on the graded cost.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.2 (prediction models), 1.4.6 (soft constraints), 1.4.8 (economic MPC and terminal value), 1.4.11 (plans coordinated by a price) and 1.6.4 (how an exercise names a run), and Chapter 5 for plans shared over a network with delays.

**The app** https://apmonitor.com/apps/district-heating/

**See also** Volume 5, Chapter 4 for a delay of hours carried as Lagrangian slugs and seen late by its analyzer. Volume 17, Chapter 6 for a solar field whose transit shows as a lag when its flow is set by hand, and Volume 17, Chapter 2 for one fuel valve whose heat and power arrive on different clocks.



## 6.1 Introduction

Water that leaves the simulated plant at 13:05 with 50 kg/s flowing reaches the hospital at the far end of the town 4.48 h later. The same 15 K step sent at 02:00, when the town's valves draw 78 kg/s, arrives in 2.83 h. The path between them is 5.95 km of pipe holding 598 $m^3$ of supply water, and neither delay equals that volume over the plant's flow (3.2 h at 50 kg/s), because less water passes each junction (35, 19 and 10 kg/s at 13:05) and every segment is crossed at its own flow. The controller sets the flow, and the flow sets the delay.

Two classes of control problem meet in this plant. The first is transport by plug flow through a line whose throughput the controller manipulates, so that the dead time a step test identifies at one throughput is wrong at the next. The second is the allocation of a limited resource among independent users, which a Lagrange multiplier separates into problems each user solves alone [1] and which a coordinating program can drive with prices [2]. Distributed and hierarchical model predictive control (MPC) gives the second class its control form [3, 4], and a heat network adds the first class to it, because the heat a price buys arrives hours after the price is posted.

Two beliefs about heat networks fail on this one. The first holds that a transport delay is a dead time, a property of the pipe. A boost of 15 K sent by hand at 04:00 reaches the hospital at 06:53, on time for the morning peak, and still costs $202 more than the plain supply curve, because the cheap source was already full and the charge was bought from the peak boiler. The second holds that a posted price coordinates as soon as it exists. At the shipped price step, 177 of the 216 control steps of a cold morning settle within 30 rounds, none settles at a step of 0.8, and after the plant's main source trips on the graded cold snap none of the 55 remaining steps settles, because the price that would ration the heat on the same planning problem is about $77,000 per MWh while the iteration reaches $984 at most.

A student meets these pieces in separate places. Network models for simulation and operation describe the delay and the loss [5–8], reviews of network control and predictive control of a network decide the supply temperature [9, 10], and distributed MPC is surveyed as an architecture and analyzed as an algorithm [3, 11], with no plant on which the supply curve, a centralized optimizer and

a price can be run against the same weather. This chapter compares how a heating curve, a hand-timed boost, a centralized MPC and a price-coordinated distributed MPC use a network whose delay follows its flow, and finds where the price stops coordinating, on one declared town in a browser simulator. The plug-flow pipe, the effectiveness relation, the two-capacity building, dual decomposition and the Hankel predictor are established methods. The results are measurements on one town and its seeded weather, among them that the price matches the centralized multiplier wherever it settles and cannot follow it when a capacity binds. The simulated network closes its energy balance to $6 \times 10^{-7}$ of the heat produced, the planning model is linear with the delays frozen at the measured flows, and the dual iteration reports a gap that is the distance between a feasible plan and a proven lower bound. Each of the ten cases credits the decision it asks for rather than a state the town reaches (Section 1.6.4).

Section 6.2 traces the domain. Section 6.3 sets out the town: its pipes, hydraulics, substations, buildings, plant, weather and instruments. Section 6.4 explains why the flow sets the delay and why the network is a store. Section 6.5 gives the loops, the supply curve, the planning model, the centralized and distributed MPCs and the Hankel predictor. Section 6.6 reports the ten cases as measured, the graded challenge and the code verification, Section 6.7 describes the meters, pumps and control platforms of real networks, Section 6.8 states the limitations, and the exercises in Section 6.10 run the cases.

## 6.2 Hot Water Networks and the Price of Heat

District heating distributes heat from central sources to many buildings through insulated pipe. A 2014 statement of the fourth generation of the technology traces its earlier generations from steam systems through pressurized hot water to prefabricated pre-insulated pipe, and sets lower supply and return temperatures as the direction of the field [12]. An international review in 2017 surveyed where district heating and cooling are used and how much heat they deliver [13]. A lower supply leaves less margin over what the buildings need, and a lower return is worth fuel to the plant, so the simulated network carries both a return limit and comfort bands that the supply must meet.

> ⧖ HISTORY **1877: Steam Sold by the Meter**
>
> In a town in New York State, the inventor Birdsill Holly set out to avoid the inefficiency of heating every building with its own small boiler. His system, which the landmark dates to 1877, ran a large central boiler plant that sent steam at moderate pressure to the buildings around it through a loop of supply and return mains, heavily insulated to cut the heat lost on the way. Each customer paid for the steam used, found by metering the water that condensed. The idea succeeded, and before the end of the century district heating companies had formed, mainly in large cities. The landmark marks the site of the first district steam heating system in the United States [14]. In this chapter a price also decides when each building draws.

Running a network against fuel prices and cogeneration made its operation an optimization problem. Operational optimization of a district heating system was formulated in 1995 [5], and aggregated dynamic models that reduce a network to a few equivalent pipes for simulation followed in 2002 [6]. Temperature dynamics in a working network were modeled and compared with its measurements in 2007 [7], and an equation-based thermo-hydraulic pipe model for district heating and cooling was presented in 2017 [8]. Predictive control of a complex district heating network was reported in 2005 [10], and a 2018 review of control that unlocks the flexibility of heating and cooling networks treats the pipes and the connected buildings as stores a controller can charge [9]. The buildings' own mass had been evaluated as a store under optimal control in 2004 [15], and models of building heat dynamics with one or more thermal capacities were identified from data in 2011 [16].

Coordination by prices is older than the networks' control. The decomposition principle for linear programs of 1960 lets a coordinating program generate prices for the parts of a problem, which answer with proposals [2], and the generalized Lagrange multiplier method of 1963 allocates limited resources among independent activities [1]. Architectures for distributed and hierarchical MPC were

**Table 6.1** Variable Classification

| Role | Symbol | Meaning | Range or unit |
|---|---|---|---|
| Objective | \$ | fuel cost, stores valued | dollars |
| CV | $T_s$ | supply temperature | 70 to 115 °C |
| | $\Delta p_i$ | least differential pressure | SP 100 kPa |
| | $T_{z,i}$ | zone of consumer $i$ | its band |
| Constrained | $T_{pr,i}$ | consumer's return | $\leq 55$ °C |
| | $T_{rp}$ | plant's return (challenge) | $\leq 55$ °C |
| | $T_s$ | firing high limit | $\leq 120$ °C |
| MV | $P_c$, $P_b$ | CHP and boiler heat | $\leq \bar{C}$, $\leq \bar{B}$ |
| | $P_t$ | tank power, + out | $\lvert P_t \rvert \leq \bar{S}$ |
| | $H_p$ | pump head | $\leq \bar{H}_p$ |
| | $q_i$ | consumer's space heat | valve $m_i \leq \bar{m}_i$ |
| DV | $T_{out}$ | outdoor temperature | seeded |
| | $g_i$, $Q_{d,i}$ | gains, hot-water draws | schedules |
| | events | CHP trip, burst, surgery, forecast error | |
| States | $T_s$, $P_c$, $P_b$ | node, two source lags | °C, MW |
| | $V_h$, $T_h$, $T_c$ | tank thermocline, layers | $m^3$, °C |
| | $T_{z,i}$, $T_{m,i}$ | zone and mass, six each | °C |
| | parcels | 18 pipes' contents | $m^3$, °C |
| Measured | $T_s$, $T_{rp}$, $Q$ | plant | $\sigma_T$, $\sigma_F$ |
| | $T_{ps,i}$, $T_{pr,i}$, $m_i$ | each consumer | $\sigma_T$, $\sigma_F$ |
| | $Q_{sub,i}$, $T_{z,i}$ | meter, zone | $\sigma_F$, $\sigma_T$ |
| | $T_{out}$ | outdoor | 0.2 K |
| Estimated | $\hat{T}_{z,i}$, $\hat{T}_{m,i}$ | observer per consumer | °C |

reviewed in 2009 [3], the coordination of several optimization-based controllers on one plant was set out as an opportunity and a challenge for process control in 2008 [4], and accelerated gradient methods were applied to dual decomposition in distributed MPC in 2013 [11]. On the modeling side, data-enabled predictive control replaced a parametric model by a Hankel matrix of recorded data in 2019 [17], resting on the 2005 result that a persistently exciting record of a linear system spans its trajectories [18].

## 6.3 A Plant, Three Trunks and Six Substations

The simulated town is drawn in Figure 6.1. A plant on the left holds a CHP unit, a peak boiler and a 150 $m^3$ storage tank above a hall whose mixing node sets the supply temperature $T_s$. A pump sends the water down a trunk of three segments, T1, T2 and T3 (2.0, 2.0 and 1.5 km, 5.5 km in all), which feeds six service branches of 400 to 450 m at junctions J1, J2 and J3, every pipe laid twice as supply and return. Six consumers hang at the branch ends: three apartment blocks, two office blocks and a hospital, each a substation feeding a lumped building. Every length, bore, loss coefficient, building constant, price and limit is a teaching value for a small town, chosen to be of plausible order and taken from no utility's data. The variables are classified in Table 6.1.

### 6.3.1 Plug Flow Keyed by Volume

Each pipe segment $s$, of length $L_s$, bore $D_s$, cross-section $A_s$ and volume $V_s = A_s L_s$, holds its water as a first-in, first-out queue of parcels. The kraft digester of Volume 5, Chapter 4 carries its chip column the same way, as Lagrangian slugs. A parcel that entered at time $t_0$ at temperature $T_0$ leaves at the time $t_1$ at which the volume pumped in behind it equals the pipe's volume, and its excess over the

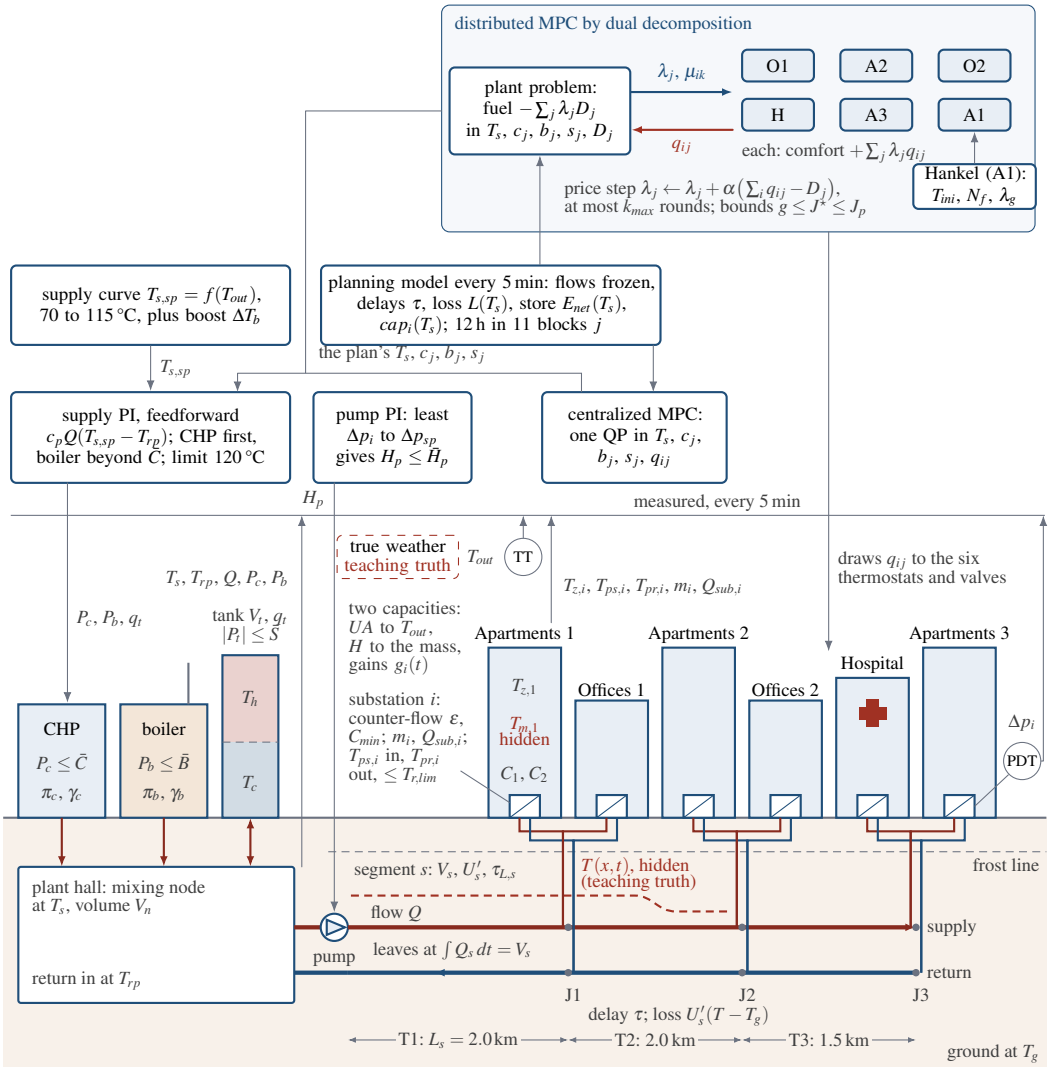



**Figure 6.1** The town's heat network as a section with its control layers above it. The plant's units feed the hall's mixing node, the pump sends the water down trunk T1 below the frost line, and branches rise from J1 to J3 to the six substations and their buildings, with the symbols written once on Apartments 1. Every layer reads the measurements on the gray line, every five minutes. In the distributed MPC, drawn as a row of its own above the planning model, the plant problem posts prices and the six consumer problems answer with draws. The supply and return pipes are labeled at the trunk's end. The water's true temperature along the pipe, the buildings' mass and the true weather are teaching truth, drawn dashed or marked hidden, and feed no controller.

ground decays at the pipe's own rate on the way, as shown in Equation (6.1).

$$\int_{t_0}^{t_1} Q_s\,dt = V_s, \qquad \tau_{L,s} = \frac{\rho c_p A_s}{U'_s}$$
$$T(t_1) = T_g + (T_0 - T_g)\,\mathrm{e}^{-(t_1 - t_0)/\tau_{L,s}} \tag{6.1}$$

where $Q_s$ is the segment's flow, $U'_s$ the loss of the insulated pipe per meter and kelvin and $T_g = 5\,^\circ\mathrm{C}$ the ground. At a constant flow the residence is $V_s/Q_s$ and the outlet follows the familiar decay of

Equation (6.2).

$$T_{out} - T_g = (T_{in} - T_g)\exp\big(-U_s' L_s/(\rho c_p Q_s)\big) \tag{6.2}$$

Each step pushes $Q_s \Delta t$ at the inlet and pops the same volume at the outlet, whole parcels and a share of the last, so nothing is interpolated and the energy that leaves is what entered less each parcel's decay. Trunk T1, 2.0 km of 0.45 m bore, holds 318 $m^3$ with $\tau_L = 224$ h. The segment flows are the sums of the consumers' flows they carry, $Q_s = \sum_{i \in s} q_i$, so the delay to a consumer is the sum of its segments' residences, each at its own flow. The true temperature profile $T(x,t)$ along every pipe is teaching truth that no instrument reads.

### 6.3.2 Pressure, Valves and the Least Consumer

The pump sets a head $H_p$. Each segment drops $k_s Q_s |Q_s|$ with $k_s = f L_s \rho/(2 D_s A_s^2)$ and a Darcy factor $f = 0.018$, and a consumer sees what is left after its supply and return paths, as shown in Equation (6.3).

$$\begin{aligned} \Delta p_i &= H_p - \textstyle\sum_{s \in \mathscr{P}_i} 2 k_s Q_s |Q_s| \\ m_i &\le \bar{m}_i \sqrt{\Delta p_i / \Delta p_{ref}} \end{aligned} \tag{6.3}$$

where $\mathscr{P}_i$ is the path from the plant to consumer $i$ and $\bar{m}_i$ is 1.6 times its design primary flow at $\Delta p_{ref} = 100$ kPa, a warm-up margin. A valve passes the flow it is asked for up to that limit, as a pressure-independent valve does. Because every drop depends on every flow, each step finds the flows by a damped fixed point (relaxation 0.4, at most 300 rounds, to $1 \times 10^{-5}$ kg/s).

### 6.3.3 Substations Rated by Effectiveness

Each substation is a counter-flow plate exchanger between the primary water and a radiator loop of fixed capacity flow $C_s$, whose radiators of conductance $UA_r$ face the room. Eliminating the two secondary temperatures gives the space heat as one conductance from the arriving supply to the zone, as shown in Equation (6.4).

$$\begin{aligned} Q_{sub,i} &= G(C_p)\,(T_{ps,i} - T_{z,i}), \quad T_{pr,i} = T_{ps,i} - \frac{Q_{sub,i}}{C_p} \\ G &= \frac{\varepsilon C_{min}}{1 + \varepsilon C_{min}(1/\varepsilon_r - 1)/C_s} \\ \varepsilon &= \frac{1 - \mathrm{e}^{-N(1-C_r)}}{1 - C_r\,\mathrm{e}^{-N(1-C_r)}} \end{aligned} \tag{6.4}$$

where $C_p = m_i c_p$ is the primary capacity flow of the space-heating share, $C_{min}$ and $C_r$ the smaller capacity flow and the ratio of the two, $N = UA_h/C_{min}$ and $\varepsilon_r = 1 - \mathrm{e}^{-UA_r/C_s}$. The conductances are sized so that the design duty passes with a 95 °C supply, a 50 °C primary return and a 70/45 °C radiator loop. $G$ rises with the flow and saturates, so a consumer short of heat at a low supply temperature opens its valve and its return rises. Above the return limit $T_{r,lim} = 55$ °C a limiter closes the valve until the return is back at the limit, giving up duty. A hot-water exchanger on the same connection meets the group's draw $Q_{d,i}$ and returns its primary water at 25 °C, with flow $Q_{d,i}/(c_p(T_{ps,i} - 25\,^\circ\mathrm{C}))$.

### 6.3.4 Two-Capacity Buildings

Each consumer is a block of buildings lumped into a zone at $T_{z,i}$ and a heavy mass at $T_{m,i}$, as shown in Equation (6.5).

$$\begin{aligned} C_1 \frac{\mathrm{d}T_{z,i}}{\mathrm{d}t} &= Q_{sub,i} + g_i(t) - UA\,(T_{z,i} - T_{out}) \\ &\quad - H\,(T_{z,i} - T_{m,i}) \\ C_2 \frac{\mathrm{d}T_{m,i}}{\mathrm{d}t} &= H\,(T_{z,i} - T_{m,i}) \end{aligned} \tag{6.5}$$

where $UA = Q_{des}/36\,\mathrm{K}$ carries the design duty at −15 °C outside and 21 °C inside, $H = r_H UA$ links the room to the structure, $C_1 = \tau_z(UA + H)$, $C_2 = \tau_{slow}/(1/UA + 1/H)$ and $g_i$ is the internal gain on its schedule plus the sun on the glass. The fast time constants are 0.48 to 0.94 h and the slow ones 26 to 53 h. The slow one makes a building a store that a price can move. Chapter 9 spends a comfort band against the same kind of slow store, an office's slab. The apartments' bands are 20 to 23 °C, the offices' 20 to 24 °C from 07:00 to 18:00 and 16 to 26 °C otherwise, and the hospital's 21 to 23 °C. A block is judged in breach when its zone lies more than 0.5 K outside its band.

### 6.3.5 The Plant, Its Tank and Its Fuel

Return water at $T_{rp}$ enters the hall's mixing node of volume $V_n$, where the CHP's and the boiler's heat are added, and a two-node stratified tank hangs between supply and return. The node, the thermocline and the sources follow Equation (6.6).

$$\begin{aligned} \rho c_p V_n \frac{\mathrm{d}T_s}{\mathrm{d}t} &= \rho c_p \left[ (Q - q_t^+)(T_{rp} - T_s) + q_t^+ (T_h - T_s) \right. \\ &\qquad \left. + q_t^- (T_c - T_s) \right] + P_c + P_b \\ \frac{\mathrm{d}V_h}{\mathrm{d}t} &= q_t^- - q_t^+, \qquad \tau_c \frac{\mathrm{d}P_c}{\mathrm{d}t} = \min(P_{c,sp}, \bar{C}) - P_c \end{aligned} \tag{6.6}$$

where $Q$ is the network flow, $q_t^+$ the tank's discharge (hot water from its top into the node, return water into its bottom) and $q_t^-$ its charge, $V_h$ the hot layer's volume at $T_h$ above the cold layer at $T_c$, and the boiler follows its setpoint likewise with $\tau_b$ and $\bar{B}$. The hot layer obeys $V_h \dot{T}_h = q_t^- (T_s - T_h)$ and the cold layer $(V_t - V_h)\dot{T}_c = q_t^+ (T_{rp} - T_c)$, each less its share of a shell loss of 1500 W/K to 5 °C. A tank power $P_t$ is turned into a flow over the layers' temperature difference, within 3 and 97 % of the volume. The fuel cost rate is Equation (6.7).

$$\dot{\$} = f(T_{rp})\,(\pi_c P_c + \pi_b P_b) + \tfrac{1}{2}\gamma_c P_c^2 + \tfrac{1}{2}\gamma_b P_b^2 \tag{6.7}$$

with $P$ in MW, where $f = 1 + 0.1\,[1 - \mathrm{clip}((57 - T_{rp})/17, 0, 1)]$ charges up to 10 % more fuel as the return rises from 40 to 57 °C, the credit of a flue-gas condenser fading above its dew point. The CHP's price is its fuel less the power it sells, and its marginal cost $\pi_c + \gamma_c P_c$ reaches about the boiler's \$52 per MWh at its full 14 MW. Volume 17, Chapter 2 treats the other side of such a unit, one fuel valve whose electricity and heat arrive on different clocks.

### 6.3.6 Weather, Forecast and Instruments

A case declares its days by a daily mean and a swing with its minimum at 06:00 and maximum at 15:00. The true outdoor temperature adds colored noise of standard deviation 0.8 K and correlation time 3 h, drawn once from the case's seed, and the controllers' forecast corrects the declared weather by the latest measurement, as shown in Equation (6.8).

$$\hat{T}_{out}(t + \ell) = T_{dec}(t + \ell) + \big(T_{out,m}(t) - T_{dec}(t)\big)\,\mathrm{e}^{-\ell/3\,\mathrm{h}} \tag{6.8}$$

Gains and hot-water draws carry 10 % colored noise. Every five minutes the controllers receive one packet: at the plant $T_s$, $T_{rp}$, the flow $Q$, the pump head, the source meters and the tank's layers, at each consumer $T_{ps,i}$, $T_{pr,i}$, $m_i$, $\Delta p_i$, $Q_{sub,i}$, the hot-water meter and the zone $T_{z,i}$, and $T_{out}$, with white noise of $\sigma_T = 0.1$ K on temperatures (0.2 K outdoors), $\sigma_F = 0.5$ % on flows and meters and 2 kPa on pressures. The pipe profiles, the buildings' mass, the true weather and the load noise never enter the packet, and the app's tests search every controller file for a read of any of them.

### 6.3.7 Constants, Line-Out and Numerics

Table 6.2 lists the parameters. Each case starts lined out at the hour its weather declares: every zone at its preferred temperature with the heat that holds it there, the flows and arrival temperatures iterated

**Table 6.2** Parameters, Defaults and Their Standing

| Symbol | Meaning | Value | Unit, standing |
|---|---|---|---|
| $\rho$, $c_p$ | water near 80 °C | 970, 4190 | SI |
| $L_s$, $D_s$ | trunks T1, T2, T3 | 2.0, 2.0, 1.5 | km |
| | bores | 0.45, 0.35, 0.25 | m, T |
| $U'_s$ | trunk loss per pipe | 0.80, 0.70, 0.60 | W/(m K), T |
| | branch loss per pipe | 0.40 to 0.45 | W/(m K), T |
| $f$, $T_g$ | Darcy factor, ground | 0.018, 5 | °C, T, A |
| $Q_{des}$ | design duty A1, O1, A2 | 4.0, 3.0, 4.0 | MW, T |
| | O2, H, A3 | 3.0, 3.5, 3.5 | MW, T |
| $\tau_z$, $\tau_{slow}$ | apartments | 0.75, 40 | h, T |
| | offices | 0.5, 25 | h, T |
| | hospital | 1.0, 50 | h, T |
| $r_H$ | $H/UA$ | 3, 2.5, 3 | T |
| $\bar{m}_i$ | valve rating | 1.6 | × design, T |
| $\bar{C}$, $\bar{B}$ | CHP, boiler capacity | 14, 16 | MW, T, A |
| $V_t$, $\bar{S}$ | tank, its power | 150, 8 | $m^3$, MW, T, A |
| $V_n$ | mixing node | 40 | $m^3$, H |
| $\tau_c$, $\tau_b$ | CHP and boiler lags | 0.15, 0.10 | h, T, A |
| $\pi_c$, $\gamma_c$ | CHP heat price | 22, 2.1 | $/MWh, /MW, A |
| $\pi_b$, $\gamma_b$ | boiler heat price | 52, 1.5 | $/MWh, /MW, A |
| $\bar{H}_p$, $\Delta p_{sp}$ | pump limit, setpoint | 600, 100 | kPa, T, A, C |
| $T_{r,lim}$ | return limit | 55 | °C, T, A |
| $T_{s,des}$ | curve at −15 °C | 95 | °C, C |
| $\sigma_T$, $\sigma_F$ | instrument noise | 0.1, 0.005 | K, relative, A |
| $K_T$, $\tau_{I,T}$ | supply loop | 0.4, 1.0 | MW/K, h, C |
| $K_p$, $\tau_{I,p}$ | pump loop | 0.5, 0.25 | kPa/kPa, h, C |
| $K_z$, $\tau_{I,z}$ | thermostats | 1.0, 2 | MW/K, h, C |
| $\alpha$, $k_{max}$ | price step, round cap | 0.5, 30 | $/MWh/MW, C |
| $T_{ini}$, $N_f$, $\lambda_g$ | Hankel predictor | 4, 12, 0.001 | slots, C |

T teaching value, A adjustable in the Model dialog, H hidden, C a control on the faceplates. The Model dialog also scales the pipe loss, the exchangers, the buildings' mass and Apartments 2's radiators. The challenge and case 10 ship a 15 MW boiler.

twice, every pipe filled at its inlet temperature, and the tank half charged. The engine's base is the hour (Section 1.6.1). A step of 10 s pops and pushes the pipes, then integrates the node, the tank, the sources, the buildings and eighteen ledgers by the fourth-order Runge–Kutta method with the step's boundary values held, and then enters the plant's outflow and each consumer's return into their pipes at the temperatures the ledgers integrated. The controllers run every 5 min, and the historian logs every minute.

### 6.3.8 Assumptions

The fronts stay sharp because the pipe has no mixing or diffusion, while a real front smears over a few hundred meters. The ground holds one temperature, and the loss coefficients are teaching values rather than a manufacturer's series. Each consumer is one zone and one mass with one radiator loop at a fixed flow and one hot-water exchanger, so a block's average temperature stands for its occupants'. The hydraulics are algebraic at every step, with no pressure waves, and pumping costs nothing. The CHP has no minimum load and its electrical side enters only through its net heat price. The weather is one declared record with seeded noise, and the forecast never sees the noise's future.

## 6.4 Why the Flow Sets the Delay

The delay a consumer sees is a sum of residences at flows the consumers choose. In case 1 the 13:05 step left at 50 kg/s while T2, T3 and the hospital's branch carried 34.9, 19.2 and 9.9 kg/s. Integrating

each segment's own flow until it has passed the segment's volume puts the front at the hospital 4.43 h after the setpoint jumped (1.69, 1.48, 0.92 and 0.35 h for T1, T2, T3 and the branch), or 4.58 h after the plant's own supply had risen by half the step 0.13 h later, and these bracket the measured 4.48 h. Frozen at the departure flows the sum is 4.62 h, and the path's 598 $m^3$ over the plant's whole flow would be 3.23 h. At 02:00, with 78.2 kg/s leaving, the integral gives 2.81 h against 2.83 h measured. The two delays stand in the ratio 1.58 while the plant's flows stand in the ratio 1.57, because the flow divides almost the same way at both hours.

**Pause.** Try Exercise 6.1 before reading on.

A slow delay makes the network a store. Raising $T_s$ fills the pipes with hotter water that the consumers draw hours later, while the hotter pipe loses more on the way by Equation (6.2). The buildings hold more, since a block warmed a fraction of a kelvin in the evening still carries the heat at the morning peak when its slow time constant is 26 to 53 h. Timing either store needs the delay, which the flow sets, and the price of the charging heat, which the dispatch sets. A consumer whose radiators are too small opens its valve, raising its flow and its return together, and the plant pays through the condenser's credit of Equation (6.7).

**DID YOU KNOW? Tanks That Bridge a Winter Peak**

In one geothermal network, winter peaks call for about 18 400 $m^3$/h of hot water for space heating, while the geothermal sources typically give 7200 $m^3$/h. Beside an oil-fired peak plant, 101 000 $m^3$ of water tanks were built to bridge those peaks [9]. The tank of this chapter's town holds 150 $m^3$ (Table 6.2).

## 6.5 Loops, Planners and a Price

Every loop is the velocity-form proportional-integral (PI) law of Section 1.6.2, with back-calculation on the integral and the clamp on the emitted output, run every $\Delta t_c = 5$ min on the packet of measurements. In Manual the plant's heat and the pump's head come from hand stations, and the PI laws track them so that a return to Auto is bumpless.

### 6.5.1 The Regulatory Layer

The supply loop computes the heat the plant must add, a PI on the measured $T_s$ whose bias is the feedforward $c_p Q_m (T_{s,sp} - T_{rp,m})$ from the measured flow and return. Its ceiling is the plant's capacity and a firing high limit that admits no more heat than brings the supply to $T_{hi} = 120\,°C$ within two control periods, $h_{hi} = c_p Q_m (T_{hi} - T_{rp,m}) + \rho c_p V_n (T_{hi} - T_{s,m})/(2\Delta t_c)$. The limit is a teaching value above the operating range, the continuous rating for which bonded pre-insulated pipe to the standard EN 253 is made, and with the sources' lags 40 MW by hand peaks at 126.7 °C. The dispatch then takes the tank's power $P_t$ off the request, gives the CHP up to its available capacity and the boiler only what lies beyond it, so the boiler does not fire to cover the CHP's 9 min lag. Under the curve a tank rule charges on the CHP's spare heat and discharges to spare the boiler. The pump loop sets $H_p$ from the measured path drops plus the setpoint $\Delta p_{sp}$, trimmed by a PI on the least measured $\Delta p_i$.

Each consumer runs its own thermostat, a PI on $T_{z,i}$ with its building's heat-loss feedforward, which sets its space heat $q_i$ under the curve. The valve flow that delivers $q_i$ is found by bisection on the monotone $G$ of Equation (6.4), with the declared exchanger and radiators, at the measured supply and zone, and the return limiter cuts it when $T_{pr,i}$ would exceed $T_{r,lim}$. Under either MPC the plan's draw is applied and the thermostat becomes a guard that adds $K_z$ times the shortfall when a zone comes within 0.3 K of its band's lower edge, a band read one hour ahead. The plan does not see the guard.

### 6.5.2 The Supply Curve and a Boost

The heating curve gives the setpoint of Equation (6.9), held between 70 and 115 °C.

$$T_{s,sp} = 70 + \frac{10 - T_{out,m}}{25}\,(T_{s,des} - 70) \tag{6.9}$$

with temperatures in °C, so the supply is 70 °C at 10 °C outside and $T_{s,des}$ at −15 °C. The manual boost adds $\Delta T_b$ for a set length from a clock hour every day. Neither looks ahead.

### 6.5.3 A Planning Model With Frozen Delays

Both MPCs solve one linear planning model, built every five minutes from the packet and the declared constants, a prediction model in the sense of Section 1.4.2. Twelve hours are split into 24 slots of $\Delta = 0.5$ h and eleven move blocks $j$ (four of 0.5 h, four of 1 h, three of 2 h), with decisions $T_{s,j}$, $c_j$, $b_j$, $s_j$ (CHP, boiler and tank, MW) and the six draws $q_{ij}$. The flows are frozen at their measured values, so each segment's residence $\tau_s = V_s/Q_s$, each consumer's delay $d_i$ and its decay $\phi_i = \prod_s \mathrm{e}^{-\tau_s/\tau_{L,s}}$ are fixed for the solve and recomputed at the next. The supply pipes' stored heat $E_{net}$ and their loss $L$ are then affine in the plant's supply history, the controller's own record before now and the plan's blocks after it, and each block's heat balance and each slot's capacity are Equation (6.10).

$$\begin{aligned} n_j\Delta\Big(c_j + b_j + s_j - \textstyle\sum_i q_{ij}\Big) &= \textstyle\sum_{k\in j}\Delta\,(d_k + L_k(T_s)) \\ &\quad + E_{net}(t_{j+1}) - E_{net}(t_j) \\ q_{ik} &\le cap_i\big(T_{arr,ik}(T_s)\big) \end{aligned} \tag{6.10}$$

where $n_j$ is the block's slot count, $d_k$ the declared hot-water draws plus the return pipes' loss held at the measured returns, and $T_{arr,ik} = \phi_i \bar{T}_s(t_k - d_i) + (1-\phi_i)T_g$. The capacity is the tangent of the exchanger's wide-open duty (with the limiter's flow cut when it is on) at the measured arrival temperature. Each consumer's zone and mass follow Equation (6.5) discretized exactly at $\Delta$ from the declared constants, started at the observer's estimate (a steady-state Kalman filter per consumer on the same model, driven by its heat meter, $T_{out,m}$ and the declared gains, with process noise of 0.05 and 0.01 K per five minutes on zone and mass and the declared zone measurement noise, $\sigma_T = 0.1$ K). The objective is Equation (6.11).

$$\begin{aligned} J = {} & \textstyle\sum_k \Delta\big[\pi_c' c + \tfrac{1}{2}\gamma_c c^2 + \pi_b' b + \tfrac{1}{2}\gamma_b b^2 + \tfrac{1}{2}\gamma_s s^2\big]_k \\ & + \textstyle\sum_{i,k} w_i (T_{z,ik} - r_{ik})^2 + \sum_{i,j}(\rho_s \sigma_{ij}^2 + \rho_l \sigma_{ij}) \\ & + \textstyle\sum_j s_T (\Delta T_{s,j})^2 + \sum r_Q q^2 - \pi_T (E_{t,N} + E_{net,N}) \end{aligned} \tag{6.11}$$

where $\pi' = f(T_{rp,m})\pi$ freezes the condenser factor at the measured return, $r_{ik}$ is the preferred temperature, $w_i$ is 20, 10 and 100 \$ per K$^2$ and slot for apartments, offices and the hospital, $\sigma_{ij}$ is the worst excursion of block $j$ outside the band backed off by 0.5 K and met an hour early, priced at $\rho_s = 2\times10^4$ \$/K$^2$ and $\rho_l = 2\times10^3$ \$/K as the soft constraint of Equation (1.7), $\gamma_s = 0.5$, $s_T = 1$ \$/K$^2$ and $r_Q = 0.05$ \$/MW$^2$. The terminal price $\pi_T$ values heat left in the tank and the pipes at the horizon's end, the terminal value of Section 1.4.8, at the boiler's base price if the forecast load then exceeds the CHP's capacity, and at the CHP's marginal cost otherwise. The CHP's and boiler's lags are not in the plan, and the supply loop dispatches.

### 6.5.4 Centralized MPC

The centralized MPC solves Equation (6.11) subject to Equation (6.10), the tank's charge inside its limits at every slot and the boxes on every input, as one quadratic program on the framework's linear MPC and its dense dual active-set solver. It has 176 variables (110 decisions and 66 comfort slacks) and 788 rows, and its Hessian is positive definite, with a smallest eigenvalue of $3.73\times10^{-2}$. The first block's supply temperature, tank power and draws are applied, and a failed solve keeps the previous plan.

### 6.5.5 Distributed MPC by Dual Decomposition

The distributed MPC solves the same problem by prices, the dual decomposition that Section 1.4.11 sets out on two tanks sharing a feeder. Chapter 5 divides a plan among gates that exchange plans

instead of prices. The plant keeps its own copy $D_j$ of the heat it delivers in block $j$, and two couplings are relaxed: the balance $\sum_i q_{ij} = D_j$ with a price $\lambda_j$ per block, and each capacity with a surcharge $\mu_{ik} \geq 0$ per consumer and slot. The Lagrangian then separates. Consumer $i$ minimizes its comfort terms of Equation (6.11) plus $\sum_k \Delta(\lambda_{j(k)} + \mu_{ik}) q_{ik}$ on its own building model, and the plant minimizes its fuel, terminal and move terms less $\sum_j n_j \Delta \lambda_j D_j$ and the surcharges it earns by raising the capacity, subject to the balance in $D_j$, the tank and its boxes. The sum of the seven optimal values, less the surcharges' constant part, is the dual function $g(\lambda, \mu)$, a lower bound on the centralized optimum $J^\star$. Each round moves the prices on the imbalances, as shown in Equation (6.12).

$$\begin{aligned} \lambda_j &\leftarrow \lambda_j + \alpha\Big(\textstyle\sum_i q_{ij} - D_j\Big) \\ \mu_{ik} &\leftarrow \max\big(0,\ \mu_{ik} + 60\alpha\,(q_{ik} - cap_{ik}(T_s))\big) \end{aligned} \tag{6.12}$$

with $\alpha$ in \$/MWh per MW, warm-started from the previous control step. The balance update is Equation (1.12) with one price per block. A step settles when every block's imbalance and every capacity excess is under 0.1 MW and the gap below is under 1 % of the plan, or stops at $k_{max} = 30$ rounds. The primal plan is recovered by letting the plant solve once more with $D_j$ pinned to the draws and rows that make its supply plan carry every draw. That plan meets every constraint of the centralized problem, so its cost $J_p$ on the centralized objective is an upper bound and $g \leq J^\star \leq J_p$. The gap $J_p - g$ the app reports is therefore a real one. When no supply plan can carry the draws, each draw is first cut to the capacity of the plant's last plan, and when even that fails the cut draws and the plant's last plan are applied and the step is flagged NO PLAN FITS.

### 6.5.6 A Building Learned From Its Own Record

Apartments 1 can plan on a Hankel-matrix predictor built only from the run's record. Every 0.5 h slot the controller stores the block's metered heat, the measured $T_{out}$ and the declared gain as $u$ and the measured zone as $y$, and the last 72 slots (36 h) form block Hankel matrices of depth $T_{ini} + N_f = 16$, 57 columns wide. The predictor is Equation (6.13).

$$\begin{bmatrix} U_p \\ Y_p \\ U_f \end{bmatrix} \theta = \begin{bmatrix} u_{ini} \\ y_{ini} \\ u_f \end{bmatrix}, \quad y_f = Y_f \theta, \quad \theta = (\mathscr{H}^\top \mathscr{H} + \tilde{\lambda}_g I)^{-1} \mathscr{H}^\top b \tag{6.13}$$

where $\mathscr{H}$ stacks $U_p$, $Y_p$ and $U_f$, $b$ stacks the data on the right, and $\tilde{\lambda}_g = \lambda_g \operatorname{tr}(\mathscr{H}^\top \mathscr{H})/57$. The least-squares weights $\theta$ make the predicted zone affine in the future heat, which is the regularized form of data-enabled predictive control with no slack on the past outputs [17], a data-driven predictor of the kind Section 1.4.2 describes. It is rebuilt every hour, and the consumer's problem on it has eight free blocks and one band slack per slot.

### 6.5.7 Three Models and the Grade

The plant is the whole of Section 6.3. The planning model freezes the delays at the measured flows and the return pipes at the measured returns, and it carries the buildings' declared constants and a 5 °C ground, which are the plant's own until a case or the learner changes the plant, as case 6 halves a radiator: the valves, the observers and both MPCs keep the declared design, and only the return limiter sees the plant's. The Hankel predictor knows only the record. The fuel each run buys is graded with the heat it leaves in its stores valued, as shown in Equation (6.14).

$$\$_{graded} = \$ + \textstyle\sum_k p_k \big(E_k(0) - E_k(t_f)\big) \tag{6.14}$$

where $k$ runs over the tank, the network (pipes and mixing node) and the buildings, and $E_k$ is the heat each holds. In the challenge and case 10, which end with the CHP out, every MWh short or over is valued both ways at the boiler's marginal price at its cap, $p = \pi_b + \gamma_b \bar{B} = \$74.5$ per MWh, which equals comparing two runs with their end states matched. The other cases value deficits at $\pi_b$ and surpluses at $\pi_c$, a conservative rule under which no run gains by emptying or filling its stores.

**Table 6.3** The Ten Cases

| # | Start, weather | Shipped settings, events | Run |
|---|---|---|---|
| 1 | 12:00, W1 | boost +15 K at 13:00 for 2 h | 18 h |
| 2 | 18:00, W1 | supply curve | 18 h |
| 3 | as case 2 | boost +15 K at 04:00 for 3 h | 18 h |
| 4 | as case 2 | centralized MPC | 18 h |
| 5 | as case 2 | distributed MPC, $\alpha$ 0.5 | 18 h |
| 6 | as case 2 | curve, A2 radiators $\times$0.5, limiter off | 18 h |
| 7 | as case 2 | distributed MPC, CHP trip at 07:00 | 18 h |
| 8 | 00:00, W2 | curve with A1 excited, then distributed MPC | 48 h |
| 9 | 12:00, W3 | challenge: curve, boiler 15 MW, CHP trip at 07:30 on the second morning | 48 h |
| 10 | 12:00, W3 | distributed MPC, boiler 15 MW, CHP trip at 00:00 (infeasible) | 48 h |

Declared daily means, °C: W1 −2, −4, W2 −3, −5, −4, W3, the snap, −4, −7, −10, each with a 4 K swing, lowest at 06:00. Seeds: 11 (case 1), 12 (cases 2 to 7), 13 (case 8), 14 (cases 9 and 10).

**Table 6.4** Measured Results of Cases 2 to 8

| Case | Setting | Boiler (MWh) | Valued ($) | Worst (K) | Measured |
|---|---|---|---|---|---|
| 2 | curve | 28.2 | 10 282 | 0 | peak 10.9 MW, 06:25 |
| 2 | curve, tank idle | 36.2 | — | — | |
| 3 | boost 04:00 | 35.6 | 10 484 | 0 | hospital at 06:53 |
| 4 | centralized MPC | 19.4 | 9 953 | 0 | guard 1.5 h |
| 5 | distributed, $\alpha$ 0.5 | 17.8 | 9 901 | 0.46 | 177 of 216 settle |
| 5 | distributed, $\alpha$ 0.8 | — | 9 909 | — | none settles |
| 5 | distributed, $\alpha$ 1.5 | 25.3 | 9 936 | — | none settles |
| 6 | limiter off | 29.9 | 10 487 | 0 | A2 return 68.7 °C |
| 6 | limiter on | — | — | 1.15 | A2 7.3 h out |
| 7 | distributed, trip | 78.8 | 11 530 | 0.46 | curve $12 088 |
| 8 | distributed | 59.2 | 26 168 | 0.17 | 238 of 288 settle |
| 8 | A1 on Hankel | — | +136 | 0.20 | 170 of 288 settle |

Valued: the fuel cost with deficits in the stores at $\pi_b$ and surpluses at $\pi_e$. Worst: the largest distance of any zone outside its declared band (case 6 on: Apartments 2, case 8 Hankel: Apartments 1). Case 8 Hankel: valued cost above the two-capacity run. A dash: not measured.

## 6.6 What the Cases Measure

The ten cases are listed in Table 6.3 and their outcomes in Table 6.4. Each number in this section was taken from the application's engine stepped without a browser, each case on its own seed and over its own run, with a named setting entered before Run. The same scripts find all 168 numbers that the app's own texts quote, with none missing.

### 6.6.1 Arrival at Two Flows

Case 1 starts at noon and sends one step. Under the shipped boost the step left at 13:05 and reached Apartments 1 in 2.06 h, Apartments 2 in 3.49 h and the hospital in 4.48 h, as shown in Figure 6.2a. With the boost moved to 02:00 and the case reset, the step left with 78 kg/s flowing and reached the same three in 1.34, 2.21 and 2.83 h. A step is counted as arrived when the supply reaching a consumer

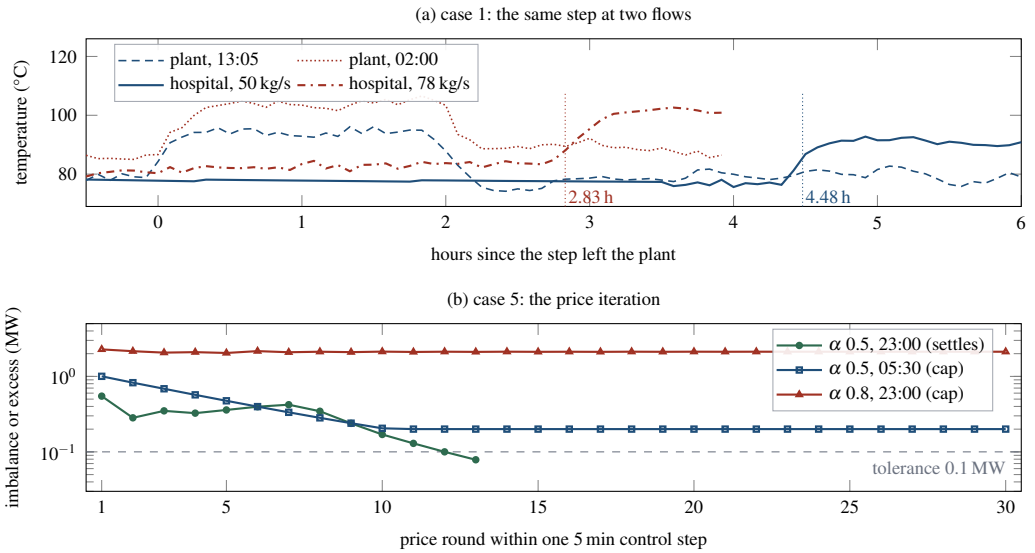



**Figure 6.2** The two halves of the lesson, from the engine. (a) Case 1, the true supply temperature leaving the plant (dashed for the 13:05 step, dotted for the 02:00 step) and arriving at the hospital (solid and dash-dotted) after the 15 K step, at 13:05 with 50 kg/s and at 02:00 with 78 kg/s (the second run ends at 06:00). (b) Case 5, the larger of the imbalance and the capacity excess in each round of one control step, for a step that settles at 23:00 (circles), one that stops at the round cap at 05:30 (squares), and the same 23:00 step with $\alpha = 0.8$ (triangles).

has risen by half the step. The flow is lower in the afternoon, when the declared swing is near its daily maximum and the buildings need less heat, and higher at 02:00, when the night is colder.

> **TRY IT IN THE APP** **1 · The Delay at Two Flows**
>
> Load the scenario and press Run. Watch the 15 K step that leaves the plant at 13:05 with 50 kg/s flowing. It reaches Apartments 1 after 2.06 h, Apartments 2 after 3.49 h and the hospital only after 4.48 h, as the front crosses segments that carry less and less water.

### 6.6.2 Sending Heat Ahead by Hand and by Plan

The cold morning of case 2 drives the supply curve past the CHP's 14 MW from 01:17, and the boiler fires to 10.9 MW at 06:25, making 28.2 MWh over 18 hours for a valued cost of $10,282. The tank, charged on the CHP's spare heat in the evening, spares the boiler 8.0 MWh, for with the tank idle the boiler makes 36.2 MWh. Case 3's boost leaves at 04:00 with 79 kg/s flowing and reaches the hospital 2.88 h later, at 06:53, as the peak arrives. It still raises the boiler's energy to 35.6 MWh and the valued cost by $202, because the CHP was already at its capacity all night and the boost's heat came from the boiler, while the hotter pipe lost more to the ground. A charge pays only with cheaper heat to spare when it is made. The centralized MPC of case 4 finds that heat in the evening. It raises the supply while the CHP has room and lets it fall at the peak, so the consumers draw down the pipes, the tank and their own mass. The boiler makes 19.4 MWh, and the run costs $9,953, $329 or 3.2 % less than the curve, with the offices on their thermostats' guard for 1.5 h of the 18.

**Pause.** Try Exercise 6.2 before reading on.

### 6.6.3 A Price That Settles and One That Swings

At the shipped step $\alpha = 0.5$ the price of case 5 settles in 177 of 216 control steps, in 9.8 rounds on average, and the run costs \$9,901, within \$52 of the centralized MPC. The 39 steps that stop at the cap fall between 03:35 and 07:55, when the offices' warm-up meets a consumer's capacity. At 23:00 the imbalance falls below 0.1 MW in 13 rounds with a gap of 0.10 % of the plan, as shown in Figure 6.2b. At 05:30 the imbalance is under 0.1 MW from round 5 and under 0.01 MW from round 11, but one consumer's capacity excess stays at 0.20 MW from round 11 on, while its surcharge climbs by only $60\alpha \times 0.20\,\text{MW} = 6\,\$/\text{MWh}$ a round, and the recovered plan shows a gap of 73 %. At $\alpha = 0.8$ the same 23:00 step keeps 2.1 MW of imbalance through all 30 rounds while the price alternates, and no step of the run settles. At 1.5 none settles either, and the boiler makes 25.3 MWh instead of 17.8 for \$9,936. The recovery keeps every applied plan feasible, but it cannot make an unsettled price a good one.

> **TRY IT IN THE APP** **5 · The Price Converging, and a Step That Does Not**
>
> Load the scenario and run its 18 hours with the shipped Price step of 0.5. The price settles in 177 of the 216 control steps, and the run costs \$9 901. Reset, set Price step to 0.8 before Run and repeat: no step settles, and the run costs \$9 909.

Where the price settles it is the right price. At every one of case 5's 177 settled steps the first block's price equals the multiplier of the same block's balance in the centralized solution of the same planning problem to within 0.71 \$/MWh (median 0.11), and any block's to within 1.00 \$/MWh. The bracket $g \le J^\star \le J_p$ holds at all 216 recovered steps, and the recovered plan exceeds the centralized optimum by at most 0.91 % where the step settled. At the round cap the reported gap is an upper bound and can be a loose one. The recovered plan exceeded the optimum at that step by a median 13 % and at most 162 % in case 5, and by at most 604 % in the challenge.

### 6.6.4 A Hot Return, a Trip and a Learned Building

Case 6 gives Apartments 2 radiators half the size its load needs, with the limiter off. Its valve opens wide and its return reaches 68.7 °C, lifting the plant's mean return from 35.4 to 42.8 °C and the valued cost by \$202. With the limiter on, the plant is spared but the block spends 7.3 h more than 0.5 K below its band, at worst 1.15 K. In case 7 the CHP trips at 07:00 as the offices open, and the price rises from about 57 to 73 \$/MWh as the boiler becomes the marginal source. At that price the offices cut their draws by about a quarter and settle 0.1 to 0.9 K above their 20 °C edge, while the hospital, whose comfort weighs ten times as much, keeps 21.9 °C. The run costs \$11,530 against the curve's \$12,088.

Case 8 records the first day under the curve while Apartments 1's thermostat is moved ±0.5 K at random every two hours, then runs the distributed MPC. On the second day the Hankel predictor's 6 h prediction of the block's zone has an error of 0.43 K rms in open loop and 0.51 K in the loop, against 0.13 K for the two-capacity model, which has the plant's own structure and constants and is a bound rather than a fair rival. Planning on the predictor, which predicts the block colder than it is, the block asks for more heat than its exchanger passes. The price settles at 170 of 288 steps (238 with the two-capacity model), the block's guard adds heat for 4.5 h of the day, and the block stays within 0.20 K of its band for \$136 more over the two days. No step uses data recorded after it.

**Pause.** Try Exercise 6.3 before reading on.

### 6.6.5 The Cold Snap, Graded

The challenge runs two days of a cold snap from noon, to −14 °C at 06:00 on the second morning, with a 15 MW boiler, and trips the CHP at 07:30 that morning. A run passes with no block more than 0.5 K outside its declared band at any instant, the return at or below its limit for 95 % of the time, and a graded cost of at most \$31,553, 1.25 % below the supply curve's \$31,952, as listed in Table 6.5.

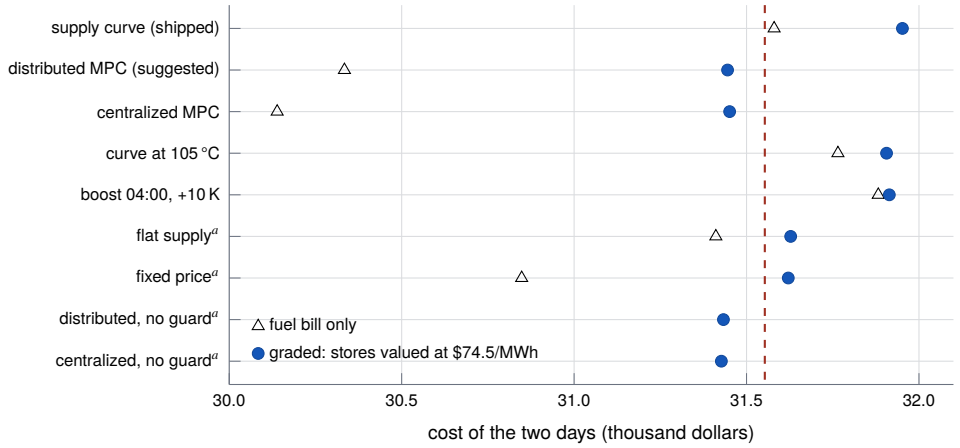



**Figure 6.3** The challenge on seed 14 redrawn from Table 6.5: each policy's fuel bill (open triangles) and its graded cost with the stores valued (filled circles), with the $31 553 target as a dashed line. [a]Teaching ablations, not panel settings, as in the table. Notice that the two MPCs buy the least fuel but leave the most heat unpaid in their stores, so grading closes most of their lead, and that a run passes only if its circle lies left of the line and its comfort and return limits hold.

**Table 6.5** The Challenge on Its Weather (Seed 14) and Seven Seeds

| Policy | Bill ($) | Stores (MWh) | Graded ($) | Worst (K) | Verdict | Seeds |
|---|---|---|---|---|---|---|
| supply curve (shipped) | 31 580 | −5.0 | 31 952 | 0.20 | fail | 0 of 7 |
| **distributed MPC** | **30 334** | **−14.9** | **31 445** | **0.37** | **pass** | **5 of 7** |
| centralized MPC | 30 139 | −17.6 | 31 451 | 0.16 | pass | 6 of 7 |
| curve at 105 °C | 31 765 | −1.9 | 31 906 | 0 | fail | 0 of 7 |
| boost 04:00, +10 K | 31 882 | −0.4 | 31 914 | 0 | fail | 0 of 7 |
| flat supply[a] | 31 411 | −2.9 | 31 628 | 0 | fail | 0 of 7 |
| fixed price[a] | 30 848 | −10.4 | 31 621 | 0 | fail | 0 of 7 |
| distributed, no guard[a] | — | — | 31 433 | 0.59 | fail | — |
| centralized, no guard[a] | — | — | 31 427 | 0.33 | pass | — |

Stores: the change of tank, network and buildings over the run, valued at $74.5/MWh. Seeds: passes against each weather draw's own supply curve. The distributed MPC (bold) is the suggested setting. [a]Teaching ablations in the app's tests, not settings on the panel.

Changing a model parameter or firing a disturbance by hand during the challenge leaves the attempt ungraded, and the app says so at once. The curve's bill is $31,580 with its stores 5.0 MWh short at the end. The suggested distributed MPC charges the network and the buildings in the evenings and draws them down at the peaks. Its bill is $30,334, 3.9 % less, but its stores end 14.9 MWh short, and graded at $74.5 per MWh it passes at $31,445, $108 inside the target, with the worst block 0.37 K out. Both halves of the strategy are needed. In two teaching ablations that no panel setting reaches, the distributed MPC held to the curve's supply temperature saves 1.01 %, and with a fixed price instead of a coordinating one 1.04 %, and both fail. The state a quarter hour after the trip is shown in Figure 6.4. Figure 6.3 draws each policy's bill beside its graded cost.

The verdict belongs to one weather. On seven draws, each against its own curve, the suggested setting saves 1.17 to 1.64 % and passes on five (one draw saves 1.17 %, another leaves a block 0.62 K out), the centralized MPC saves 1.15 to 1.62 % and passes on six, the two ablations save 0.48 to

1.15 % and pass on none, and the curve variants save 0.06 to 0.22 % and pass on none. The ranking holds on every draw. The valuation matters. With deficits charged at the boiler's $52 no-load price and surpluses credited at the CHP's $22, the asymmetric rule the other cases use, the verdict would turn on which store a run left empty, and valued both ways at one price the suggested setting reaches a 2 % margin only below 61.4 $/MWh. The suggested setting's comfort also depends on the consumers' guard, which holds Offices 1 and Offices 2 at their edges for 4.2 and 3.4 h over the two days. Without the guard, Offices 2 falls 0.59 K below its band and the attempt fails, while the centralized MPC passes without it at 0.33 K.

> **+ DID YOU KNOW? A Smaller Bill, a Larger Graded Cost**
>
> From the simulator: in the challenge the centralized MPC buys fuel for $30 139, less than the distributed MPC's $30 334, but leaves its stores 17.6 MWh short against 14.9 MWh. Graded with the stores valued, it costs $31 451 and the distributed one $31 445 (Table 6.5), so the bill alone ranks them wrongly.

After the trip the coordination does not converge. The boiler's 15 MW cannot carry every block's ask, and none of the 55 control steps after the trip settles. On 54 of them no plan fits, since the draws, cut to their capacities, still exceed what the plant can deliver in some block by a median of 2.2 MW and at most 3.3 MW. The price climbs to a median of $398 and at most $984 per MWh, while the multiplier of the first block's balance on the same planning problem is a median of $76,993 and at most $138,567, set by the comfort band's soft penalty. A fixed step of 0.5 cannot climb that far in 30 rounds. The plant gives what it can, the guard holds the offices, and the hospital keeps its heat. With no plan that fits, the tank charges only from spare heat while the supply is within 0.5 K of its setpoint, which holds or trims nine planned charges.

### 6.6.6 A Snap No Price Holds

Case 10 trips the CHP at midnight as the snap begins and keeps it out, so the 15 MW boiler and the tank carry two cold nights. Under the shipped distributed MPC the offices fall 4.67 K below their band, the apartments 2.76 K and the hospital 2.16 K. A logged sweep of fourteen settings a learner can reach (each strategy, the curve at 95 and 115 °C, an evening boost, an idle tank, two price steps, four band settings, Manual at the plant's limit, and the network charged in Manual before the trip) holds the hospital and every apartment block within 0.5 K in none. The best, Manual at the limit with raised edges and the quickest thermostats, leaves the apartments 1.52 K and the hospital 0.88 K out, with the offices 3.82 K below theirs. The claim is scoped to the shipped boiler and tank. At 17 MW the shipped distributed MPC holds both, at 0.45 and 0 K. Here the price after the trip sits at a median of $16,809 per MWh against a multiplier of $905,251.

### 6.6.7 Code Verification

Code verification tests the program against closed forms and conservation. A single pipe at constant flow matches Equation (6.2) at 0.02, 0.05 and 0.15 $m^3/s$ on T1 and a branch with errors of $3.5 \times 10^{-10}$ to $3.0 \times 10^{-8}$ K. The parcel queue agrees with $V/Q$ at constant flow and with $\int Q\,dt = V$ when the flow changes in transit to within one 10 s step, and the delays of case 1 agree with the segment-by-segment integral as reported in Section 6.4. Over the 18 hours of case 2 the plant produces 274.84 MWh, of which 257.64 MWh is delivered, 9.755 MWh is lost from the pipes (an integral of $U'(T - T_g)$ over every parcel, independent of the ledger), 1.174 MWh is lost from the tank and 6.268 MWh is stored, a residual of $-1.65 \times 10^{-4}$ MWh, or $-6.0 \times 10^{-7}$ of production. The plant's $\rho c_p Q(T_s - T_{rp})$ equals the sources' mean heat to 0.2 % in each of seven runs checked, the plant flow spans 36 to 191 kg/s, the fastest water moves at 1.53 m/s, and the pump reaches its 600 kPa limit only in case 10. The duty never exceeds $C_{min}$ times the available temperature difference, at 38 880 substation-steps each of cases 2 and 6, where the largest excess is −4.67 W. Halving the step to 5 s moves the suggested challenge's bill from $30,334 to $30,331 and its saving from 1.59 to 1.58 %, with both verdicts unchanged. With the plant and its state wrapped during every controller call of cases 7, 8 and the

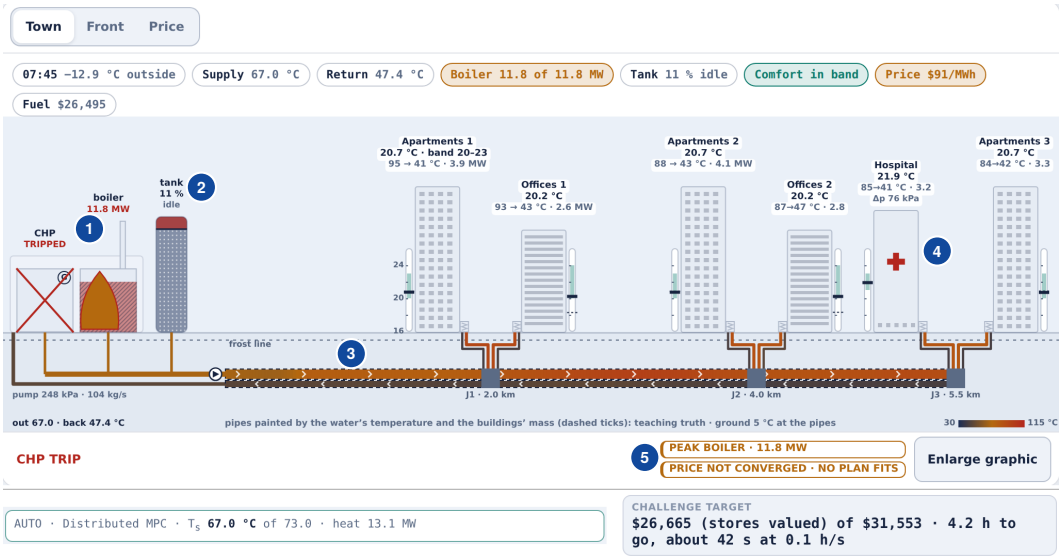



**Figure 6.4** The challenge with the suggested distributed MPC ($\alpha = 0.5$) at $t = 43.75$ h, shown by the clock as 07:45 on the second morning, a quarter hour after the CHP trip, on the Town view. The figure joins two parts of the screen: the Town view above, and below it the supply faceplate's readout line and the challenge tile, with the header, the control panel, the trends and the rest of the Learn panel cut. (1) The CHP is crossed out, and the boiler, ramping toward its 15 MW cap (Table 6.3), gives 11.8 MW of the 13.1 MW asked. The supply leaves at 67.0 °C against a setpoint of 73.0 °C, so (2) the tank, 11 % full, stands idle although the plan, answering prices that did not settle, would by the record charge it at about 1.4 MW. (3) The trunk's water and the buildings' mass are drawn as teaching truth. The return is 47.4 °C at −12.9 °C outside, Offices 1 stands at 20.2 °C on its guard and (4) the hospital at 21.9 °C, and the first block's price is about 91 $/MWh. (5) The strip reads CHP TRIP, PEAK BOILER and PRICE NOT CONVERGED with NO PLAN FITS, and the challenge tile values the run at $26,665 against the $31,553 target with 4.2 h to go.

challenge, the controller read only the measurement packet. Stepped headless, a control step takes a median of 8.9 ms on case 5 (87 ms at the 95th percentile, 101 ms at most) and the centralized QP a median of 31.5 ms. In a desktop browser at the top speed of 0.25 h per second, an 18-hour case takes about 72 s. The challenge's two days take 196 s of wall time with every price round solved (a median of 9.5 ms per control step), or 6.6 s for the Fast-forward helper that runs the first day's same steps and 99 s for the second day.

### 6.6.8 Agreement With Published Relations

Agreement with published relations is a separate matter. The counter-flow effectiveness of Equation (6.4) agrees with the framework's implementation and with a marched exchanger of 20 000 cells at three points (0.377589 against 0.377592 at $N = 0.5$, $C_r = 0.25$, and 0.750000 three ways at $N = 3$, $C_r = 1$), and every substation gives its design duty with a 50.00 °C return at 95 °C and the design flow. The pipe loss coefficients are teaching values, but conduction through a polyurethane shell and the soil gives 0.43 to 0.58 W/(m K) per pipe for these bores, matching the branches' 0.40 to 0.45 and below the trunks' 0.60 to 0.80, as older pipe would be. The network loses 3.55 % of the heat produced in case 2. At a medium-weight construction's 165 kJ/(m$^2$ K) of floor, the buildings' heavy capacities imply 32.5 to 79.5 thousand m$^2$ and design loads of 44 to 92 W/m$^2$, the order of a block of existing buildings. The cases have not been evaluated with students, so this chapter makes no claim about learning.

## 6.7 Technology in Practice

### What Measures It

At the plant the supply and return temperatures and the flow give the heat sent out, and at each building a heat meter gives the heat taken. Such meters are inspected monthly for leaks, loud running, blocked strainers and, on an ultrasonic meter, the seat of its transducers and the state of the coupling gel, and they are calibrated each year or when the trended data show a drift [19]. A meter reports by a 4 to 20 mA signal, a fieldbus or calibrated pulses, and a pulse rate set too high can saturate the data system and lose counts [19]. The chapter's packet stands for these instruments: every five minutes it brings each consumer's supply, return, flow, pressure difference, heat meter and zone, with 0.1 K of noise on the temperatures and 0.5 % on the flows and meters.

> **TECHNOLOGY Sensor: The Heat Meter**
>
> A heat meter finds the energy delivered as the water's mass flow times the temperature difference between supply and return. A flow computer, built into the meter or separate, combines the flow reading with the two temperatures and integrates the energy over each interval. The flow element may be a differential pressure, turbine, vortex or ultrasonic meter. A transit-time ultrasonic meter times sound pulses sent with and against the flow, suits clean water and covers a flow range of about 20 to 1. Its coupling gel must be rated for the pipe's temperature or it can melt away and weaken the signal, and a meter on hot water must stay below saturation so that it does not cavitate [19]. The chapter's meters carry 0.5 % noise and never fail.

### What Moves It

The plant's final elements are the fuel to the cogeneration unit and the boiler, the charge and discharge of the storage tank, and the circulating pump, while each building moves its own substation valve. A substation valve is a throttle, and a throttle valve controls flow by raising the back pressure and dissipating the fluid's energy, which a pump on a variable speed drive avoids by slowing [20]. In the chapter the sources follow their setpoints through first-order lags, the tank's power is turned into a flow, the pump's head is set by a loop on the least measured pressure difference, and each valve passes the flow it is asked for up to its limit, as a pressure-independent valve does.

> **TECHNOLOGY Final Element: The Variable-Speed Circulation Pump**
>
> A pumped loop's flow can be set by a bypass line, a throttle valve, several pumps or the pump's speed. A bypass usually wastes the most energy, and a throttle valve raises the pump's back pressure and dissipates energy. Speed is the most efficient. A variable frequency drive, by far the most popular adjustable speed drive, changes the frequency of the motor's supply, and the pump's head and power curves drop down and to the left as it slows. Where static head is a large part of the total, a pump slowed too far can vibrate as it does against its shutoff head [20]. The chapter's pump sets a head of up to 600 kPa (Table 6.2) and costs nothing to run.

### Where the Controller Runs

A network spread over a town is run by supervisory control and data acquisition. A central control server polls its field devices at defined intervals, for example every 5 or 60 s, sends them new set points and watches for alarms, and supervisory and lower-level loops run with cycle times from milliseconds to minutes [21]. The chapter's plant steps every 10 s and its loops and planners every 5 min (Figure 6.5). The planners belong in the supervisory layer, and a price sent to each building is one of the simplest forms of distributed control in a heat network. Multi-agent schemes let consumer agents sell load shedding to producer agents within their comfort bounds [9].

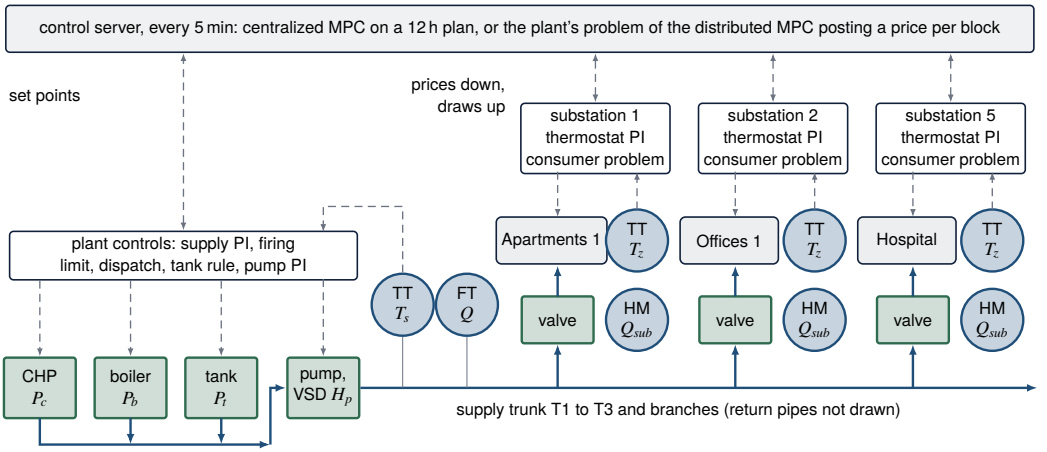



**Figure 6.5** The town's loops on field hardware, drawn from Sections 6.3 and 6.5. Green boxes are final elements, blue circles transmitters (TT temperature, FT flow, HM heat meter), dashed arrows signals, and three of the six substations are drawn. Notice that the price travels to the substations at once, while the heat it buys travels through the trunk for hours.

> **IN PRACTICE** **Four Loops of Classic Network Control**
>
> Classic network control has four parts. The circulating pump holds the pressure difference between supply and return high enough for flow at the network's far edge. Each building's heat demand control meets its comfort and hot-water needs, and flow control opens the substation valve far enough to pass that heat. The plant holds the supply temperature on a curve of the outdoor temperature. Pressure and supply control are central, and the other two sit in the substations. This control is robust and easy to extend, but in a heat deficit the consumers near the plant take what they need and the far ones go short [9]. The chapter's regulatory layer and supply curve are these four loops.

### What Is Modernizing It

Three developments bear on the chapter's method. Large water tanks beside a cogeneration plant cut the peak and let the plant run at its best ratio of heat to power, and one simulated network saved up to 12 % of its primary energy that way. Hybrid control, in which a central optimization sets the load profile and agents in the buildings carry it out, is gaining interest and was judged very promising. Smart meters installed for electricity that also measure heat use and room temperatures could share the cost of the communication network that an advanced controller needs [9]. The chapter's centralized and distributed MPCs are the two ends of that range.

## 6.8 Limitations

The pipes, buildings and prices are teaching values. The loss coefficients are of the order conduction predicts and the buildings of a plausible order, but neither is fitted to a network, so the delays and costs are those of a declared town. Both MPCs plan with the delays frozen at the measured flows and the return pipes held, so a front can arrive sooner or later than planned when the flow changes, which the next solve five minutes later corrects. The CHP's lag is not in the plan. The comfort the MPC runs report includes the consumers' guard, which the plan does not see, and the suggested setting's pass in the challenge depends on it. The planning model carries the buildings' declared constants, which are the plant's in every case but case 6, so model error appears only through the frozen network, the Hankel predictor and a plant that a case or the learner changes.

**OPEN CHALLENGE Sensing a Whole Network**

An advanced network controller needs the valve positions, the flows and temperatures in the pipes, and the state of the buildings, the plants and the stores, possibly several times an hour. A network can span a street or a region, so the sensors and the communication network cost much to install and to run. Estimators and hierarchical agents can cut the data sent, and the fewest sensors that reveal a network's state, and where to place them, remain open questions [9]. This chapter's controllers read every consumer every five minutes. Progress would be plans that keep their quality on a few well-placed meters.

The fixed-step price iteration settles at most steps of a cold morning but does not converge where a capacity binds hard or the plant is short. At the round cap the reported gap is only an upper bound, and the recovered plan exceeded the optimum by up to 604 % at one step. When no plan fits, the applied plan is the plant's last answer to unsettled prices, kept from charging the tank at the consumers' expense by a rule, not a plan. The graded margin rests on one weather draw (passed on five of seven) and on a valuation rule, both ways at $74.5 per MWh, that the other cases do not use. Their rule values deficits at the boiler's price and surpluses at the CHP's, so their valued costs are conservative rather than neutral to the end state. A simulated pass certifies no network, contract or plant.

**OPEN CHALLENGE Charging the Pipes Ages the Pipes**

Raising and lowering the water temperature stores a limited amount of heat in a network and shaves a peak. Cycling more often between high and low temperatures may speed fatigue in steel pipes, and cracks can grow until a pipe fails. A simple and accurate way to state how much flexibility a network and its buildings offer is still lacking [9]. The chapter's plan charges a supply move only 1 $/K$^2$ and counts no fatigue. Progress would be a plan that weighs the fuel saved against the life of the pipe.

## 6.9 Conclusion

A town's heat network delays its heat by the flow its consumers draw, and this simulator measures what that does to four ways of running the plant. The same 15 K step reaches the far end in 4.48 h at 50 kg/s and 2.83 h at 78 kg/s, because every segment passes its volume at its own flow. A boost timed by hand arrives with the peak and still costs $202 more than the curve, while a centralized MPC that charges the network on the CHP's spare evening heat saves 3.2 %. On the cold morning the price-coordinated run costs within $52 of the centralized one, and where the price settles it equals the centralized multiplier, but it does not converge when the plant's capacity binds, and after a trip it stays two orders below the price that would ration the heat. On the graded cold snap the distributed MPC saves 1.59 % with its stores valued and passes on five of seven weathers. The model is declared and its numbers are teaching values. A fuller model would add front dispersion, pumping energy, a CHP with a minimum load, an accelerated or scheduled price step, and consumers whose models differ from the plant's. The town makes a price of heat something a class can post, watch settle, and watch fail.

**KEY POINTS**

- A town's heat network delays its heat by the flow its consumers draw.
- The same 15 K step reaches the far end in 4.48 h at 50 kg/s and 2.83 h at 78 kg/s.
- A boost timed by hand arrives with the peak and still costs $202 more than the curve, while a centralized MPC that charges the network on the CHP's spare evening heat saves 3.2 %.
- Where the price settles it equals the centralized multiplier, but it does not converge when the plant's capacity binds.

**Where else this applies.** Plug flow keyed by volume also carries the chip column of Volume 5, Chapter 4 (Section 6.3), and plans shared over a network with delays also run the gates of Chapter 5 (Section 6.5).

## References


[1] H. Everett, "Generalized Lagrange multiplier method for solving problems of optimum allocation of resources," *Operations Research*, vol. 11, no. 3, pp. 399–417, 1963.

[2] G. B. Dantzig and P. Wolfe, "Decomposition principle for linear programs," *Operations Research*, vol. 8, no. 1, pp. 101–111, 1960.

[3] R. Scattolini, "Architectures for distributed and hierarchical model predictive control – a review," *Journal of Process Control*, vol. 19, no. 5, pp. 723–731, 2009.

[4] J. B. Rawlings and B. T. Stewart, "Coordinating multiple optimization-based controllers: New opportunities and challenges," *Journal of Process Control*, vol. 18, no. 9, pp. 839–845, 2008.

[5] A. Benonysson, B. Bøhm, and H. F. Ravn, "Operational optimization in a district heating system," *Energy Conversion and Management*, vol. 36, no. 5, pp. 297–314, 1995.

[6] H. V. Larsen, H. Pálsson, B. Bøhm, and H. F. Ravn, "Aggregated dynamic simulation model of district heating networks," *Energy Conversion and Management*, vol. 43, no. 8, pp. 995–1019, 2002.

[7] I. Gabrielaitiene, B. Bøhm, and B. Sunden, "Modelling temperature dynamics of a district heating system in Naestved, Denmark—a case study," *Energy Conversion and Management*, vol. 48, no. 1, pp. 78–86, 2007.

[8] B. van der Heijde, M. Fuchs, C. Ribas Tugores, G. Schweiger, K. Sartor, D. Basciotti, D. Müller, C. Nytsch-Geusen, M. Wetter, and L. Helsen, "Dynamic equation-based thermo-hydraulic pipe model for district heating and cooling systems," *Energy Conversion and Management*, vol. 151, pp. 158–169, 2017.

[9] A. Vandermeulen, B. van der Heijde, and L. Helsen, "Controlling district heating and cooling networks to unlock flexibility: A review," *Energy*, vol. 151, pp. 103–115, 2018.

[10] G. Sandou, S. Font, S. Tebbani, A. Hiret, and C. Mondon, "Predictive control of a complex district heating network," in *Proceedings of the 44th IEEE Conference on Decision and Control*, 2005, pp. 7372–7377.

[11] P. Giselsson, M. D. Doan, T. Keviczky, B. De Schutter, and A. Rantzer, "Accelerated gradient methods and dual decomposition in distributed model predictive control," *Automatica*, vol. 49, no. 3, pp. 829–833, 2013.

[12] H. Lund, S. Werner, R. Wiltshire, S. Svendsen, J. E. Thorsen, F. Hvelplund, and B. V. Mathiesen, "4th Generation District Heating (4GDH)," *Energy*, vol. 68, pp. 1–11, 2014.

[13] S. Werner, "International review of district heating and cooling," *Energy*, vol. 137, pp. 617–631, 2017.

[14] American Society of Mechanical Engineers, "Holly district heating system, landmark 122," https://www.asme.org/about-asme/engineering-history/landmarks/122-holly-district-heating-system, landmark ceremony 1987; page read 2026-11-22.

[15] G. P. Henze, C. Felsmann, and G. Knabe, "Evaluation of optimal control for active and passive building thermal storage," *International Journal of Thermal Sciences*, vol. 43, no. 2, pp. 173–183, 2004.

[16] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy and Buildings*, vol. 43, no. 7, pp. 1511–1522, 2011.

[17] J. Coulson, J. Lygeros, and F. Dörfler, "Data-enabled predictive control: In the shallows of the DeePC," in *2019 18th European Control Conference (ECC)*, 2019, pp. 307–312.

[18] J. C. Willems, P. Rapisarda, I. Markovsky, and B. L. M. De Moor, "A note on persistency of excitation," *Systems & Control Letters*, vol. 54, no. 4, pp. 325–329, 2005.

[19] S. A. Parker, W. D. Hunt, K. McMordie Stoughton, B. K. Boyd, K. M. Fowler, T. M. Koehler, W. F. Sandusky, G. P. Sullivan, and R. Pugh, "Metering best practices: A guide to achieving utility resource efficiency, release 3.0," Pacific Northwest National Laboratory for the U.S. Department of Energy Federal Energy Management Program, Richland, WA, Tech. Rep. PNNL-23892, 2015.

[20] U.S. Department of Energy, "Improving pumping system performance: A sourcebook for industry, second edition," U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Industrial Technologies Program, Washington, DC, Tech. Rep. DOE/GO-102006-2079, 2006.

[21] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication 800-82r3, 2023.

## 6.10 Exercises

**6.1** Concept check · 2 min

In Scenario 1 the 15 K step leaves the plant at 13:05 with 50 kg/s flowing and reaches the hospital 4.48 h later, although the 598 $m^3$ of pipe on its path would pass at the plant's whole flow in about 3.2 h. Why does the step take longer?

(a) The pipe loses heat on the way, so the front cools and is slow to rise by half the step

(b) Less water passes each junction, so each later segment is crossed at its own smaller flow

(c) The pump's head falls in the afternoon, which slows the water in every segment by one factor

(d) The front smears along the path, so its middle arrives well after its leading edge

**6.2** Concept check · 1 min

In Scenario 3 a 15 K boost leaves at 04:00 and reaches the hospital at 06:53, as the morning peak arrives, yet the valued cost rises $202 above the supply curve's. Why?

(a) The front arrives after the peak has passed, so the heat stored in the pipes is wasted

(b) The hotter supply pushes the returns past their limit, and the limiters cut the duty

(c) The CHP was already at its capacity, so the charge was bought from the peak boiler

(d) The boost switches off the tank's rule, so the tank no longer spares the boiler

**6.3** Concept check · 2 min

In Scenario 8 the Hankel predictor, built only from Apartments 1's own record, predicts the block's zone 6 h ahead with an error of 0.43 K rms in open loop, against 0.13 K for the two-capacity model. Why does the chapter call the two-capacity result a bound rather than a fair rival?

(a) It is fitted to the same record with fewer parameters, so it cannot overfit the data

(b) It is rebuilt every hour from the last 36 h, so it always has the newest data

(c) It predicts with the true outdoor temperature, which the Hankel predictor never sees

(d) It has the plant's own structure and declared constants, so no fitted model can do better

**6.4** Theory and calculation · 6 min

The hospital's path runs through trunks T1, T2 and T3 and its own branch, 598 $m^3$ in all. Use Table 6.2 ($\rho = 970\,\mathrm{kg/m^3}$) and the flows of case 1 at 13:05: 50 kg/s in T1, 34.9 in T2, 19.2 in T3 and 9.9 kg/s in the hospital's branch.

(a) Find the volumes of T1, T2 and T3 from their lengths and bores, and the branch's volume as what remains of the 598 $m^3$. *Checkpoint:* 318, 192, 74 and 14 $m^3$.

(b) Find each segment's residence at its 13:05 flow, held frozen, and their sum. *Checkpoint:* 1.71, 1.49, 1.03 and 0.38 h, 4.61 h in all.

(c) Find the path's volume over the plant's whole flow. Compare both results with the measured 4.48 h, and say why the frozen sum comes out late. *Checkpoint:* About 3.22 h.

**6.5** Theory and calculation · 6 min

Trunk T1 is 2.0 km of 0.45 m bore with a loss of 0.80 W/(m K) per pipe, buried in ground at 5 °C.

(a) Find T1's loss time constant $\tau_L$ of Equation (6.1). *Checkpoint:* About 224 h.

(b) Water enters T1 at 95 °C and spends 1.71 h in it at 50 kg/s. How much does it cool on the way? Check the result with Equation (6.2). *Checkpoint:* About 0.69 K.

(c) Find the heat T1's supply pipe loses with water at 95 °C along its length, and at 110 °C. What does the difference amount to over a boost of three hours? *Checkpoint:* 144 and 168 kW, about 0.07 MWh more over three hours.

**6.6** Theory and calculation · 7 min

Use the fuel prices of Table 6.2 with the condenser factor $f = 1$.

(a) Find the CHP's marginal cost $\pi_c + \gamma_c P_c$ at its full 14 MW and compare it with the boiler's base price. *Checkpoint:* 51.4 \$/MWh against 52 \$/MWh.

(b) Find the boiler's marginal cost at the challenge's 15 MW cap, and say why the challenge values every MWh its stores end short or over at this price. *Checkpoint:* 74.5 \$/MWh.

(c) After the challenge's trip the price sits at a median of \$398 per MWh while the centralized multiplier is a median of \$76,993. With $\alpha = 0.5$ and an imbalance of 2.2 MW in every round, how many rounds, how many control steps of 30 rounds and how many hours would the price need to climb that far? How far could it climb in the 55 control steps after the trip? *Checkpoint:* About 69 600 rounds, 2320 steps and 193 h. About 1800 \$/MWh.

**6.7** App challenge · 8 min

Load Scenario 1, *The delay at two flows*, at its default seed and change nothing. Run at 0.25 h per second until the run pauses at 06:00 ($t = 18$ h, about 72 s) and read the *Supply step arrived* tile. Then set Boost start to 2, press Reset and run to the pause again.

(a) For each run give the hospital's arrival time (H) and the plant's flow when the step left, to 0.02 h and 1 kg/s. *Checkpoint:* 4.48 h at 50 kg/s. 2.83 h at 78 kg/s.

(b) Find the ratio of the two hospital delays and the ratio of the two flows. Scale the afternoon delay by the flow ratio and compare with the night run. *Checkpoint:* 1.58 and 1.56, 2.87 h against 2.83 h.

(c) Read Apartments 1 (A1) and Apartments 2 (A2) in both runs. Why does every consumer's delay shrink by nearly the same factor? *Checkpoint:* A1 2.06 and 1.34 h, A2 3.49 and 2.21 h.

**6.8** App challenge · 9 min

Load Scenario 2, *The supply curve on a cold morning*, at its default seed and change nothing. Run at 0.25 h per second until the run pauses at 12:00 ($t = 18$ h, about 72 s) and read the tiles *Fuel cost*, *Fuel cost, stores valued* and *Peak boiler*. Do the same with Scenario 4, *Centralized MPC: when to send heat*, and Scenario 5, *The price converging, and a step that does not*, as loaded.

(a) Give case 2's boiler energy, its peak and its valued cost. *Checkpoint:* 28.2 MWh, 10.9 MW, \$10,282.

(b) Give the boiler energy and the valued cost of cases 4 and 5, and each one's saving against case 2 in dollars and percent. *Checkpoint:* Case 4 19.4 MWh and \$9,953, \$329 or 3.2 %. Case 5 17.8 MWh and \$9,901, \$381 or 3.7 %.

(c) From case 4's *Fuel cost* and the tank, pipes and buildings on its valued tile, rebuild the valued cost with Equation (6.14), deficits at the boiler's 52 \$/MWh and surpluses at the CHP's 22 \$/MWh. *Checkpoint:* \$9,696, tank −3.2, pipes −4.3, buildings +6.0 MWh: \$9,954.

**6.9** App challenge · 10 min

Load the challenge, *Challenge: a two-day cold snap and a CHP trip*, at its default seed.

(a) Change nothing. Press Fast-forward day 1, then run at 0.25 h per second to the end at $t = 48$ h (about 100 s). Read the tiles *Challenge target*, *Fuel cost* and *Fuel cost, stores valued*. *Checkpoint:* Not met, \$31,952 of \$31,553. Bill \$31,580, stores −5.0 MWh.

(b) Set Strategy to Distributed MPC, leave Price step at 0.5, press Reset, press Fast-forward day 1 and run to the end again. Read the same tiles. *Checkpoint:* Met, \$31,445 of \$31,553. Bill \$30,334, stores −14.9 MWh.

(c) Show that each graded cost is the bill plus the stores' shortfall at 74.5 $/MWh. Give the distributed MPC's saving against the curve on the bill and on the graded cost, and say which one the target uses and why. *Checkpoint:* $31,952 and $31,444. 3.9 % on the bill, 1.59 % graded.

**6.10** Application area · 4 min

In Scenario 6 Apartments 2 has radiators half the size its load needs, and its return limiter is off. Its return reaches 68.7 °C and lifts the plant's mean return from 35.4 to 42.8 °C.

(a) Evaluate the condenser factor $f(T_{rp})$ of Equation (6.7) at the two mean returns, and say what each kelvin of return above 40 °C costs. *Checkpoint:* 1.000 and 1.017. About 0.6 % more fuel per kelvin.

(b) With the limiter on, the plant is spared. What does Apartments 2 pay instead, and what cure would a network engineer ask the owner for? *Checkpoint:* 7.3 h more than 0.5 K below its band, at worst 1.15 K.

**6.11** Application area · 3 min

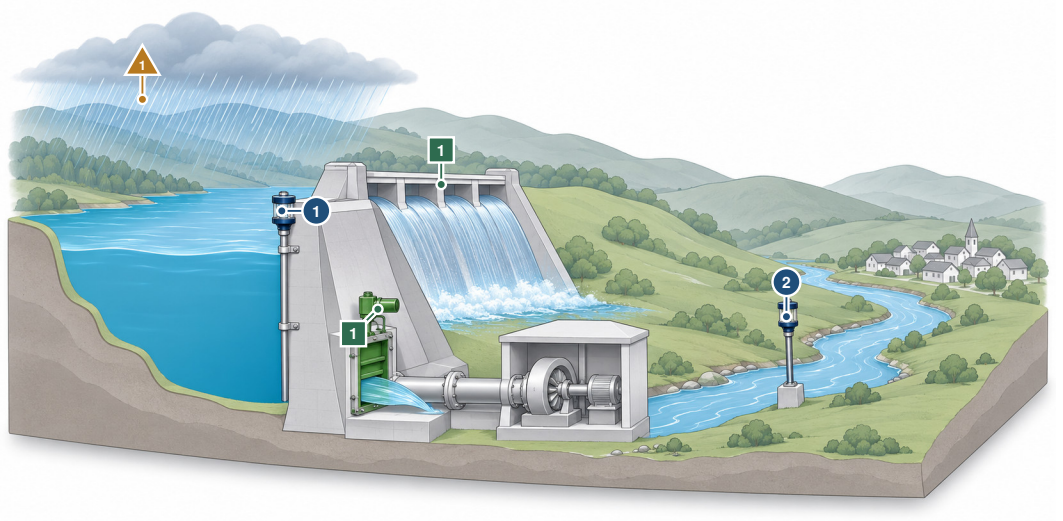
In Scenario 10 the CHP trips at midnight as the cold snap begins. Under the shipped distributed MPC the offices fall 4.67 K below their band, the apartments 2.76 K and the hospital 2.16 K, and none of fourteen settings a learner can reach holds the hospital and every apartment block within 0.5 K. As the network's engineer, what would you propose, and what does the result not prove?

**6.12** Review the agent · 5 min

An AI assistant reviews Scenario 5 and writes: *"At the shipped price step of 0.5, 39 of the 216 control steps stop at the 30-round cap, so the price moves too slowly. Raise the Price step to 1.5 and the price will reach its settled value in a third of the rounds, every step will converge and the run will cost less."* Find the flaw in the claim and say how you would check it in the app.

CHAPTER 7

# Reservoir Flood Control on an Ensemble Forecast With Scenario Trees




**The controller moves** 1 the release, through the outlets and the spillway gates

**The controller reads** 1 lake level, read as storage 2 river gauge at the town

**What disturbs it** 1 storm rain on the hills above the lake

*Water can leave the lake only as fast as the river below can carry it, so room for a storm must be made days before the rain. Illustration, not to scale.*

IN PLAIN TERMS **Making Room in a Lake Before the Rain**

**The main idea.** A dam above a town keeps part of its lake empty in winter to catch a flood. Water can leave the lake only as fast as the river below can carry it. So the room must be made days before the storm.

**The challenge.** The operator decides on a forecast that gives several possible futures. Sometimes the promised storm never falls. Planning for the wettest future wastes water, and planning for a milder one can leave too little room.

**Why it matters.** Every dam that guards a town must choose how much flood risk to accept for the water and power it keeps.

THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 7.2, the try-it segments in Sections 7.4 and 7.6, the two facts, and the key points |
| **Study** | 50 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 20 min | the first page, the results (Section 7.6), Technology in Practice (Section 7.7), Limitations (Section 7.8) and the open challenges |

**The control question** A reservoir empties its flood pool only as fast as the channel below the dam carries water, so it must release days before a storm, on a forecast that sometimes promises a storm that never comes. What do a rule curve, a value-function policy and a scenario tree each cost and buy in flood safety, supply and energy?

**Learning objectives**

- Compute how fast a channel-limited release empties a flood pool, and the lead that rate demands.
- Explain why a tributary below the dam floods the town before the pool is full.
- Describe what a scenario tree commits today and how its chance constraint lets the wettest member go.
- Compare three deciders over a water year in days over the limit, supply and energy.
- Judge an energy result against a perfect-foresight bound at the run's own end storage.

**Prerequisites** Sections 1.4.6 (slack prices), 1.4.8 (terminal value), 1.4.9 (dynamic programming), 1.4.12 (scenario tree, chance constraint) and 1.6.4, and Chapter 5 for releasing ahead of a delay.

**The app** https://apmonitor.com/apps/reservoir-flood-control/

**See also** Volume 5, Chapter 4 (every useful move is a prediction), Volume 5, Chapter 5 (a quality fixed before it is measured, and a supervisor that says when no move can meet the limit) and Volume 5, Chapter 9 (a slow inventory refilled through a small, capped valve).

## 7.1 Introduction

On 13 February of the simulated water year a 440 mm atmospheric river begins to fall on the 2500 $km^2$ catchment above an 1100 $hm^3$ reservoir. Operated on its rule curve, the lake meets the storm at 798 $hm^3$, two cubic hectometers below its winter target, rises above its crest to 1137 $hm^3$, spills 111 $hm^3$ over it, and on 21 February the river at the town downstream breaks a levee declared to fail above 650 $m^3/s$ with 666 $m^3/s$. On the same weather a chance-constrained scenario tree that reads the morning ensemble raises the release from 12 to 440 $m^3/s$ between 8 and 12 February, meets the storm at 688 $hm^3$, fills to 1095 $hm^3$ without touching the crest, and never lets the town's river exceed 425 $m^3/s$ against a channel limit of 450 $m^3/s$. What separates the two in February is what each does in the five days before the rain.

Reservoir operation has been posed as an optimization problem since the 1950s [1], and its methods have been reviewed repeatedly [2–4]. The control problem underneath it is general. An inventory receives an uncertain inflow, a downstream unit accepts only a limited rate, and the only defense against a large inflow is room made before it arrives. A stormwater basin, a surge tank ahead of a treatment plant, a buffer between a variable feed and a capacity-limited unit and a battery dispatched against a solar forecast share the shape. Volume 17, Chapter 10 dispatches such a battery against a forecast of the net load, and Volume 17, Chapter 11 plans the inventory of an air store with a water value. Chemical process control met the same structure in chance-constrained predictive control [5] and in multi-stage predictive control of a polymerization reactor on a tree of uncertain futures [6], both instances of multi-stage stochastic programming [7].

This chapter tests two beliefs. The first is that a reservoir floods its town because its flood pool is full. In case 2 the town's river first exceeds the channel limit on day 17.4 with the lake at 866 $hm^3$, 234 $hm^3$ below the crest, because the tributary between the dam and the town rises first and the dam's release reaches the town a day after it is made, and a flood pool enlarged from 300 to 400 $hm^3$ leaves the town's peak at the same 575 $m^3/s$. The second is that the cautious use of a forecast is to plan for its wettest member. In case 4 a storm that the forecasts promise from day 1 for day 10 never falls, and a rule that chases the wettest member sends 23 $hm^3$ through the gates for nothing, against 2 $hm^3$ for a rule that hedges on the median member.

These ideas usually meet a student in different courses. Stochastic dynamic programming is taught with water resources planning [4, 8], ensemble forecasts with hydrology [9], and scenario trees and chance constraints with optimization [7, 10]. This chapter runs a rule curve, a stochastic dynamic

program and a chance-constrained scenario tree on one declared reservoir and one ensemble forecast in a browser simulator, and measures what each one costs and buys in flood safety, supply and energy. The value function [1, 8], tree-based predictive control on an ensemble [11, 12] and the discarding of scenarios for a chance constraint [10] are established methods. The measurements are made on one declared plant and one weather year, and among them is the finding that the graded water year is passed by forecasting and not by emptying the lake. The simulator underneath has a 19-state model whose water balance closes to $10^{-10}$ hm$^3$, a value function on 73 000 states solved in under a second in a browser and checked against an independent Bellman backup, a tree linear program checked against brute force, and a measurement boundary that keeps every decider away from the true future. Ten cases carry the measurements, each with a completion rule in the app (Section 1.6.4), and the exercises at the end of the chapter are built on them.

Section 7.2 follows the domain from the operating diagram to ensemble forecasts. Section 7.3 declares the basin from the rain to the town, and Section 7.4 shows why the pool must be emptied before the storm. Section 7.5 sets out the deciders: the rule curve with its forecast pre-release, the stochastic dynamic program, the scenario tree with its chance constraint and fallbacks, and a perfect-foresight bound. Section 7.6 measures the ten cases, the last of them a graded water year, and verifies the code against its own equations and against published hydrology. Section 7.7 describes the gauges, gates and forecasts of real operation, Section 7.8 states what the model leaves out, and Section 7.9 draws the results together. The scenario tree, its shared first decisions and the chance constraint are introduced on the heated tank of Section 1.4.12, and the value function and its slope in Section 1.4.9.

## 7.2 Flood Pools, Value Functions and Ensemble Forecasts

Multipurpose dams have long been run from an operating diagram drawn when the dam was designed. A rule curve gives the target storage for each day of the year, the space between the target and the top of the dam is reserved for floods in the wet season and returned to supply and power in the dry season, and a flood rule releases as much as the channel downstream can carry whenever the lake rises into that reserve [4]. The diagram needs no forecast and is easy to audit.

> ⧖ HISTORY **1959: A Guide Curve Drawn Without Forecasts**
>
> At one flood-control reservoir in California, the water control manual was issued in 1959, with minor revisions in 1986, without the benefit of modern weather and streamflow forecasts. Its guide curve sets the storage during a flood and the safe releases soon after it that make room for the next one, and it rests on the historical pattern of wet winters and dry summers, not on the weather in sight. Decades later a group of agencies tested releases recommended from a 15-day ensemble streamflow forecast, as approved deviations from that manual. In the 2020 water year, the third driest in a 127-year record, the lake ended the winter with 19 percent more water stored [13]. This chapter's rule curve plays the part of such a diagram, and its trees play the forecast.

The optimization of reservoir releases under uncertain inflow has early roots in hydroelectric storage. In 1955 Little treated the use of storage water in a hydroelectric system as a sequence of decisions under a random inflow [1]. Stochastic dynamic programs with inflow classes in a Markov chain were then set out as reservoir operation models [8], and reviews in 1985 and 2004 record how linear, dynamic and nonlinear programming and simulation were applied to single reservoirs and to systems of them [2, 3]. The control of a water reservoir under economic, social and environmental constraints was set out for the automatic control community in 2008 [14].

Forecasts changed what the optimizer could know. Ensemble streamflow prediction conditioned a sampling stochastic dynamic program on the forecast traces [15], and flood forecasting driven by ensembles from numerical weather prediction was reviewed in 2009 [9]. Short-term operation then moved to receding-horizon optimization over a tree of the ensemble members, with the first decisions shared by every member [11], and real-time flood control of a reservoir by tree-based predictive control has been reported with forecast uncertainty included [12]. The chance constraint

**Table 7.1** Variables by Role, With Their Values on 17 February of the Challenge on the Suggested Settings

| Variable | Role | Value | Range |
|---|---|---|---|
| $Q_{town}$ | CV, true flow at the town | 418 $m^3/s$ | limit 450, levee 650 |
| $S$ | CV, storage, measured exactly | 824 $hm^3$ | 125–1100 |
| $R$ | MV, release through the outlets | 124 $m^3/s$ | 0–1780 $m^3/s$ |
| $Q_t$, $Q_g$ | split, turbines then gates | 124, 0 $m^3/s$ | $Q_t \leq 180$ |
| $Q_c$ | output, crest spill | 0 $m^3/s$ | above 1100 $hm^3$ |
| $P$, $P_L$ | DV, rain (teaching truth) | 8.2 mm/d | storms declared |
| $I$ | DV, inflow to the lake | 1087 $m^3/s$ | gauge ±2 %: 1064 |
| $L$ | DV, local tributary | 348 $m^3/s$ | teaching truth |
| $\hat{I}_{m,d}$, $\hat{L}_{m,d}$ | measured, ensemble | $M = 7$, 10 days | issued daily |
| $T(d)$ | SP, rule curve target | 800 $hm^3$ | 800–1000 |
| $D$, $D_s$ | supply diverted, scheduled | 6, 6 $m^3/s$ | 6–28 $m^3/s$ |
| $I_r$ | river below the diversion | 118 $m^3/s$ | $\geq Q_{env} = 6$ |
| $E$ | evaporation | 0.03 $hm^3/d$ | — |
| $H$, $P_e$ | head, power | 88 m, 94 MW | to 100 m, 156 MW |
| $n_{1..3}$, $g$ | states, catchment stores | — | mm |
| $W$ | state, water in the reach | — | $m^3/s\,d$ |

arrived from process control, where it states the probability with which an output limit must hold [5], and from optimization, where discarding a declared number of sampled scenarios gives a tractable approximation [10]. The hydrology underneath is older. Nash derived the unit hydrograph of a catchment by routing rain through a cascade of equal linear reservoirs [16], and the Muskingum storage law for a river reach was related to the physics of flood propagation by Cunge [17]. The simulator places the rule curve, the value function and the tree on one basin built from those two models.

> ✦ **DID YOU KNOW? Sixty-One Members on an Hourly Step**
>
> The streamflow ensembles prepared for one reservoir's forecast-based operation run 15 days ahead on an hourly step, and the hindcast used to test them carries 61 members [13]. This chapter's forecast issues $M = 7$ members for ten days, one value a day (Table 7.1).

## 7.3 A Declared Basin From Rain to the Town

The simulated basin is a declared teaching design with no real river, dam or operator behind it, as shown in Figure 7.1. Rain falls on a catchment of $A = 2500\,km^2$ above the dam and on a local catchment of $A_L = 600\,km^2$ between the dam and the town. The lake holds 1100 $hm^3$ at its uncontrolled crest and 120 $hm^3$ below its intake. The dam releases through a 180 $m^3/s$ powerhouse and a gated spillway, a diversion below the powerhouse takes the supply schedule, an environmental flow stays in the river, and a one-day reach carries the rest to a town whose channel carries 450 $m^3/s$ and whose levee fails above 650 $m^3/s$. The operator decides one release a day. The time base is the day, volumes are in cubic hectometers (1 $hm^3$ is $10^6$ $m^3$), flows in $m^3/s$, rain in mm/d, power in MW, and 1 $m^3/s$ held for a day is 0.0864 $hm^3$.

Each variable carries one role in Table 7.1. Nineteen states are integrated: the eight cascade and groundwater stores of the two catchments, the storage $S$, the water $W$ in the reach, the energy generated, and eight running volumes that exist so the water balances can be checked exactly.

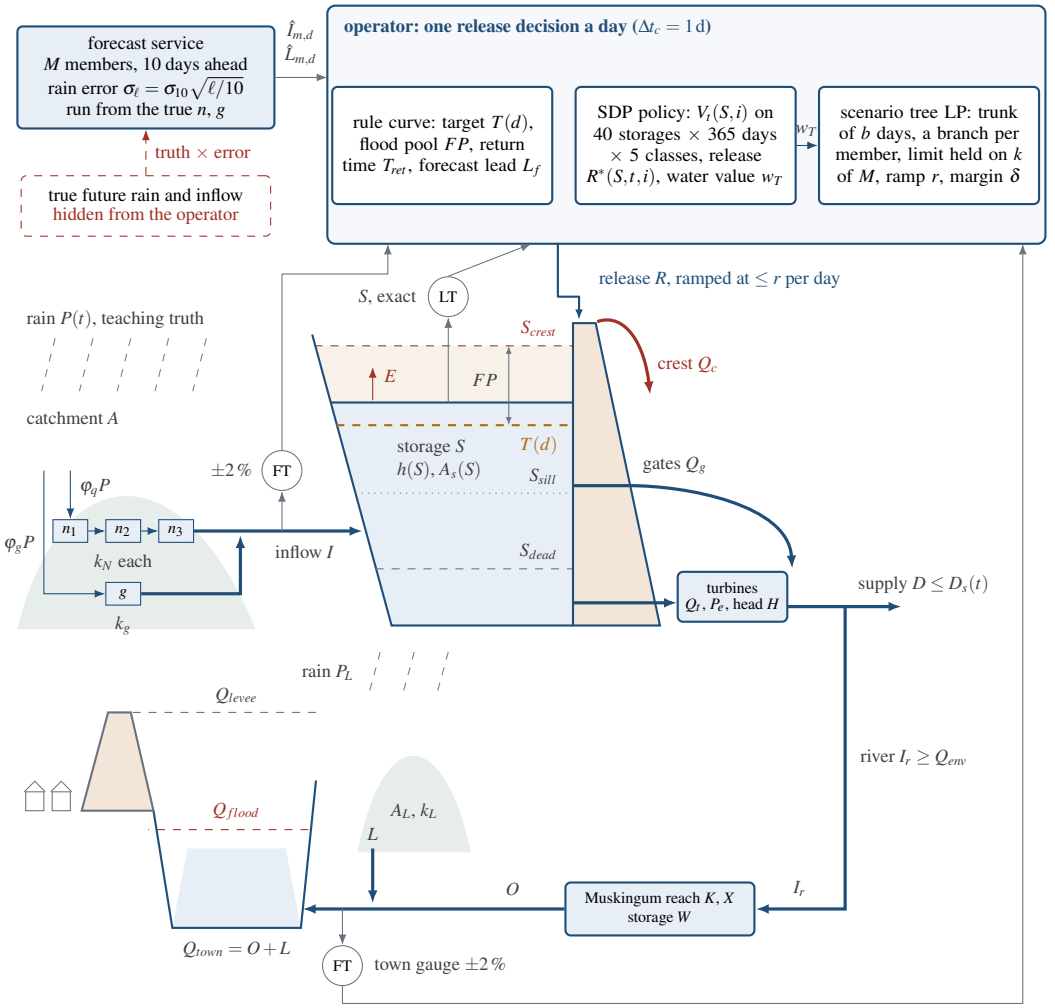



**Figure 7.1** Reservoir, river and town as modeled, inside the daily decision loop, after the application's Basin view, with the river below the dam turned back so that the town sits at the lower left. Rain $P$ (teaching truth) fills the cascade stores $n_1$ to $n_3$ (constant $k_N$) and the groundwater store $g$ ($k_g$) of the catchment $A$, giving the inflow $I$. The lake holds storage $S$ at depth $h(S)$ and area $A_s(S)$ between the dead pool $S_{dead}$ and the crest $S_{crest}$, loses $E$ to evaporation, and keeps the flood pool $FP$ above the rule curve $T(d)$. The turbines pass $Q_t$ at head $H$ for power $P_e$, the gates $Q_g$ above the sill $S_{sill}$, and the crest $Q_c$. The diversion takes $D$, the river keeps $Q_{env}$, the reach ($K$, $X$, storage $W$) gives $O$, and the tributary ($A_L$, $k_L$) adds $L$ to make $Q_{town}$ against $Q_{flood}$ and $Q_{levee}$. The operator reads $S$ exactly, the two gauges to ±2 % and the ensemble, never the true future. Not to scale.

### 7.3.1 Rain and Runoff

Each catchment is a cascade of three equal linear reservoirs of quick runoff beside a groundwater store that feeds the base flow, as shown in Equation (7.1).

$$\begin{aligned} \frac{dn_1}{dt} &= \varphi_q P - \frac{n_1}{k_N}, & \frac{dn_i}{dt} &= \frac{n_{i-1} - n_i}{k_N} \quad (i = 2,3) \\ \frac{dg}{dt} &= \varphi_g P - \frac{g}{k_g}, & I &= \frac{A}{86.4}\left(\frac{n_3}{k_N} + \frac{g}{k_g}\right) \end{aligned} \tag{7.1}$$

where the stores are in mm, $P$ is in mm/d, $A/86.4$ converts mm/d over $A$ in km$^2$ to m$^3$/s, and the local catchment has the same form with $P_L$, $A_L$, $k_L = 1.0\,\mathrm{d}$ and outflow $L$. Of the rain, $\varphi_q = 0.40$ runs off quickly and $\varphi_g = 0.25$ recharges the groundwater, which drains with $k_g = 40\,\mathrm{d}$. The response of the cascade to a unit impulse of quick runoff is Nash's gamma-shaped unit hydrograph, Equation (7.2) [16].

$$u(t) = \frac{1}{k_N\,\Gamma(n)} \left(\frac{t}{k_N}\right)^{n-1} \mathrm{e}^{-t/k_N} \tag{7.2}$$

With $n = 3$ and $k_N = 1.5\,\mathrm{d}$ it peaks at $(n-1)k_N = 3\,\mathrm{d}$ with a mean lag of $nk_N = 4.5\,\mathrm{d}$, while the smaller local catchment peaks at 2 d. The tributary below the dam therefore rises a day before the reservoir's own inflow on the same storm.

### 7.3.2 The Lake, Its Outlets and Its Crest

The storage integrates the inflow less the three outflows and the lake's evaporation, the depth at the dam follows a valley's storage curve, and the turbines produce power on the head above a fixed tailwater, as shown in Equation (7.3).

$$\begin{aligned} \frac{\mathrm{d}S}{\mathrm{d}t} &= 0.0864\,(I - Q_t - Q_g - Q_c) - E(S,t) \\ h &= 6.69\,S^{0.4}, \qquad H = h - 10, \qquad A_s = \frac{S}{0.4\,h} \\ E &= 10^{-3} e_{m(t)} A_s, \qquad P_e = 9.81 \times 10^{-3}\,\eta\,Q_t H \end{aligned} \tag{7.3}$$

where $h$ and $H$ are in m, $A_s$ in km$^2$, $e_{m(t)}$ is the month's evaporation of 0.8 to 7.0 mm/d, $\eta = 0.88$, the factor $9.81 \times 10^{-3}$ is the weight of water in MN per m$^3$, and $P_e$ is in MW. At the crest the head is 100.1 m and the full powerhouse makes 155.6 MW. The release $R$ follows the operator's command as a straight line in time, from the release at the last decision toward the new target at the declared rate, and is capped by what the outlets pass at the present storage, as shown in Equation (7.4).

$$\begin{aligned} R &= \min\left(R_{cmd}(t),\ R_{cap}(S)\right) \\ Q_t &= \min(R, Q_{t,max}), \qquad Q_g = R - Q_t \\ R_{cap} &= c(S)\left[Q_{t,max} + \gamma \min\left(Q_{s,full},\ a\,(S - S_{sill})^{+}\right)\right] \\ Q_c &= C_w \left(h(S) - h(S_{crest})\right)^{3/2} \quad (S > S_{crest}) \end{aligned} \tag{7.4}$$

where $a = Q_{s,full}/(S_{crest} - S_{sill}) = 2.29\,\mathrm{m^3/s}$ per hm$^3$ makes the gated rating linear from the sill at 400 hm$^3$ to 1600 m$^3$/s at the crest, $\gamma$ is 1 or 0.5 after a gate fault, $c(S)$ falls from 1 to 0 over the 5 hm$^3$ above the dead pool as the intake uncovers, and $C_w = 255\,\mathrm{m^{1.5}/s}$ is a 150 m weir. The turbines take the release first and the gates the rest, and the crest spills on its own.

### 7.3.3 Supply, the Reach and the Town

The diversion takes the month's scheduled supply $D_s$ (6 to 28 m$^3$/s, 462 hm$^3$ a year) from the dam's outflow above the environmental flow, and the river below it is routed by Muskingum's storage law to the town, as shown in Equation (7.5).

$$\begin{aligned} D &= \min\left(D_s,\ \max(0,\ R + Q_c - Q_{env})\right) \\ I_r &= R + Q_c - D \\ W &= K\,[X I_r + (1 - X)\,O], \qquad \frac{\mathrm{d}W}{\mathrm{d}t} = I_r - O \\ Q_{town} &= O + L \end{aligned} \tag{7.5}$$

where $K = 1\,\mathrm{d}$ and $X = 0.2$. Solved for $O$, the law gives $\mathrm{d}W/\mathrm{d}t = (I_r - W/K)/(1-X)$ and $O = (W/K - XI_r)/(1-X)$, integrated at the physics step. A sudden rise in $I_r$ makes $O$ dip by $X/(1-X) =$

0.25 of the step before it rises, a property of the law itself, which the ramped release keeps small. The daily form of the same law has coefficients 0.231, 0.538 and 0.231, which are all positive because $2KX = 0.4 \le \Delta t = 1 \le 2K(1-X) = 1.6$, the range of the step discussed for the method by Cunge [17].

### 7.3.4 Weather and the Forecast Service

A case's rain is drawn when it loads. A colored-noise index $x$ with a two-day correlation becomes a day's rain on the season's mean, and the local catchment uses 0.9 of the same index plus its own noise, as shown in Equation (7.6).

$$P(d) = \bar{P}(d)\, w\, \frac{\max(0,\, x_d - \theta)^2}{\kappa(\theta)} \tag{7.6}$$

where $\bar{P}$ is interpolated from monthly means of 70, 130, 170, 180, 150, 120, 70, 35, 12, 4, 4 and 20 mm (October first, 965 mm a year), $w$ is a case's wetness, $\theta = 0.3$, and $\kappa(\theta) = 0.30206$ is the mean of $\max(0, x-\theta)^2$ for a unit normal, so the long-run mean is $\bar{P}w$. Storms are declared by each case as three days carrying 30, 45 and 25 % of their total, on the local catchment at 0.9 of that unless the case says otherwise. Over 2000 generated years the mean is 966 mm, and seed 365, which every case uses, gives 919 mm with no natural inflow above 249 $\mathrm{m}^3/\mathrm{s}$.

Each morning the forecast service issues $M$ members for the next ten days. Member $m$ scales the true rain of lead day $\ell$ by a mean-one lognormal factor whose spread grows with the lead, and runs the declared runoff model from the catchments' true state that morning, as shown in Equation (7.7).

$$\begin{aligned} \hat{P}_{m,\ell} &= P(d+\ell-1)\, \exp\!\left(\sigma_\ell z_{m,\ell} - \sigma_\ell^2/2\right) \\ \sigma_\ell &= \sigma_{10}\sqrt{\ell/10} \\ z_{m,\ell} &= 0.7\, z_{m,\ell-1} + \sqrt{0.51}\, \varepsilon_{m,\ell} \end{aligned} \tag{7.7}$$

where $\varepsilon$ is a unit normal deviate and $\sigma_{10} = 0.6$. The members' daily inflows $\hat{I}_{m,d}$ and local flows $\hat{L}_{m,d}$ are the ensemble. A forecast bust is a declared storm added to every member, scaled by 0.15, 0.4, 0.7, 0.9, 1.1, 1.35 and 1.7 for the seven members, on every morning before its reveal day, and absent from the truth. Over 355 mornings of seed 365 the relative mean error of the ensemble mean is 0.0000, 0.0037 and 0.0002 at leads of 1, 5 and 10 days, and the members' relative spread grows from 0.002 to 0.060 and 0.151.

### 7.3.5 What the Operator Measures

The deciders read one packet each morning: the storage exactly, the inflow and town gauges with a colored error of standard deviation 2 % and correlation time 0.5 d, the dam's own exact records of its release, outflow, diversion and river flow, the gates' status, the public supply schedule and the ensemble. The rule estimates the tributary as the town gauge less the output of its own Muskingum model run on its own record of $I_r$. The rain, the true inflow ahead and the tributary are never in the packet, and a check in the application's own test suite rejects any version of the controller, the dynamic program or the tree that reads anything else. The exported file writes each gauge beside the true flow it reads, under a `truth_` prefix that keeps a replay from using the true values.

### 7.3.6 Constants, Starting State and Integration

Table 7.2 lists the constants. Every value is a declared teaching value chosen so that the basin is consistent, not one measured on a real river. The channel limit stands for a reach's rated capacity from a hydraulic study, the levee's failure flow for a levee's design flow, the environmental flow for the instream requirement of a dam's license or permit, the dead pool for an intake's invert, and the ramp limit for a permit's ramping rate, and the application's model dialog says so. Each case starts at its declared day of the water year, storage and release, with the catchment stores run through a

**Table 7.2** Constants. B and A Mark Controls in the Beginner and Advanced Panels, F Marks Fixed Plant Constants

| Symbol | Meaning | Value | Unit | |
|---|---|---|---|---|
| $A, A_L$ | catchments, above the dam and local | 2500, 600 | $km^2$ | F |
| $\varphi_q, \varphi_g$ | quick and groundwater shares | 0.40, 0.25 | — | F |
| $k_N, k_L$ | cascade constants, above and local | 1.5, 1.0 | d | A |
| $k_g$ | groundwater recession | 40 | d | F |
| $S_{dead}, S_{sill}$ | intake, gate sill | 120, 400 | $hm^3$ | F |
| $S_{crest}$ | uncontrolled crest | 1100 | $hm^3$ | F |
| $Q_{t,max}, \eta$ | powerhouse, efficiency | 180, 0.88 | $m^3/s$, — | F |
| $Q_{s,full}$ | gated spillway at the crest | 1600 | $m^3/s$ | F |
| $C_w$ | crest weir | 255 | $m^{1.5}/s$ | F |
| $Q_{flood}$ | channel limit at the town | 450 | $m^3/s$ | B |
| $Q_{levee}$ | levee failure flow | 650 | $m^3/s$ | A |
| $Q_{env}$ | environmental flow | 6 | $m^3/s$ | B |
| $K, X$ | Muskingum reach | 1.0, 0.2 | d, — | A |
| $FP$ | winter flood pool of the rule | 300 | $hm^3$ | B |
| $T_{ret}$ | return time of the rule | 10 | d | B |
| $r$ | ramp limit on the release | 150 | $m^3/s$ per d | B |
| $\delta$ | planning margin below the limit | 7 | % | A |
| $L_f$ | pre-release lead of the rule | 0 | d | B |
| $M, \sigma_{10}$ | members, rain error at 10 days | 7, 0.6 | — | A |
| $b$ | tree trunk | 2 | d | A |
| $k$ | members held | 6 | — | B |
| $c_F$ | flood damage over the limit | $5 \times 10^4$ | $/($m^3$/s d) | A |
| $c_S, c_E$ | supply, environmental shortfall | 0.10, 0.20 | $/$m^3$ | A |
| $p_t$ | energy value by month | 34–56 | $/MWh | A |
| — | SDP grid | 40×365×5 | — | A |
| — | gauge error, correlation time | 2, 0.5 | %, d | F |

year of the previous year's weather and the reach in steady flow at the starting release. The engine of Section 1.6.1 integrates by fourth-order Runge-Kutta on a step of 0.1 d, the decision is made every 1 d, and a row is logged every 0.2 d. Rain is constant within a day, so the derivatives read the day from the start of the step and stay pure.

### 7.3.7 Assumptions

The runoff model is linear, so a fixed share of rain runs off whatever the soil's wetness, and the zero mean error of the forecast relies on that linearity. Rain is uniform over each catchment and constant within a day. The lake is one well-mixed storage with no wind setup, and the reach is one Muskingum element with no backwater or floodplain storage. The forecast service knows the catchments' true state and errs only in the rain, while real ensembles also err in timing and in the initial state. The levee's failure latches a stop, and the breach is not routed. The water stage drawn at the town is a schematic stage, linear in the true flow between the channel's bank-full line and the levee's crest, not a rating curve. The powerhouse is a static map from flow and head to power: the water column, the turbine and its governor are left out, and Volume 1, Chapter 7 treats them.

## 7.4 Why the Pool Must Be Emptied Before the Storm

Three facts make the problem one of timing. The first is a rate. With the planning margin the dam may send $(1-\delta)Q_{flood} - L + D_s$ toward the town, and at about 400 $m^3/s$ the channel takes 34.6 $hm^3$ a day, so emptying a 300 $hm^3$ flood pool with no inflow at all takes 8.7 days. The second is an order of arrival. The local catchment's unit hydrograph peaks a day before the reservoir's inflow, and a release reaches the town a day after it is made because the reach's centroid lag equals $K$. By the time the tributary is high the channel has no room left for the dam. The third is the forecast. A storm seen ten days ahead is seen through a rain error of $\sigma_{10} = 0.6$ on its logarithm, which in case 7's setting of $\sigma_{10} = 0.9$ puts the wettest member's ten-day rain at 2.01 times the median member's.

The cases make each fact visible. On run of river in case 1 the lake rises only 31 $hm^3$ through a 300 mm storm and the town receives the dam's outflow plus the tributary, so the levee fails on day 16.8 at 660 $m^3/s$, while the rule curve holds the same storm with 137 $hm^3$ of its pool and a town peak of 444 $m^3/s$. In case 3 the rule is told the forecast, counting the storm's volume above the channel as if it were already in the lake. With two days of lead the town peaks at 552 $m^3/s$, with three at 525, with four at 489, and with five at 425 $m^3/s$, under the limit, so the second storm of a wet December needs five days of warning. A dam can be operated well only by acting on water that has not yet arrived. Chapter 5 meets the same need on a canal, where each gate releases a published order ahead of the demand by the travel time of the water wave.

> **TRY IT IN THE APP** **1 · Run of River**
>
> Load the scenario and press Run. Watch the town's river against the channel's limit on the top trend: on run of river the levee fails on day 16.8 at 660 $m^3/s$, and the run pauses there. Then choose Rule curve as the Strategy (Auto), press Reset and run again. The town now peaks at 444 $m^3/s$ (Table 7.4).

## 7.5 Deciding the Morning's Release

The decision is made once a day, at $t = 0, 1, 2, \ldots$ days, at the start of the physics step, and handed to the plant as a ramp. The rules and the dynamic program move the release toward their target at the full ramp rate $r$, and the tree's plan is a straight line to its first boundary value, which its own ramp rows keep within $r$. In Manual the release station sets the target.

### 7.5.1 The Rule Curve

The rule curve's target $T(d)$ is 1000 $hm^3$ from May to September, falls linearly to $S_{crest} - FP$ by 15 November, holds there to the end of February and refills through March and April. The rule returns the lake to the curve over $T_{ret}$ days and switches to the channel's limit whenever the lake is more than 5 $hm^3$ into its flood pool, as shown in Equation (7.8).

$$\begin{aligned} R &= R_{ch} \quad \text{if } S + V_f - T(d) > 5\,\text{hm}^3 \\ R &= \min\left(I_m + \frac{S - T(d)}{0.0864\,T_{ret}},\ R_{ch}\right) \quad \text{otherwise} \\ R_{ch} &= (1-\delta)\,Q_{flood} - \hat{L} + D_s \end{aligned} \tag{7.8}$$

where $I_m$ is the inflow gauge and $\hat{L}$ the tributary estimate. While the lake is more than 30 $hm^3$ above the dead pool, $R$ is held at least at $D_s + Q_{env}$ unless $R_{ch}$ is lower. Without a forecast $V_f = 0$. With a pre-release lead $L_f$, $V_f = \sum_{\ell=1}^{L_f} 0.0864 \max_m \left(\hat{I}_{m,\ell} - R_{ch,m,\ell}\right)^+$ counts, day by day, the largest excess among the members of the forecast inflow over what the channel lets the dam release with that member's tributary (or the median excess, to hedge), and the rule's tributary is then the largest (or median) forecast local flow over the next three days. A margin of 5 % let the rule touch 453 $m^3/s$ while emptying case 1's pool, which is why the margin is 7 %.

### 7.5.2 Stochastic Dynamic Programming

The dynamic program turns the year into a value function on the storage, the day and the inflow class. Each day the inflow falls in one of five classes of its half-month, split at the 35th, 65th, 85th and 96th percentiles so that the top class holds the storm days, and the next day's class follows a counted transition matrix $P_{s(t)}(i,j)$. The recursion is the stochastic form of Equation (1.11), run backward over 365 daily stages, as shown in Equation (7.9).

$$\begin{aligned} V_t(S,i) &= \max_{R \le R_{cap}(S)} \left[ r_t + \textstyle\sum_j P_{s(t)}(i,j)\, V_{t+1}(S',j) \right] \\ S' &= S + 0.0864\,(\bar{q}_{s,i} - R) - E(S,t) \end{aligned} \tag{7.9}$$

where $r_t = r_t(S,i,R)$ is the day's reward, $\bar{q}_{s,i}$ the class's mean inflow, a release the storage cannot supply is cut at the dead pool, and storage above the crest leaves over it within the day. The reward is the day's energy at the month's price less the costs of a shortfall, as shown in Equation (7.10).

$$\begin{aligned} r_t = {} & 24\,p_t P_e\big(\min(R, Q_{t,max}),\ \tfrac{S+S'}{2}\big) \\ & - 0.0864\left[c_S(D_s - D) + c_E(Q_{env} - I_r)^+\right] \\ & - c_F(Q_{town} - Q_{flood})^+ - 40\,c_F(Q_{town} - Q_{levee})^+ \end{aligned} \tag{7.10}$$

where $Q_{town} = I_r + \bar{l}_{s,i}$ uses the class's mean tributary flow without routing, and reaching the dead pool costs \$2 M a day. Storage is discretized to 40 levels from 120 to 1100 $\mathrm{hm}^3$, interpolated linearly, and the release to a grid of 2 to 50 $\mathrm{m}^3/\mathrm{s}$ steps cut at the outlets' capacity. The statistics are counted from a declared record of 100 synthetic years of the same weather generator. Its storm days are one-day bursts drawn at 1/120 a day from November to April, each 30 % of a lognormal 230 mm total, so the record holds 1.66 bursts a year of 71 mm on average and a mean inflow of 59.3 $\mathrm{m}^3/\mathrm{s}$. The year is repeated, $V_0$ of one pass becoming $V_{365}$ of the next, until the yearly increment is the same at every state to within \$1. It converges in four years, with spans of $5.3 \times 10^7$, $2.2 \times 10^3$, 3.2 and $4.9 \times 10^{-3}$ dollars and a gain of \$15.73 M a year. The policy's release $R^*(S,t,i)$ is read each morning at the storage, the day and the class of the ensemble mean for that day. The water value $w(t,S,i) = \sum_j P(i,j)\,\partial V_t(S,j)/\partial S$ is the price the tree puts on stored water, the slope of the value function that Section 1.4.9 reads as a costate.

### 7.5.3 The Scenario Tree

Each morning the tree solves one linear program over $N = 10$ days and the $M$ members. Its decisions are the release at each day boundary, $B_{m,j}$, with the release ramping linearly between boundaries so that the day's mean release is $(B_{m,d} + B_{m,d+1})/2$, and the turbine flows $Q_{t,m,d}$. For the first $b = 2$ days the decisions are one shared variable for every member, because they are taken before the members can be told apart (the non-anticipativity constraint of Equation (1.13)), and from day 3 each member has its own branch, as shown in Equation (7.11).

$$\begin{aligned} \max\ & \textstyle\sum_{d<b} p_d k_H Q_{t,d} + \frac{1}{M}\sum_m J_m \\ & J_m = \textstyle\sum_{d\ge b} p_d k_H Q_{t,m,d} + w_T S_{m,N} \\ & \qquad - c_m \textstyle\sum_j x_{m,j} - c_\sigma \sum_d \sigma_{m,d} \\ & \text{s.t. } B_{m,j} = B_j\ (j \le b), \qquad |B_{m,j} - B_{m,j-1}| \le r \\ & S_{m,j} = S_0 + \textstyle\sum_{d<j}\left[0.0864\,(\hat{I}_{m,d} - \bar{B}_{m,d}) - E_d - \sigma_{m,d}\right] \\ & \qquad S_{dead} + 5 \le S_{m,j} \le S_{crest} \\ & \qquad D_s + Q_{env} \le B_{m,j} \le Q_{t,max} + a\,(S_{m,j} - S_{sill}) \\ & \qquad 0 \le Q_{t,m,d} \le \min\left(Q_{t,max},\ \bar{B}_{m,d}\right) \\ & \qquad O_{m,j} + \max(\hat{L}_{m,j-1}, \hat{L}_{m,j}) \le (1-\delta)\,Q_{flood} + x_{m,j} \end{aligned} \tag{7.11}$$

where $\bar{B}_{m,d} = (B_{m,d} + B_{m,d+1})/2$ is the day's mean release, $k_H$ the energy of 1 $\mathrm{m}^3/\mathrm{s}$ for a day at today's head, $p_d$ the price, and $O_{m,j}$ the reach outflow at boundary $j$, an affine function of the boundary releases because the Muskingum law of Equation (7.5) is integrated exactly for a release that ramps linearly through each day. The rating row applies where the member's storage cannot fall below the sill, and $B_{m,j} \le Q_{t,max}$ applies elsewhere. The crest variable $\sigma_{m,d}$ exists only on the days the ramp lets that member's lake reach the crest. It costs $c_\sigma = 2c_F/0.0864 = \$1.16 \times 10^6$ per $\mathrm{hm}^3$, twice the damage the same water would do over the town's limit, and it is never credited with a negative water value, so the plan spills only where the gates at their rating and ramp cannot hold the lake. The terminal price $w_T$, the terminal value of Section 1.4.8, is the dynamic program's water

value ten days ahead at today's storage and class, one linear price for every member's end storage. The tree is solved by the bounded-variable simplex method of the apps' shared linear programming module, after a presolve that drops rows the ramp makes unreachable. On the 365 mornings of the challenge with the suggested settings the program has 193 to 291 rows and 126 to 224 columns.

🕮 **Pause.** Try Exercise 7.2 before reading on.

### 7.5.4 The Chance Constraint and the Fallbacks

The chance-constrained tree holds the channel on $k$ of the $M$ members, the chance constraint of Section 1.4.12. Before the solve, the $M-k$ members with the largest forecast volume reaching the town are let go, and their excess $x_{m,j}$, a slack priced linearly as in Section 1.4.6, carries a token cost $c_m = 0.02\,c_F$ so that their branches stay sensible, while every other member has no $x_{m,j}$ at all. This is the scenario-discarding approach of Campi and Garatti [10] applied to seven members, which is far too few for its probabilistic guarantee, so $k/M$ is a declared fraction of the ensemble and not a measured probability. With $k = M$ the program is the scenario tree in which every member is held. When the members to be held cannot all be held, the program is solved again with every member soft at the full damage cost $c_m = c_F$, which minimizes the expected damage while the energy and a positive water value still count, and a negative water value is taken as zero for any soft member so that it cannot buy flooding. If that fails the storage bounds and the supply floor are relaxed, and if no solve succeeds the rule curve decides that day. The Forecast view labels every morning on which no plan held the required members.

🕮 **Pause.** Try Exercise 7.3 before reading on.

### 7.5.5 A Perfect-Foresight Bound and Three Models

The challenge measures energy against the most any operator could make from the year's realized inflow, as shown in Equation (7.12).

$$\begin{aligned} E_{PF}(S_f) &= \max_{Q_t} \sum_t 24\,P_e\big(Q_{t,t},\ \tfrac{S_t+S_{t+1}}{2}\big) \\ \text{s.t. } S_{t+1} &= \min\big(S_{crest},\ S_t + V_{in,t} - E_t - 0.0864\,Q_{t,t}\big) \\ 0 &\le Q_{t,t} \le Q_{t,max}, \qquad S_N \ge S_f \end{aligned} \tag{7.12}$$

where $V_{in,t}$ is the day's recorded inflow volume and $S_f$ is the run's own end storage. It relaxes the channel, the ramp and the supply, so it bounds every policy that ends at or above $S_f$. A forward dynamic program solves it on 1 and 0.5 $\mathrm{hm}^3$ storage grids, and the two answers are extrapolated to zero spacing. It is a teaching bound, never a controller. Three models are kept apart. The plant is everything in Section 7.3. The dynamic program's model is a daily Markov chain of class means with no routing and no ramp. The tree's model is the plant's own storage, rating, ramp and Muskingum laws driven by the forecast members, with energy valued at today's head.

## 7.6 Measurements on Ten Cases and a Water Year

The cases appear in Table 7.3, and the measured comparisons in Table 7.4. All values below come from headless runs of the application's own engine on each case's declared weather (seed 365), advanced through the same step sequence the browser uses, and a setting that a row changes is applied before Run, which primes the run exactly as Reset would (Section 1.6.4). A day over the limit is a day on which the true town flow exceeded 450 $\mathrm{m}^3/\mathrm{s}$ at any physics step, the pool used is the largest storage above the rule curve, and gate volumes are integrated over time.

### 7.6.1 Routing, Lead and a Forecast That Busts

Case 2's two storms show the order of arrival at work. The first storm leaves 92 $\mathrm{hm}^3$ in the pool when the second begins on day 16, the town exceeds the limit from day 17.4 and peaks at 575 $\mathrm{m}^3/\mathrm{s}$ on day

**Table 7.3** The Ten Cases

| # | Strategy, start and weather | Teaching point |
|---|---|---|
| 1 | Run of river, 1 Dec, 300 mm on day 12 | storage is flood control |
| 2 | Rule, wet December, 260 and 360 mm ten days apart | the tributary wins |
| 3 | Rule + forecast, lead 2 d, case 2's weather | the lead a channel needs |
| 4 | Rule + forecast, lead 5 d, 450 mm bust for day 10 | what chasing costs |
| 5 | SDP policy, 1 Nov, 380 mm on day 40 | a season's table |
| 6 | Scenario tree, 420 mm on day 12 | trunk and branches |
| 7 | Scenario tree, mid March, $\sigma_{10} = 0.9$ | letting a member go |
| 8 | Rule, ramp 40 $m^3/s$ a day, cloudburst on the tributary | permit against flood rule |
| 9 | Rule, $FP = 100\,hm^3$, 360 mm on day 1 | an infeasible flood |
| 10 | Challenge: rule, 1 Oct, storms 20 Dec and 13 Feb, bust for 19 Jan, dry spring | the whole year |

**Table 7.4** Measured Comparisons on the Cases' Own Windows

| Case | Run | Town ($m^3/s$) | Days over | Pool ($hm^3$) | Gates ($hm^3$) | Energy (GWh) |
|---|---|---|---|---|---|---|
| 1 | run of river[a] | 732 | 7 | 31 | 209 | 57.2 |
| | rule curve | 444 | 0 | 137 | 196 | 63.9 |
| 2 | rule, $FP$ 300 | 575 | 4 | 304 | 465 | 130.2 |
| | rule, $FP$ 400 | 575 | 4 | 273 | 515 | 131.7 |
| 3 | lead 2 d | 552 | 4 | 322 | 430 | 130.0 |
| | lead 5 d | 425 | 0 | 307 | 444 | 133.5 |
| 4 | wettest member | 314 | 0 | 2 | 23 | 23.4 |
| | median member | 185 | 0 | 2 | 2 | 24.8 |
| | tree, all held | 429 | 0 | 0 | 66 | 46.0 |
| | tree, $k = 6$ | 279 | 0 | 0 | 13 | 55.1 |
| 5 | SDP policy | 530 | 4 | 38 | 171 | 224.4 |
| | rule curve | 452 | 2 | 310 | 368 | 164.9 |
| 6 | scenario tree | 434 | 0 | 115 | 139 | 123.7 |
| | rule curve | 442 | 0 | 285 | 301 | 93.6 |
| 7 | tree, all held | 412 | 0 | 173 | 64 | 67.6 |
| | tree, $k = 6$ | 340 | 0 | 215 | 24 | 78.0 |
| 8 | rule, ramp 40 | 642 | 5 | 157 | 220 | 64.5 |
| | rule, ramp 150 | 565 | 3 | 157 | 219 | 60.2 |
| | tree, ramp 40 | 429 | 0 | 249 | 105 | 56.7 |
| | tree, $Q_{env}$ 60 | 459 | 1 | 242 | 87 | 58.0 |
| 9 | rule curve[a] | 752 | 7 | 139 | 156 | 65.9 |
| | manual 300 $m^3/s$ | 560 | 7 | 120 | 258 | 89.9 |
| | tree, perfect forecast[a] | 651 | 8 | 105 | 269 | 85.9 |

[a]The levee failed. The browser pauses there, and the measurement runs on to the end of the window.

19.3, and only then does the lake fill, reaching 1104 $hm^3$ and spilling over the crest for 1.7 days from day 24.3. With a 400 $hm^3$ pool the lake stays at 973 $hm^3$ or below and the town's peak is unchanged, as Table 7.4 shows. In case 3 the lake's highest level falls only from 1122 to 1107 $hm^3$ between a two-day and a five-day lead, so what the lead buys is not volume but timing, room made while the channel is still empty. In case 4 a 450 mm storm promised for day 10 vanishes from the forecasts on day 9. Chasing the wettest member, which put it at 765 mm, the rule sends 23 $hm^3$ through the gates and the lake stands at 755 $hm^3$ on day 10, against 786 $hm^3$ when hedging on the median member and 801 $hm^3$ with no forecast. At the case's head 1 $hm^3$ through the turbines is worth 0.21 GWh, so the chase costs water and not safety.

🕮 **Pause.** Try Exercise 7.1 before reading on.

### 7.6.2 A Season's Table and a Storm's Plan

The dynamic program draws the lake down early. In case 5 it holds the turbines at their full 180 $m^3/s$ for 55 % of the 40 days before the storm, opens no gate, and stands at 559 $hm^3$ on day 40, when the storm falls, where the rule holds 794 $hm^3$. It makes 224.4 GWh over 90 days against 164.9 GWh,

uses 38 $hm^3$ of its pool against 310 $hm^3$, and ends at 636 $hm^3$ against 797 $hm^3$, so part of its energy is water the rule kept. Its table, however, neither routes the river nor sees the ramp, and its storm release puts the town at 530 $m^3/s$ on day 44.6, over the limit on 4 days, while the rule touches 452 $m^3/s$. Across the flood season the policy releases early at and below the rule's target, a mean of 225.8 $m^3/s$ against the rule's 128.1 $m^3/s$ from 100 $hm^3$ below the target up to it, and less than the rule above the target, 276.7 $m^3/s$ against 380.0 $m^3/s$, where the rule releases at the channel's limit. The water value of Figure 7.2(b) shows why. It is positive all year at low storage and rises to 87 k$ per $hm^3$ at 305 $hm^3$ in June, when summer supply and energy depend on the water, and it turns negative in the winter near the crest, reaching −72 k$ per $hm^3$ at 1000 $hm^3$ on 1 February in the middle class. Over the 4800 grid points of half-month, storage and class it is negative at 448, never below 472 $hm^3$.

The scenario tree is the storm's plan. In case 6 it lowers the lake from 800 to 664 $hm^3$ by day 12, mostly through the turbines, uses 115 $hm^3$ of its pool, holds the town at 434 $m^3/s$ and makes 123.7 GWh in 40 days. The rule waits for the storm, uses 285 $hm^3$ of the pool, peaks the town at 442 $m^3/s$, makes 93.6 GWh and sends 301 $hm^3$ through the gates to the tree's 139 $hm^3$, and the two end within 10 $hm^3$ of each other. A morning of the challenge, 9 February, is drawn in Figure 7.2(c) and (d). Its seven members put the inflow peak between 536 and 1523 $m^3/s$, and the inflow that came peaked at 1121 $m^3/s$. The trunk commits 198 $m^3/s$ and then 162 $m^3/s$, identical in every branch. Over days 3 to 6 the held members' releases fall to the 12 $m^3/s$ floor of supply and environmental flow as their tributaries fill the channel, and five of the seven plans, four held members and the let-go one, end the horizon with the lake at the crest. The let-go wettest member releases up to 355 $m^3/s$ and plans the town at 642 $m^3/s$. The crest bound is active on those members at day 10, so the trunk's release is what keeps them under it.

### 7.6.3 The Chance Constraint and the Permit

Holding every member means planning for the wettest. In case 7, a spring storm with $\sigma_{10} = 0.9$, the tree that holds all seven sends 64 $hm^3$ through the gates and makes 67.6 GWh. Holding six, it sends 24 $hm^3$ and makes 78.0 GWh, and the town peaks at 340 $m^3/s$ against 412 $m^3/s$. Both end at 1082 $hm^3$. The all-member tree finds no plan that holds the wettest member on 6 of its 30 mornings and weighs that member's excess at the full damage cost, while the chance-constrained tree weighs it at 2 % of it, which is what frees the water. On the bust of case 4 the same contrast is 13 against 66 $hm^3$ through the gates and 55.1 against 46.0 GWh. The storm that came was no wetter than the six held members, and the measurement says nothing about the year in which it is the seventh.

> **TRY IT IN THE APP** **7 · The Chance Constraint**
>
> Load the scenario and run it on Scenario MPC, which holds every member. Then choose Chance-constrained MPC with Members held (k) at 6, press Reset and run again. Over the case's window the gates pass 24 $hm^3$ instead of 64 $hm^3$, the energy rises from 67.6 to 78.0 GWh, and the town peaks lower (Table 7.4).

Case 8 puts a permit against the flood rule. The lake starts 160 $hm^3$ into its pool with the release near the channel's limit, and a cloudburst falls mostly on the tributary, which peaks at 394 $m^3/s$. The rule at a ramp of 40 $m^3/s$ a day lets the town reach 642 $m^3/s$, 5 days over the limit, and at 150 $m^3/s$ a day still 565 $m^3/s$, because it starts the cut only when the gauge rises. The tree at 40 $m^3/s$ a day starts the cut before the rain, holds 249 $hm^3$ in the pool and keeps the town at 429 $m^3/s$, although on 9 of its 20 mornings no plan held every member. Raising the environmental flow to 60 $m^3/s$ puts a floor under that cut, and the town reaches 459 $m^3/s$ on 1 day.

### 7.6.4 A Flood No Release Holds

Case 9 is the infeasible variant. The flood pool is at its smallest, 100 $hm^3$, the lake at the top of the rule curve, and a 360 mm atmospheric river begins on day 1. The rule fills the lake to 1139 $hm^3$, spills over the crest, and the levee fails on day 7.1 with the town peaking at 752 $m^3/s$. Constant releases

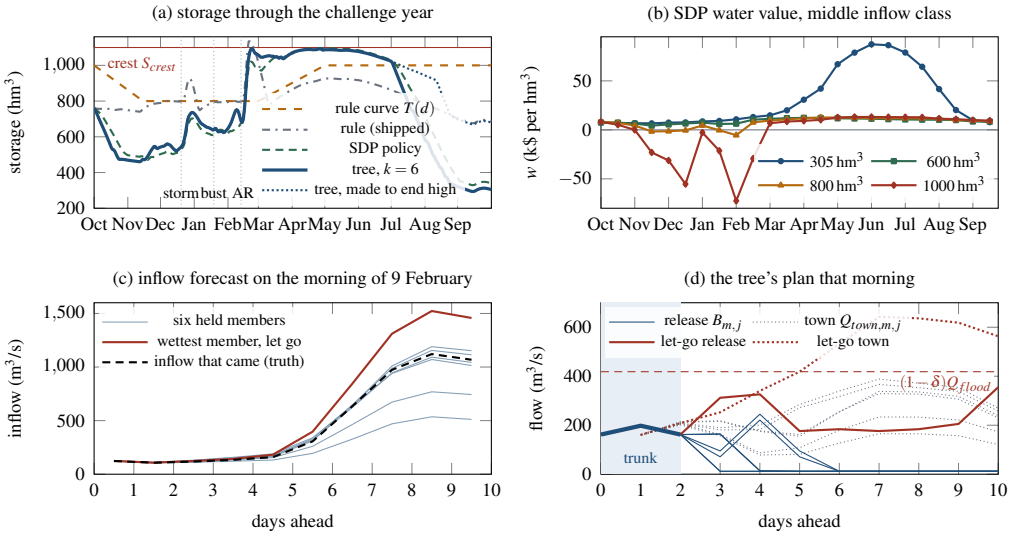



**Figure 7.2** Three deciders measured on the challenge. (a) Storage through the water year under the shipped rule curve (dash-dotted), the dynamic program's policy (densely dashed), the suggested chance-constrained tree ($k = 6$, heavy solid) and the same tree made to end the year high by raising its terminal water value by 2000 \$ per hm$^3$ from 1 July (dotted), with the rule curve's target dashed. Vertical dotted lines mark the 20 December storm, the bust promised for 19 January and the 13 February atmospheric river (AR). (b) The dynamic program's water value $w$ at four storages on the first day of each half-month, in the middle inflow class, one marker shape per storage. (c) The ensemble issued on the morning of 9 February (six held members thin, the wettest heavy) and the daily inflow that then came (dashed, teaching truth, never read by the tree). (d) The tree's plan that morning: boundary releases $B_{m,j}$ (solid), with the two-day trunk shared (heaviest line), and each member's town flow (dotted) against the planning limit of 418.5 m$^3$/s. Six members are held, and the wettest is let go (heavy lines).

held from the storm's first day give peaks of 744, 643, 560, 660 and 759 m$^3$/s for 100, 200, 300, 400 and 500 m$^3$/s. The scenario tree peaks at 663 m$^3$/s, and on a perfect forecast, a teaching bound, at 651 m$^3$/s. A wider search through the plant does no better: 245 two-step schedules reach 522 m$^3$/s at best and a coordinate descent on the first 12 daily releases 503 m$^3$/s. This is a search, not a proof of infeasibility. Linear programs with perfect foresight that check the town only at day ends or let the crest overflow instantly do reach 450 m$^3$/s, but their schedules peak at 554 to 671 m$^3$/s when run through the plant.

### 7.6.5 The Water Year

The challenge runs from 1 October with a 280 mm storm on 20 December, a 300 mm storm forecast for 19 January that never falls, a 440 mm atmospheric river on 13 February and rain at 35 % of its mean for 92 days from 1 March. Its target is no day over the 450 m$^3$/s limit, supply of at least 95 % of the schedule, and energy of at least 80 % of the perfect-foresight energy for the same year ending at the run's own storage. The verdict holds only for the case's own weather and water: moving the rain slider, pressing an event button or using the helper that sets the lake on the rule curve voids it, the last because it moves water outside the balance. Table 7.5 lists eight configurations, and the storage paths of four appear in Figure 7.2(a). The shipped rule curve fails in February, as the opening paragraph describes, with 400.4 GWh, 75.6 % of its bound. The dynamic program makes 89.5 % but puts the town over the limit on 9 days. Both trees pass, and the suggested chance-constrained tree, holding six members, makes 528.9 GWh, 89.5 % of 590.8 GWh, with a town peak of 425 m$^3$/s and

**Table 7.5** The Challenge Water Year at Seed 365. Share of the Perfect-Foresight Energy at the Run's Own End Storage

| Configuration | Energy (GWh) | Share (%) | End ($hm^3$) | Days over | Town ($m^3/s$) | Seeds passed |
|---|---|---|---|---|---|---|
| rule curve (shipped)[a] | 400.4 | 75.6 | 687 | 8 | 710 | 0/7 |
| rule, lead 5 d, wettest[a] | 399.1 | 75.4 | 687 | 5 | 651 | 0/7 |
| SDP policy | 524.6 | 89.5 | 341 | 9 | 600 | 1/7 |
| tree, all seven held | 528.8 | 89.5 | 305 | 0 | 428 | 7/7 |
| tree, $k = 6$ (suggested) | 528.9 | 89.5 | 305 | 0 | 425 | 7/7 |
| tree, $k = 5$ | 532.8 | 90.4 | 318 | 4 | 525 | 6/7 |
| tree, $k = 6$, made to end high | 469.7 | 88.5 | 681 | 0 | 425 | — |
| run of river[a,b] | 377.4 | 76.3 | 861 | 16 | 962 | 0/7 |

[a]The levee failed. [b]Supply 45.0 %, the others 100.0 %. All start at 760 $hm^3$.

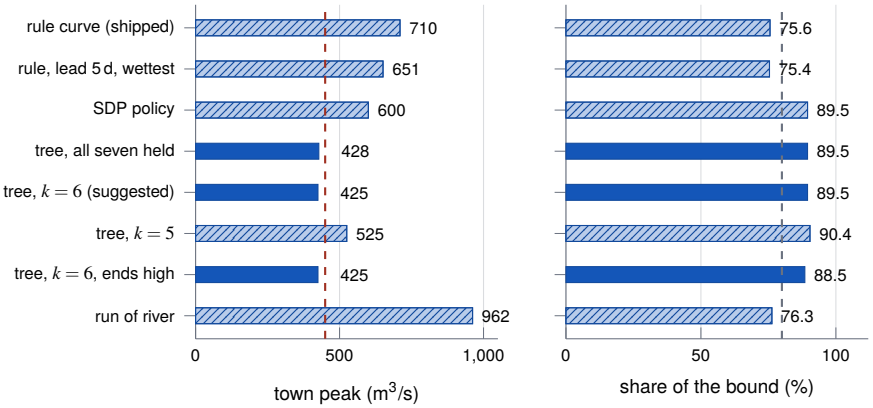



**Figure 7.3** The challenge water year redrawn from Table 7.5: the town's peak flow against the 450 $m^3/s$ channel limit (red dashed line, left) and the share of the perfect-foresight energy against the 80 % target (gray dashed line, right). Dark bars pass the challenge, hatched bars fail it. Notice that the shares of the dynamic program, the trees and the tree that holds five members lie within two points of each other, while only the trees that hold six or seven members keep the town under its limit.

the supply in full. Holding five members puts the town over the limit on 4 days. Over seven weather seeds the suggested configuration passes 7 times with 84.7 to 91.8 %, the rule passes none, and the dynamic program passes once. On 9 of the 365 mornings, between 6 and 15 February, no plan held six members and the tree minimized the expected damage instead, and on every such morning of the shipped cases the plan's expected excess is within 2 % of the least any plan reaches.

Both trees end the year at 305 $hm^3$ from a start of 760 $hm^3$, while the rule ends at 687 $hm^3$, which raises the question of whether the pass was bought by emptying the lake. It was not. Made to end the year at 681 $hm^3$, by raising its terminal water value by 2000 $ per $hm^3$ from 1 July, the same tree makes 469.7 GWh, 88.5 % of the 530.8 GWh its own end storage allows, and still passes with the town at 425 $m^3/s$. The drawdown is worth one point of the share. The comparison at the run's own end storage was kept because a common floor is not a bound. Perfect foresight from 760 $hm^3$ gives 515.5 GWh, which the suggested run exceeds, at 102.6 %. The trees spend the lake in the late summer, when energy is dearest, because the water value says the winter refills it, and run on into a second year they refill it to 1096 $hm^3$ with the supply in full and make 398.1 GWh against the rule's 352.5 GWh. Figure 7.3 draws the town peaks and energy shares of Table 7.5 side by side.

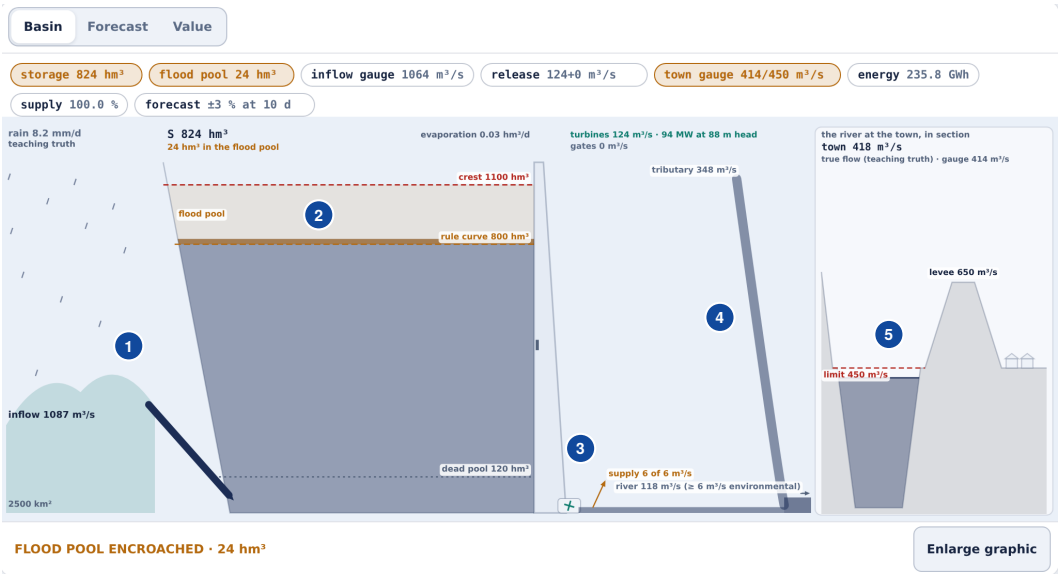



**Figure 7.4** The challenge on the suggested settings (chance-constrained tree, $k = 6$) at day 139.0, 17 February, the state of Table 7.1, cropped to the Basin view. (1) The atmospheric river brings 1087 $m^3/s$ into the lake (the gauge reads 1064). (2) The lake stands at 824 $hm^3$, 24 $hm^3$ into its flood pool above the 800 $hm^3$ rule curve, so the status strip reads FLOOD POOL ENCROACHED. (3) The dam releases 124 $m^3/s$ through the turbines (94 MW at an 88 m head) with the gates shut, and 118 $m^3/s$ stays in the river below the 6 $m^3/s$ diversion. (4) The tributary brings 348 $m^3/s$. (5) The town's true flow is 418 $m^3/s$ (gauge 414) against the 450 $m^3/s$ limit and the 650 $m^3/s$ levee. The chips read 235.8 GWh and the supply in full, and the year ends with no day over the limit and a town peak of 425 $m^3/s$ (Table 7.5).

**+ DID YOU KNOW? One Member More Let Go Buys Little**

From the simulator: over the challenge year the tree that holds five of the seven members makes 532.8 GWh against 528.9 GWh for six, but puts the town over the limit on 4 days with a peak of 525 $m^3/s$ and passes 6 of 7 seeds, against 0 days, 425 $m^3/s$ and 7 of 7 (Table 7.5).

The state of the challenge on 17 February, four days after the atmospheric river began, is shown in Figure 7.4.

### 7.6.6 Code Verification and Agreement

Code verification asks whether the program solves its own equations. At every 0.1 d step of the challenge year the storage equals its start plus inflow less outflow and evaporation to $1.3 \times 10^{-10}$, $3.4 \times 10^{-11}$ and $6.3 \times 10^{-11}$ $hm^3$ under the rule, the dynamic program and the suggested tree, and the outflow equals the diversion plus the river and the reach's inflow less outflow its storage change to $2.6 \times 10^{-10}$ $hm^3$. Halving the step changes the challenge year's energy by $-1.3 \times 10^{-5}$ under the rule and $2.8 \times 10^{-4}$ under the suggested tree, and leaves every verdict and the energy ranking of the seven strategies unchanged. The perfect-foresight extrapolation from 1 and 0.5 $hm^3$ agrees with the one from 0.5 and 0.25 $hm^3$ to $2.9 \times 10^{-4}$. A Bellman backup recomputed with independent reward and transition code at all 73 000 states equals the stored value function to the dollar with the same decision at every state, and one more year of recursion changes no decision. Simulated for 2900 years on its own Markov model, the policy averages \$15.97 M a year against the reported gain

of \$15.73 M, a difference of 1.5 % from the storage interpolation. An independent brute force on a three-day, three-member tree with a binding rating and a let-go member equals the linear program to 0.000 on a 0.05 $m^3/s$ grid, and in closed loop every plan stays inside the ramp to $8 \times 10^{-12}$ $m^3/s$. With $k$ of 5 and 4 the plan exceeds the channel on exactly $M - k$ members. Runs logged at 1 and 30 days a second are identical.

Agreement with the published hydrology is claimed separately, and more narrowly. The plant's cascade, driven by a unit impulse, matches Nash's Equation (7.2) at 1.5, 3 and 6 d to $2.0 \times 10^{-6}$, $3.8 \times 10^{-7}$ and $-1.7 \times 10^{-7}$ relative, with its peak at 3.00 d, unit volume, a first moment of 4.5000 d and a second of 27.000 $d^2$, the moments of $nk_N$ and $n(n+1)k_N^2$ that Nash gives. The reach's response to a unit step from rest matches its analytic solution, $O(t) = 1 - e^{-t/K(1-X)}/(1-X)$, including the initial dip of $-0.250$, at 0.1, 0.5 and 2.0 d to $10^{-6}$, and on a 450 $m^3/s$ hydrograph the volume closes to $6.4 \times 10^{-14}$ $m^3/s$ d with a centroid lag of 1.000 00 d, equal to $K$ as Muskingum's law requires. Because no river or dam data entered the model, these checks validate it against no real basin. The value function takes 735 to 746 ms to solve on first load in a desktop browser, a tree solve averages 4.8 to 4.9 ms, and the water year plays in 12.5 to 12.6 s at 30 days a second.

## 7.7 Technology in Practice

### What Measures It

A flood-control decision reads the lake, the rivers and the forecast. Lake stage is the index from which reservoir contents are computed, and a lake gauge is held to the same accuracy as a river gauge. Satellite telemetry is now the standard link from a gauging station to the office: it delivers the stage in near real time and alerts the hydrographer to extreme or abnormal events and to instrument malfunctions [18]. The inflow gauge and the town gauge of Figure 7.1 stand for two such stations, and the exactly measured storage $S$ for a lake gauge read through the storage curve.

> **TECHNOLOGY** **Sensor: The River Gauge**
>
> A river gauge measures stage, the height of the water surface, with a float in a stilling well, a gas-purge bubbler read by a pressure transducer, a submerged pressure transducer, or a radar that times a pulse reflected from the water. The flow follows from the stage through the site's stage-discharge relation. A whole station must be accurate to 0.01 foot or 0.2 percent of the effective stage, whichever is greater, and its recorder stores a value every 5 or 15 minutes or every hour. A clogged intake makes the well lag the river, high velocity beside a pier can draw the reading down by 0.5 foot or more, and radar is upset by waves and floating debris [18]. The chapter's two gauges carry a colored 2 % error and none of these faults.

### What Moves It

A dam releases through its powerhouse, its outlet works and its spillway. A spillway needs gates or another movable crest when an uncontrolled crest cannot pass the required flood, and also when it must release storage below the normal water surface. Radial gates and wheel-mounted slide gates are the most common on large spillways [19]. In the chapter's plant the turbines take the release first and the gated spillway the rest, with a rating that is linear from the sill to the crest (Equation (7.4)), and the crest spills on its own.

> **TECHNOLOGY** **Final Element: The Radial Spillway Gate**
>
> A radial gate is a steel cylindrical segment carried on radial arms that turn on trunnion pins. Its face is centered on the pins, so the whole water load passes through them and the hoist lifts only part of the gate's weight. The small effort suits simple automatic control, and the gates of one spillway can open step by step as the lake rises. The choice of gate weighs ice and debris, whether an operator attends in a flood, and whether electric power is available. A human or mechanical failure to open gates raises the lake above its expected level [19]. The chapter's gate fault halves the gated capacity, $\gamma = 0.5$.

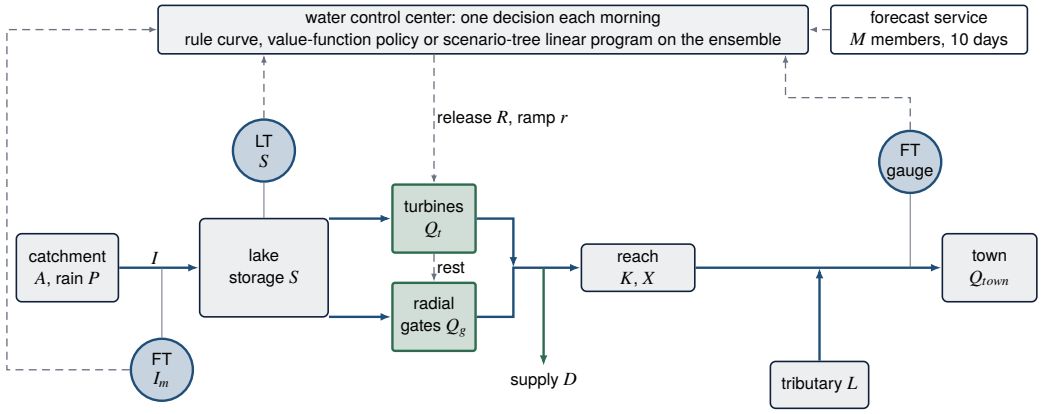



**Figure 7.5** The basin's decision loop on field hardware, drawn from Sections 7.3 and 7.5. Green boxes are the turbines and the radial spillway gates, blue circles the inflow gauge, the lake gauge and the town gauge, dashed arrows signals. Notice that the town gauge sits below the tributary, a day's travel from the dam, so only the forecast lets the morning's release act before the town's river rises.

### Where the Controller Runs

A water control plan is a well-defined set of rules and tools that guides daily decisions to release or retain water in a reservoir's flood pool. In one forecast-based study, the streamflow forecast was issued at least daily, a decision-support model recommended the flood releases, and that model's computational step was later moved from daily to hourly [13]. River gauges record every 5 or 15 minutes or every hour [18]. The chapter's deciders act once a day on the morning's ensemble, while the plant is integrated every 0.1 d, and the scenario tree is one linear program solved each morning, as Figure 7.5 draws it.

> **IN PRACTICE** **Hard Constraints Every Plan Must Meet**
>
> When one reservoir's water control plan was compared with four forecast-based alternatives, every plan had to satisfy the same hard, inviolable constraints: the limits on the release's rate of change, the maximum release schedule, the instream minimum flow, and no call for forecast updates more often than they are issued. Each plan should also avoid spillway flow as far as possible. Under the existing plan no forecast is used, and storage above the guide curve is always evacuated as quickly as feasible [13]. The chapter's ramp limit $r$, environmental flow $Q_{env}$ and one ensemble each morning stand for these rows.

### What Is Modernizing It

Three developments are changing reservoir operation. Ensemble streamflow forecasts have been operational at one river forecast center since 2012, driven by three days of forecaster precipitation followed by twelve days of a global weather ensemble. In two test years an operation added a variable buffer pool to the guide curve and let a 15-day ensemble forecast recommend the flood releases [13]. Observation is improving as well: reconnaissance fills major gaps in observations over the ocean inside atmospheric rivers, storms that can be forecast with a long lead [13], and radar stage sensors, which need no stilling well, are now in common use [18]. The chapter's chance-constrained tree belongs to the same family of forecast-based rules, on a declared basin and with seven members instead of an operational ensemble.

## 7.8 Limitations

The basin is declared, and every limit is a teaching value. The runoff is linear, rain is uniform over a catchment, the forecast errs only in the rain, and the cases run on one weather seed. At other seeds

three answers change: at seed 2 the rule in case 1 touches 469 $m^3/s$ for a day and a five-day lead in case 3 peaks at 485 $m^3/s$, 3 days over the limit, and at seed 1 the wettest and median runs of case 4 coincide.

> **OPEN CHALLENGE** **Ensembles That Run Low in a Flood**
>
> Tested on years of winter hindcasts, one reservoir's ensemble put the observed three-day inflow in its upper tail at leads of one to six days, a bias toward under-forecasting that shrinks with the lead, and its wettest forecasts at one to three days looked too narrow. The study asked for methods that produce ensembles capturing more of the meteorological and hydrological uncertainty [13]. This chapter's members err only in the rain, with no mean error. Progress would be ensembles whose spread matches their error in the floods that drive release decisions.

The dynamic program has three known mismatches with the plant. Its 100-year record holds one-day bursts of 71 mm, 1.66 a year, not the three-day atmospheric rivers of 260 to 440 mm that the cases declare, so it learns a milder winter than the plant has, and the application's model dialog and instructions say so. Trained on full three-day storms instead, with a mean record inflow of 73.4 $m^3/s$ rather than 59.3 $m^3/s$, its gain falls from \$15.73 M to −\$4.20 M a year, and the suggested tree that prices water with it draws the challenge's lake to 171 $hm^3$ by 1 November and makes 66.3 % of its bound, so that record was not adopted. A reader who compares case 5's storm with the record's bursts sees the mismatch directly. The table is daily and ignores the reach and the ramp, which is why its storm release in case 5 overshoots the channel, and it releases early at and below the rule's target, not above it.

The tree prices every member's end storage at one linear water value taken at today's storage, so on 9 February it values five members' lakes at the crest at the 10 128 \$ per $hm^3$ of a lake at 752 $hm^3$. Its let-go members carry a token 2 % of the damage cost, its fallback on a truly infeasible morning minimizes expected damage and is labeled as such, and seven members are far too few for the guarantee of the scenario-discarding method. Case 9's infeasibility rests on a search, not a proof. The band of the forecast issued five days earlier is remembered by the scene for display and never read by a decider, and a capture made without a display shows none. A simulated pass is not a flood-safety assessment, an operating permit or a certification of any decision rule.

> **OPEN CHALLENGE** **How Much Risk a Release Should Accept**
>
> One forecast-based release rule runs every ensemble member through the reservoir and limits the fraction of members whose storage exceeds a threshold, by a risk curve over the 15-day forecast. The study recommended optimizing those risk-tolerance curves by more sophisticated methods, exploring multi-objective approaches, and generating synthetic ensembles to test the rule over a wider range of hydrology [13]. This chapter's $k$ of $M$ plays the same role, a declared fraction rather than a measured probability. Progress would be a risk level chosen by a stated method and checked against long records.

## 7.9 Conclusion

A flood-control reservoir is a plant whose safety is decided before the storm, and this chapter measured that on a declared basin with Nash cascades, a Muskingum reach, a tributary below the dam and a seven-member ensemble forecast. A wet December floods the town with the lake 234 $hm^3$ below its crest, and a rule told the forecast needs five days of lead to hold the second storm. The dynamic program draws the lake to 559 $hm^3$ before a storm where the rule holds 794 $hm^3$, but its season's table puts the town at 530 $m^3/s$. The scenario tree uses 115 $hm^3$ of pool where the rule uses 285 $hm^3$ and makes 123.7 GWh against 93.6 GWh, and letting the wettest member go buys 10.4 GWh in a spring storm. Over the water year the suggested tree keeps the town at 425 $m^3/s$ and makes 89.5 % of the perfect-foresight energy, and 88.5 % when made to end the year high. The dynamic program's record is milder than the cases' storms, the tree prices stored water at one linear value, and the infeasible case

rests on a search. A fuller model would add soil moisture to the runoff, timing errors to the forecast, a ramp state to the dynamic program and a concave terminal value to the tree. The measurements show that the room a dam makes for a flood is made from a forecast, days before the flood arrives.

**KEY POINTS**

- A dam makes room for a flood only as fast as the channel below carries water, so it must act on water that has not yet arrived.
- A wet December floods the town with the lake 234 $hm^3$ below its crest, because the tributary below the dam rises first.
- Letting the wettest member go buys 10.4 GWh in a spring storm.
- Over the water year the suggested tree keeps the town at 425 $m^3/s$ and makes 89.5 % of the perfect-foresight energy.

**Where else this applies.** An irrigation canal (Chapter 5) also releases ahead of a delay: an order released ahead moves its levels about 2 cm, against up to 16 cm at its start (Table 5.6).

## References

[1] J. D. C. Little, "The use of storage water in a hydroelectric system," *Journal of the Operations Research Society of America*, vol. 3, no. 2, pp. 187–197, 1955.

[2] W. W.-G. Yeh, "Reservoir management and operations models: A state-of-the-art review," *Water Resources Research*, vol. 21, no. 12, pp. 1797–1818, 1985.

[3] J. W. Labadie, "Optimal operation of multireservoir systems: State-of-the-art review," *Journal of Water Resources Planning and Management*, vol. 130, no. 2, pp. 93–111, 2004.

[4] D. P. Loucks and E. van Beek, *Water Resource Systems Planning and Management*. Springer International Publishing, 2017.

[5] A. T. Schwarm and M. Nikolaou, "Chance-constrained model predictive control," *AIChE Journal*, vol. 45, no. 8, pp. 1743–1752, 1999.

[6] S. Lucia, T. Finkler, and S. Engell, "Multi-stage nonlinear model predictive control applied to a semi-batch polymerization reactor under uncertainty," *Journal of Process Control*, vol. 23, no. 9, pp. 1306–1319, 2013.

[7] J. R. Birge and F. Louveaux, *Introduction to Stochastic Programming*, ser. Springer Series in Operations Research and Financial Engineering. Springer New York, 2011.

[8] J. R. Stedinger, B. F. Sule, and D. P. Loucks, "Stochastic dynamic programming models for reservoir operation optimization," *Water Resources Research*, vol. 20, no. 11, pp. 1499–1505, 1984.

[9] H. L. Cloke and F. Pappenberger, "Ensemble flood forecasting: A review," *Journal of Hydrology*, vol. 375, no. 3–4, pp. 613–626, 2009.

[10] M. C. Campi and S. Garatti, "A sampling-and-discarding approach to chance-constrained optimization: Feasibility and optimality," *Journal of Optimization Theory and Applications*, vol. 148, no. 2, pp. 257–280, 2011.

[11] L. Raso, D. Schwanenberg, N. C. van de Giesen, and P. J. van Overloop, "Short-term optimal operation of water systems using ensemble forecasts," *Advances in Water Resources*, vol. 71, pp. 200–208, 2014.

[12] G. Uysal, R. Alvarado-Montero, D. Schwanenberg, and A. Şensoy, "Real-time flood control by tree-based model predictive control including forecast uncertainty: A case study reservoir in Turkey," *Water*, vol. 10, no. 3, p. 340, 2018.

[13] Lake Mendocino FIRO Steering Committee, "Lake Mendocino forecast informed reservoir operations final viability assessment," UC San Diego, Tech. Rep., Dec. 2020. [Online]. Available: https://cw3e.ucsd.edu/FIRO_docs/LakeMendocino_FIRO_FVA.pdf

[14] A. Castelletti, F. Pianosi, and R. Soncini-Sessa, "Water reservoir control under economic, social and environmental constraints," *Automatica*, vol. 44, no. 6, pp. 1595–1607, 2008.

[15] B. A. Faber and J. R. Stedinger, "Reservoir optimization using sampling SDP with ensemble streamflow prediction (ESP) forecasts," *Journal of Hydrology*, vol. 249, no. 1–4, pp. 113–133, 2001.

[16] J. E. Nash, "The form of the instantaneous unit hydrograph," in *IAHS Red Book Series, Publication 45*. International Association of Hydrological Sciences, 1958, pp. 114–121. [Online]. Available: https://iahs.info/uploads/dms/045011.pdf

[17] J. A. Cunge, "On the subject of a flood propagation computation method (Muskingum method)," *Journal of Hydraulic Research*, vol. 7, no. 2, pp. 205–230, 1969.

[18] V. B. Sauer and D. P. Turnipseed, "Stage measurement at gaging stations," U.S. Geological Survey, Reston, VA, Techniques and Methods 3-A7, 2010.

[19] U.S. Bureau of Reclamation, *Design of Small Dams*, 3rd ed. Denver, CO: U.S. Department of the Interior, Bureau of Reclamation, 1987. [Online]. Available: https://www.usbr.gov/tsc/techreferences/mands/mands-pdfs/SmallDams.pdf

## 7.10 Exercises

**7.1** Concept check · 2 min

In case 2 the rule curve meets a second storm with 92 $\mathrm{hm}^3$ already in its 300 $\mathrm{hm}^3$ flood pool. The town's river first exceeds the channel limit on day 17.4, with the lake at 866 $\mathrm{hm}^3$, far below its 1100 $\mathrm{hm}^3$ crest. Why does the town flood while the lake still has room?

(a) The tributary below the dam rises first, and a cut in the release reaches the town a day after it is made

(b) The flood pool is too small, so the rule has to release at the channel's limit through the whole storm

(c) The lake spills over its crest, and the crest flow adds to what the gates already release

(d) The diversion stops taking the supply in a storm, so its share of the release goes down the river

**7.2** Concept check · 2 min

The scenario tree plans ten days on seven forecast members, and for the first $b = 2$ days the release at each day boundary is one variable shared by every member. Why are those releases shared?

(a) The forecast's first two days carry no rain error, so every member agrees on them

(b) They are decided before anyone can tell which member will come true, so one release must serve all seven

(c) The ramp limit stops the release from changing within two days, so the members cannot differ

(d) The dynamic program's water value fixes the release for two days, and the tree adds the branches after it

**7.3** Concept check · 2 min

The chance-constrained tree holds the channel on $k = 6$ of the $M = 7$ members. What does this setting promise?

(a) The town stays under its limit with a probability of at least 6/7, as the scenario approach guarantees

(b) Nothing about the wettest member: its branch may release anything, since its excess carries no cost

(c) The limit holds on the six driest members, the wettest pays a token price, and no flood probability is stated

(d) All seven members are held, each to six sevenths of the channel's limit, to leave a margin

**7.4** Theory and calculation · 6 min

The rule curve's flood rule releases $R_{ch} = (1-\delta)\,Q_{flood} - \hat{L} + D_s$ (Equation (7.8)) with $\delta = 0.07$ and $Q_{flood} = 450\,\mathrm{m}^3/\mathrm{s}$. Take the winter supply $D_s = 6\,\mathrm{m}^3/\mathrm{s}$, and recall that 1 $\mathrm{m}^3/\mathrm{s}$ held for a day is 0.0864 $\mathrm{hm}^3$.

(a) With no tributary flow ($\hat{L} = 0$), compute $R_{ch}$ and the volume it removes from the lake in a day. *Checkpoint:* 424.5 $\mathrm{m}^3/\mathrm{s}$ and 36.7 $\mathrm{hm}^3$ a day.

(b) How many days does that release take to empty a full 300 $\mathrm{hm}^3$ flood pool with no inflow, and with a steady inflow of 100 $\mathrm{m}^3/\mathrm{s}$? *Checkpoint:* 8.2 d and 10.7 d.

(c) The tributary runs at 200 $m^3/s$. What is $R_{ch}$ now, and how much longer does the same pool take to empty with no inflow? Use the result to say why the rule of case 3 needs days of forecast lead and not hours. *Checkpoint:* 224.5 $m^3/s$, 19.4 $hm^3$ a day, 15.5 d, nearly twice as long.

**7.5** Theory and calculation · 6 min

The reach is one Muskingum element with $K = 1$ d and $X = 0.2$ (Equation (7.5)).

(a) Compute the coefficients $C_0 = (\Delta t - 2KX)/D$, $C_1 = (\Delta t + 2KX)/D$ and $C_2 = (2K(1-X) - \Delta t)/D$, with $D = 2K(1-X) + \Delta t$, for a daily step $\Delta t = 1$ d. Check that they are positive and sum to one. *Checkpoint:* 0.231, 0.538, 0.231.

(b) The continuous law answers a unit step in $I_r$ from rest with $O(t) = 1 - e^{-t/K(1-X)}/(1-X)$. Find $O$ at $t = 0$ and at 1 d, and the time at which $O$ crosses zero. *Checkpoint:* −0.25, 0.642, 0.18 d.

(c) The dam raises its release by 100 $m^3/s$ at once. By how much does the flow leaving the reach first fall, and why does the plant's ramped release keep this dip small? *Checkpoint:* 25 $m^3/s$.

**7.6** Theory and calculation · 5 min

The lake's depth at the dam is $h = 6.69\,S^{0.4}$ with $S$ in $hm^3$, the head is $H = h - 10$, and the powerhouse makes $P_e = 9.81 \times 10^{-3}\,\eta\,Q_t H$ in MW with $\eta = 0.88$ (Equation (7.3)).

(a) Compute the head and the full-powerhouse power ($Q_t = 180\,m^3/s$) at 800 $hm^3$ and at the 1100 $hm^3$ crest. *Checkpoint:* 87.0 m and 135 MW; 100.1 m and 155.6 MW.

(b) How much energy does 1 $hm^3$ make through the turbines at 800 $hm^3$, and at 305 $hm^3$, the storage at which both trees end the challenge year? *Checkpoint:* 0.209 GWh and 0.134 GWh.

(c) By what fraction does a lake held at 305 $hm^3$ lower the energy of each later 1 $hm^3$ against one at 800 $hm^3$? Say why the challenge's score compares each run with perfect foresight ending at the run's own storage. *Checkpoint:* About 36 %.

**7.7** App challenge · 8 min

Load scenario 3, "Pre-release on a forecast: how much lead the channel needs", default seed, Rule + forecast with the Pre-release lead at 2 d as loaded. Run at 7 d/s: the case pauses itself at day 45.

(a) Read the Town peak tile, the days over the limit on the Energy, supply and exceedance tile, and the Flood pool tile. *Checkpoint:* 552 $m^3/s$, 4 days, 322 $hm^3$.

(b) Set the Pre-release lead to 3, then 4, then 5 d, each time before Run or with Reset. Record the town peak and the days over the limit, and give the shortest lead that keeps the town under 450 $m^3/s$. *Checkpoint:* 525 and 4 days, 489 and 3 days, 425 and 0 days: 5 d.

(c) With the 5 d lead the Flood pool tile reads only about 15 $hm^3$ less than with 2 d. What did the extra lead buy, if not room in the lake? *Checkpoint:* 307 $hm^3$: the lead buys timing.

Read every value to $\pm 1\,m^3/s$ or $\pm 1\,hm^3$.

**7.8** App challenge · 8 min

Load scenario 7, "The chance constraint: letting the wettest member go", default seed, with Scenario MPC as loaded. Run at 7 d/s: the case pauses itself at day 30.

(a) Read the energy on the Energy, supply and exceedance tile, the Town peak tile and the storage badge. *Checkpoint:* 67.6 GWh, 412 $m^3/s$, 1082 $hm^3$.

(b) Choose Chance-constrained MPC with Members held ($k$) at 6, press Reset and run to the pause again. Read the same three values. What did letting the wettest member go buy, and what did it cost the town? *Checkpoint:* 78.0 GWh, 340 $m^3/s$, 1082 $hm^3$.

(c) Set Members held to 5, press Reset and run again. Compare the energy with part (b), and say why the second member let go buys so little here. *Checkpoint:* 78.3 GWh, 345 $m^3/s$.

Read the energy to $\pm 0.1$ GWh and the flows to $\pm 1\,\mathrm{m^3/s}$.

**7.9** App challenge · 10 min

Load the Challenge, "a water year: two storms, a bust and a dry summer", default seed, with the Rule curve as loaded. Run at 30 d/s. The run pauses if the levee fails, press Run again, and it pauses at 30 September with the verdict on the Challenge target tile.

(a) Read the verdict, the days over the limit and the share of the perfect-foresight energy. *Checkpoint:* Not met, 8 days over, 75.6 % of 529.6 GWh, levee failed.

(b) Choose SDP policy, press Reset and run the year again. Read the same values and the end storage the tile names. *Checkpoint:* Not met, 9 days over, 89.5 % of 586.2 GWh, ending at 341 $\mathrm{hm^3}$.

(c) Choose Chance-constrained MPC with Members held at 6 (the suggested settings), press Reset and run the year. Read the verdict, the share, the end storage and the Town peak tile. *Checkpoint:* Met, 0 days over, 89.5 % of 590.8 GWh, ending at 305 $\mathrm{hm^3}$, town peak 425 $\mathrm{m^3/s}$.

(d) The SDP policy and the tree reach the same share. Why does only one of them pass, and why is the perfect-foresight energy different in each run? *Checkpoint:* The SDP floods the town; each bound ends at its own run's storage.

Read the shares to $\pm 0.1$ %.

**7.10** Application area · 3 min

In case 8 the tributary below the dam peaks at 394 $\mathrm{m^3/s}$. The scenario tree, with the ramp at 40 $\mathrm{m^3/s}$ a day, starts its cut before the rain and holds the town at 429 $\mathrm{m^3/s}$. With the environmental flow raised from 6 to 60 $\mathrm{m^3/s}$ the town reaches 459 $\mathrm{m^3/s}$ and is over the limit on 1 day.

(a) Using $Q_{town} = O + L$ and a river of at least $Q_{env}$ below the diversion, find the lowest town flow the dam can give at the tributary's peak with each environmental flow, once the reach has passed the cut. *Checkpoint:* 400 $\mathrm{m^3/s}$ and 454 $\mathrm{m^3/s}$.

(b) What should an engineer tell the holder of the dam's permit?

**7.11** Application area · 4 min

The stochastic dynamic program learns its statistics from a 100-year synthetic record whose storm days are one-day bursts of about 71 mm, 1.66 a year, while the cases declare three-day atmospheric rivers of 260 to 440 mm.

(a) What does this mismatch do to the policy in case 5 and in the challenge?

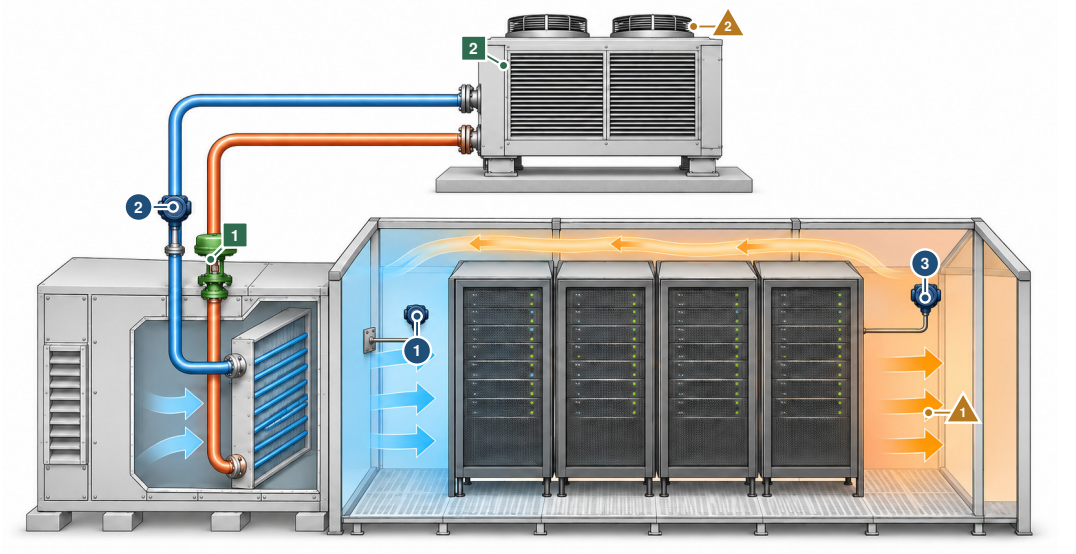
(b) Trained on full three-day storms, the program's gain falls from \$15.73 M to $-$\$4.20 M a year, and the tree that prices water with it draws the challenge's lake to 171 $\mathrm{hm^3}$ by 1 November and makes 66.3 % of its bound. Why was that record not adopted, and what would a utility check before trusting a value function?

**7.12** Review the agent · 5 min

An AI assistant reviews the challenge results and writes: "The suggested chance-constrained tree makes 528.9 GWh, more than the 515.5 GWh perfect-foresight energy for the year from 760 $\mathrm{hm^3}$, which is 102.6 %. A real controller beat perfect foresight, so the bound is loose and the 80 % target is too easy." Find the flaw and say how you would check it in the app.

CHAPTER 8

# A Data Center Cooling Simulator Whose Optimum Moves With the Wet Bulb




**The controller moves** 1 chilled water valve at the coil 2 the plant's chilled water setpoint and tower spray

**The controller reads** 1 cold-aisle (server inlet) temperature 2 chilled water supply temperature 3 hot-aisle temperature

**What disturbs it** 1 heat from the servers 2 outdoor dry and wet bulb

*One valve holds the cold aisle, while the layer above it chooses how cold the plant makes the water and how wet it runs the tower, the choice that sets the bill. Illustration, not to scale.*

### IN PLAIN TERMS A Cool Server Room at the Lowest Cost

**The main idea.** A data center is a room of computers that turn power into heat, and a cooling plant carries the heat outdoors. A simple controller holds the air the computers take in at one temperature, and a planner above it picks how the plant runs.

**The challenge.** Colder water costs more power, and a wetter cooling tower saves power but uses up water. The cheapest mix that keeps every limit moves as the outdoor air turns damp or dry.

**Why it matters.** In this chapter's hall the planner cuts the cooling bill by more than a tenth while the room stays just as cool.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in the Introduction, the facts in Sections 8.2 and 8.6, the try-it segments in Section 8.6, and the key points |
| **Study** | 65 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 25 min | the first page, the results (Section 8.6), Technology in Practice (Section 8.7), Limitations (Section 8.8) and the open challenge |

**The control question** A loop that holds the server inlet at its setpoint says nothing about what the hall costs, because the chilled water setpoint and the tower's evaporative fraction beneath it decide the electricity and the water. Where does the cheapest operating point that meets every limit sit, what holds it there as the wet bulb moves, and does a predictive layer on the same cost find a better one?

**Learning objectives**

- Compute the steady water draw, cost, chiller COP and recirculation threshold of the hall from its model equations.
- Explain why the cost is not unimodal in the evaporative fraction and why the supervisory layer therefore searches a grid.
- Identify the binding constraint at an operating point and predict how weather, coil fouling, a chiller trip and a water price move it.
- Distinguish what tuning, feedforward and a cascade improve from what only the supervisory layer improves, using cold-aisle error, energy, water and cost.
- Compare the grid search with an economic predictive layer on the same cost, before and after a chiller trip.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.6 (soft constraints), 1.4.7 (steady-state optimization), 1.4.8 (economic MPC) and 1.6.4 (how an exercise names a run), and Volume 3, Chapter 2 for feedforward and cascade structure.

**The app** https://apmonitor.com/apps/data-center-cooling/

**See also** Volume 17, Chapter 6 for an optimum outlet temperature that moves with the irradiance, and Volume 17, Chapter 13 for a supervisory layer weighed against simpler heuristics. Volume 3, Chapter 6 for a feedforward from a measurement earlier in the causal chain.

## 8.1 Introduction

The cooling plant of a data hall is a control problem whose difficulty sits one level above its regulatory loops. A chilled water valve lowers the supply air temperature, the supply air lowers the cold-aisle temperature, and a single loop settles a setpoint step to within 0.1 °C in about three minutes, so a proportional-integral controller holds the server inlet temperature at 24.00 °C against every disturbance the coil and chiller have capacity for. The regulatory loop does not decide the chilled water supply temperature, the evaporative fraction of the cooling tower, or whether the water-side economizer is in service. Those decisions set the electricity and water consumption and they do not appear on the temperature trend. In the simulator of this chapter the shipped configuration holds 24.00 °C exactly while spending $98.00 per hour against a lowest-cost feasible point of $86.18, a gap of $11.81 per hour, or roughly $103,000 per year on one 3 MW hall.

Both resources are publicly accounted. A recalibration of global data center energy use placed 2018 consumption near 1 % of world electricity and found that a large rise in computing output had been absorbed almost entirely by efficiency gains rather than by more power [1]. A spatially resolved footprint of United States data centers put them at about 1.8 % of national electricity use, attributed about 0.5 % of national greenhouse gas emissions to them, and found that one fifth of the direct water footprint of their servers is drawn from moderately to highly water-stressed watersheds [2], with the water spent indirectly in generating that electricity larger again than the water evaporated in the towers [3]. Neither quantity is decided by the hardware once the plant is built, because both are decided by setpoints.

A trend holding its setpoint is evidence about the regulatory loop and not about the operating cost of the plant, and the simulator is built to separate the two. Tuning the regulatory loop over a whole day of weather and then adding a load feedforward improves the peak cold-aisle error by a factor of four and changes the day's operating cost by 18 cents. The distinction does not appear in a block diagram, because neither the objective function nor the binding constraint is part of the loop.

> ⧖ HISTORY **2007: A Report to Congress Counts the Servers**
>
> A report to the United States Congress in 2007 estimated that the nation's data centers had used about 61 billion kilowatt-hours in 2006, 1.5 percent of national electricity, for about $4.5 billion. That was more than double the estimate for 2000. A national laboratory's update of 2016 found the growth had slowed: about 70 billion kilowatt-hours in 2014, about 1.8 percent of national use, up about 4 percent from 2010 to 2014 against nearly 90 percent from 2000 to 2005. It traced the change to slower growth in server shipments, to virtualization, and to very large data centers run at high utilization, which need fewer servers for the same services [4]. The cooling plant of every hall sits inside those totals.

The class of problem is supervisory optimization over a fast regulatory layer, in which the objective encodes a chosen price ratio and the binding constraint changes with an uncontrolled input, the steady-state optimization of Section 1.4.7. Which variables to hold constant and which to reoptimize is the plantwide question of self-optimizing control [5], the same structure appears wherever a slow layer feeds setpoints to fast loops against a live objective [6, 7], and the building services version of it is a documented retrofit practice [8, 9]. The same structure appears in any plant where a boiler steam pressure, a compressor discharge pressure or a refrigeration suction pressure was fixed at commissioning and never reoptimized, and Volume 17, Chapter 6 meets it in a solar field whose best outlet temperature moves with the irradiance.

The physical content of the model is stated with its fundamentals. Each air and water node is an unsteady energy balance over a well-mixed volume, the two heat exchangers are effectiveness models, the chiller is a Carnot term in series with a loss ceiling, and the layers are separated in time, with the regulatory loop settling in three to eight minutes and the supervisory layer executing every three minutes against a steady-state model and moving its setpoints at a rate limit [10–12]. The scenarios teach process gain that changes with operating point, feedforward against a measured load, cascade structure, anti-windup at a saturated valve and constrained steady-state optimization above regulatory control. The behavior matters industrially because chilled water and condenser water setpoints fixed at commissioning are a documented source of avoidable energy use in central cooling plants [9, 13].

The data center literature divides in a way that leaves this shape uncovered. One half is thermal, concerned with rack densities, airflow fields and the arrangement of the room [14–16], and the other half optimizes the facility as a whole with the plant treated as a set of efficiency curves [17]. Neither provides an interactive plant whose feasible region moves while the search runs. This chapter sets out a browser-hosted data hall and its cooling plant in enough detail for a reader to rebuild both, and then establishes where its constrained optimum sits, what holds it there, and which of its two decision variables the weather moves. Two measured results carry the argument. The cost surface is not unimodal in the tower evaporative fraction, so an exhaustive search is a modeling decision rather than a convenience. Over a full day of weather the chilled water setpoint takes three of the twenty-two values the search grid offers while the evaporative fraction sweeps all fifteen, from the first to the last. The plant has ten states and one steady-state evaluator shared by the cost map, the optimizer and the initializer. Above it sit a constrained grid search transcribed from the executing code and an optional predictive layer on the same cost, measured against it in Section 8.6.8. The simulator reports electricity, water, inlet margin and cost separately rather than combined into one score. The chiller correlation, the tower approach model and the search itself are established methods.

Section 8.2 reviews how a raised floor delivering 13 °C air became a hall held at 24 °C with a supervisory layer above it. Section 8.3 sets out the plant model, its parameters and its assumptions. Section 8.4 chooses the two supervisory decisions by sweeping the steady-state model, and Section 8.5 describes the regulatory loops, the grid search and the predictive layer. Section 8.6 measures the scenarios, from the hall by hand to the graded day and the predictive layer against the grid search. Section 8.7 describes the sensors, valves and controllers of real halls, Section 8.8 states the limitations, Section 8.9 concludes, and the exercises in Section 8.10 run the same cases in the app.

## 8.2 Background

The first computer rooms were cooled to a narrow band of temperature and humidity specified by the mainframe vendor, held by computer room air conditioning units around the perimeter of a raised floor. Air was pushed into the plenum, rose through perforated tiles, crossed the equipment above them and returned to the units through the room. The controlled variable was the return air temperature at the unit, an average over the whole room, so the only safe way to protect the worst rack was to make the supply air cold enough for it. Halls delivering air at 13 °C were common, and the excess cooling was paid for twice, in chiller lift and in mixing losses.

Rack power densities ended that arrangement. When a cabinet went from one or two kilowatts to ten or more, the room stopped behaving like a well-mixed volume and thermal management became a subject in its own right [15]. The response that mattered for control was the proposal to provision cooling where the heat is generated and to treat the plant as a dynamically operated system rather than a statically sized one [14]. Cold-aisle and hot-aisle rack arrangement came first and physical containment of one aisle came after it, which is what makes the recirculation term of Section 8.3 small rather than dominant.

Permission to run a hall warmer came from industry guidelines rather than from the control literature. The ASHRAE thermal guidelines for data processing environments, first issued in 2004 and revised repeatedly since, separated a recommended envelope from wider allowable classes and moved the recommended upper limit up as manufacturers agreed to it [16]. The two inlet limits in this application, 27 °C recommended and 32 °C allowable, come from that lineage, and they enter the model as constraints rather than as setpoints because that is what they are. Raising the inlet is not free. An analysis of hall temperature against facility efficiency showed the gain is not monotone, because the chiller savings from a warmer hall are eventually overtaken by the server fan power that the warmer hall causes [18], which is why the server fan ramp of Equation (8.3) is in the model at all.

A wider envelope makes free cooling reachable. Air side economizers, indirect evaporative units and water side economizers of the kind modeled here all exploit the same fact, which is that a hall permitted to run warm can often be cooled by the weather instead of by a compressor [19]. Free cooling behaves as a threshold in the wet bulb rather than as a gradient, so site selection became largely an argument about climate and the cooling tower became the component that decides how close to the wet bulb a plant can get.

Two ratios were adopted to make the argument quantitative. Power usage effectiveness, the ratio of total facility power to IT power, was promoted by the Green Grid consortium and became the standard reported metric. Water use effectiveness, in liters of make-up per kilowatt-hour of IT energy, followed as its companion once it became clear that a plant could lower one ratio by raising the other, which is the trade this application is built around. Both are reported by the application, and at the design point of Section 8.3 the hall runs at a PUE of 1.291 and a water use effectiveness of 1.77 L/kWh.

> ✦ DID YOU KNOW? **About a Third for Cooling and Power**
>
> Facility infrastructure, in which cooling dominates, took 36 percent of United States data center electricity in 2018 and 31 percent in 2024, as the national average PUE fell from 1.55 to 1.45 [20]. This chapter's hall runs at 1.291 at its design point (Table 8.5).

The setpoint question remains. Chilled water is still commonly made at a fixed temperature chosen at commissioning, condenser water is still commonly held at a fixed approach, and resetting either against load or weather is sold as an energy services retrofit rather than treated as ordinary operation [9, 13]. Present interest in direct-to-chip and immersion liquid cooling changes the hall side of the problem substantially and the plant side very little, because the heat still has to leave the building through a tower or a dry cooler, and how cold to make the water and how much of it to evaporate is the same question in every case.

## 8.3 Process Model

The plant is a 3 MW data hall with hot-aisle containment. Server fans pull air out of the cold aisle, through the racks and into the hot aisle, and computer room air handlers (CRAH) take that return air across a chilled water coil and deliver it back to the cold aisle. A water-cooled centrifugal chiller makes the chilled water, a hybrid cooling tower with wet and dry cells rejects the heat outdoors, and a plate heat exchanger across the chiller provides free cooling of the chilled water return whenever the condenser water is cold enough. The arrangement and every symbol in the equations below are shown in Figure 8.1.

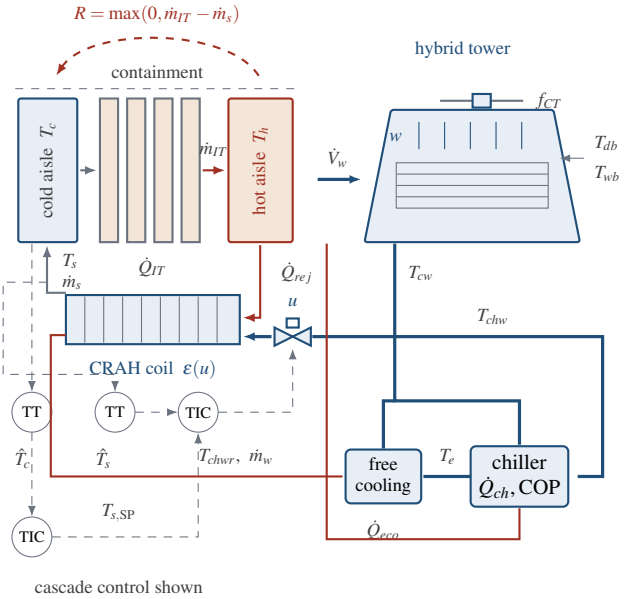



**Figure 8.1** The hall, the chilled water loop and the hybrid tower. Server fans draw $\dot{m}_{IT}$ from the cold aisle at $T_c$ through the racks into the hot aisle at $T_h$, and the CRAH delivers $\dot{m}_s$ back at $T_s$ across a coil of effectiveness $\varepsilon(u)$. Whenever the racks pull more air than the floor delivers, the shortfall $R$ is made up over the containment and lifts the inlet above the supply air. Valve $u$ is the only manipulated variable. The free-cooling exchanger takes $\dot{Q}_{eco}$ out of the return $T_{chwr}$ against condenser water at $T_{cw}$ and the chiller finishes what it can, while the tower rejects $\dot{Q}_{rej}$ to a sink set by the evaporative fraction $w$ between the dry bulb $T_{db}$ and the wet bulb $T_{wb}$, spending make-up water $\dot{V}_w$ to do it.

The controlled variable is the cold-aisle temperature, which is the server inlet temperature, and its setpoint is 24 °C. The manipulated variable is the chilled water valve at the coils, and it is the same manipulated variable in every control architecture the application offers. Table 8.1 classifies everything the loop interacts with. Ten states are integrated, and the time base is minutes throughout, with thermal capacities carried in kW min/°C so that $\mathrm{d}T/\mathrm{d}t = \dot{Q}/C$ needs no conversion factor.

### 8.3.1 Air-Side Energy Balances and Recirculation

Hot-aisle containment is never perfect, and the airflow through the racks is set by the server fans rather than by the CRAH. An unsteady energy balance over each aisle as a well-mixed node gives Equations (8.1).

$$
\begin{aligned}
C_c \frac{\mathrm{d}T_c}{\mathrm{d}t} &= \dot{m}_s c_p (T_s - T_c) + R\, c_p (T_h - T_c) \\
C_h \frac{\mathrm{d}T_h}{\mathrm{d}t} &= \dot{m}_{IT} c_p (T_c - T_h) + \dot{Q}_{hall} + B\, c_p (T_c - T_h)
\end{aligned}
\tag{8.1}
$$

**Table 8.1** Classification of Variables. Nominal Values Are Those of the Design Operating Point at 24 °C Dry Bulb and 17 °C Wet Bulb.

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $T_c$ | CV, measured | 24.00 °C | 18–27 °C SP |
| $u$ | MV | 26.7 % | 0–100 % |
| $T_{chw,\mathrm{SP}}$ | supervisory decision | 7 °C | 5–20 °C |
| $w$ | supervisory decision | 85 % | 0–100 % |
| $f_{CT}$, $f_{CRAH}$, $d_{eco}$ | operator controls | 80, 85, 100 % | see Sec. 8.4 |
| $\dot{Q}_{IT}$ | DV, measured | 3000 kW | 20–125 % |
| $T_{db}$ | DV, measured | 24 °C | −10–45 °C |
| $T_{wb}$ | DV, measured | 17 °C | −12–30 °C |
| $\phi$ fouling | DV, unmeasured | 0 | 0–40 % |
| $a_{ch}$ chiller | DV, measured | 100 % | 40–110 % |
| $a_p$ pump | DV, measured | 100 % | 45–110 % |
| $b_c$ inlet bias | DV, unmeasured | 0 °C | ±3 °C |
| $T_c, T_h, T_s$ | air states | 24.00, 38.6, 24.00 | °C |
| $T_{chw}, T_{cw}$ | water states | 7.0, 22.7 | °C |
| $\dot{Q}_{ch}$ | chiller duty | 2923 | kW, $\le \dot{Q}_{cap}$ |
| $a_v, a_w, a_f, a_d$ | actuators | 0.267, 0.85, 0.80, 1.00 | 0–1 |

where $C_c$ and $C_h$ are the two air capacities in kW min/°C, $c_p$ is the specific heat of air, $R = \max(0, \dot{m}_{IT} - \dot{m}_s)$ is air recirculated over the containment when the racks demand more air than the CRAH delivers, $B = \max(0, \dot{m}_s - \dot{m}_{IT})$ is cold air bypassing straight to the return, and $\dot{Q}_{hall}$ is every watt dissipated in the hall. Both sides carry units of kW. Solving (8.1) at steady state collapses the pair to Equation (8.2).

$$T_c = T_s + \frac{R\,\dot{Q}_{hall}}{\dot{m}_s M c_p}, \qquad M = \max(\dot{m}_{IT}, \dot{m}_s) \tag{8.2}$$

While the CRAH moves at least as much air as the racks do, $R$ is zero and the cold aisle is exactly the supply air. The moment the racks pull more, the shortfall comes out of the hot aisle and the inlet rises with the supply air unchanged. At the design point the CRAH moves 221 kg/s against 210 kg/s of server airflow, so $R$ is zero with 11 kg/s to spare.

The server fans consume that margin, because they hold a base speed until the inlet passes a knee temperature and then ramp, with fan power following the cube of speed, as shown in Equation (8.3).

$$\begin{aligned} s &= \mathrm{clamp}\big(1 + k_f(T_c - T_{knee}),\ 1,\ s_{max}\big) \\ \dot{m}_{IT} &= s\,\dot{m}_{IT,0}, \qquad P_f = s^3 P_{f,0} \end{aligned} \tag{8.3}$$

where $s$ is fan speed as a multiple of base, $k_f$ is the ramp slope and $s_{max}$ is the upper limit. The airflow term in (8.3) is stabilizing, while the power term raises the cooling load and the operating cost. Server fan power is 180 kW at the knee, 281 kW at 27 °C and 494 kW at 30 °C, and all of that power is dissipated as heat in the same hall.

### 8.3.2 Coil Effectiveness and the Chilled Water Loop

The coil is written as an air-side effectiveness that saturates in water flow, as shown in Equation (8.4), because the last third of the valve stroke adds little duty while pump power rises with the cube of flow.

$$\begin{aligned} \varepsilon(u) &= \varepsilon_{max}(1 - \phi)\,\frac{1 - \mathrm{e}^{-k_{coil}u^{0.8}}}{1 - \mathrm{e}^{-k_{coil}}} \\ \tau_{coil}\frac{\mathrm{d}T_s}{\mathrm{d}t} &= T_h - \varepsilon(T_h - T_{chw}) - T_s \end{aligned} \tag{8.4}$$

where $u$ is the valve fraction, which sets both the effectiveness and the water flow, $\phi$ is the fouling disturbance, $k_{coil}$ is a shape factor and the exponent 0.8 is the usual turbulent water-side dependence.

The saturation in (8.4) is the reason the process gain the regulatory loop sees is a function of where the supervisory layer has put the chilled water setpoint, which Section 8.6 measures.

Chilled water flow follows the valve above a minimum-flow floor,

$$\dot{m}_w = \max(\dot{m}_{w,max}\, a_p u,\ 0.06\,\dot{m}_{w,max})\,,$$

because without a bypass the return temperature grows without bound as the valve closes. The chilled water loop and the chiller's own capacity control are Equations (8.5).

$$\begin{aligned} C_{chw}\frac{\mathrm{d}T_{chw}}{\mathrm{d}t} &= \dot{m}_w c_w (T_e - T_{chw}) - \dot{Q}_{ch} \\ \tau_{ch}\frac{\mathrm{d}\dot{Q}_{ch}}{\mathrm{d}t} &= \min\big(\dot{Q}_{req},\ \dot{Q}_{cap}\big) - \dot{Q}_{ch} \end{aligned} \tag{8.5}$$

where $T_e$ is the water temperature leaving the free-cooling exchanger, $\dot{Q}_{req} = \dot{m}_w c_w (T_e - T_{chw,\mathrm{SP}})$ is the duty that would land the supply exactly on setpoint, $\dot{Q}_{cap}$ is the capacity available now, and $\tau_{ch}$ is the vane or drive lag. The min in (8.5) is the capacity constraint and nothing else in the model enforces it. Unsaturated it returns $T_{chw} = T_{chw,\mathrm{SP}}$ exactly, and saturated it lets the supply drift up until the warmer evaporator hands the machine enough capacity back to hold.

The economizer removes the first part of the return duty before the chiller sees it, as shown in Equation (8.6).

$$\begin{aligned} T_e &= T_{chwr} - \varepsilon_{eco} a_d \max(0,\ T_{chwr} - (T_{cw} + \Delta T_{HX})) \\ \dot{Q}_{eco} &= \dot{m}_w c_w (T_{chwr} - T_e) \end{aligned} \tag{8.6}$$

with $T_{chwr} = T_{chw} + \dot{Q}_{coil}/(\dot{m}_w c_w)$. Free cooling is a threshold rather than a gradient, because the max in (8.6) is zero until the condenser water plus the exchanger approach falls below the chilled water return. With 16 °C chilled water and a fully wet tower the chiller switches off entirely at any wet bulb below 9.3 °C. At 10 °C free cooling still carries 92 % of the load, at 12 °C it carries 70 %, and at 14 °C it carries 48 %. Two degrees of wet bulb near that threshold change the operating cost more than any control action available to the plant.

### 8.3.3 Chiller Performance and Capacity

Chiller performance is written as a fixed fraction of the Carnot value between evaporator and condenser, in series with a lift-independent ceiling, as shown in Equation (8.7). The two-resistance form is the reduced thermodynamic account of a vapor compression machine, in which a reversible term carries the lift dependence and a lumped internal loss carries what does not scale with lift [21].

$$\begin{aligned} \frac{1}{COP} &= \frac{T_{cond} - T_{chw}}{\eta_C (T_{chw} + 273.15)} + \frac{1}{COP_\infty} \\ T_{cond} &= T_{cw} + \Delta T_{cd} \end{aligned} \tag{8.7}$$

where $\eta_C$ is the fraction of Carnot achieved, $COP_\infty$ is the ceiling and $\Delta T_{cd}$ is the condenser approach. The correlation reproduces both of the usual rules of thumb, since one degree of warmer chilled water improves the $COP$ by 3.8 % at the design condition and one degree of colder condenser water improves it by 3.6 %, with the lift dependence carried entirely by the Carnot term. The Carnot term alone extrapolates poorly at low lift, because at 16 °C chilled water it passes a $COP$ of 40 at about 4.5 °C of lift and keeps climbing, which no machine does, since motor, drive, oil and control losses do not shrink as the lift shrinks. The ceiling term is quantitatively necessary. At the design point the Carnot term alone gives 8.52 against 6.06 for the combined model, and at the optimum 15.57 against 8.94, so the error would be largest at the operating points the supervisory layer selects.

Capacity moves with both ends of the machine, as shown in Equation (8.8).

$$\dot{Q}_{cap} = a_{ch}\, \dot{Q}_{ch,0}\, [1 + k_{ev}(T_{chw} - 7) - k_{cd}(T_{cw} - 30)] \tag{8.8}$$

where $a_{ch}$ is the fraction of the machine in service and $\dot{Q}_{ch,0}$ is the nameplate at 7 °C chilled water and 30 °C condenser water. The rated 3600 kW becomes 3960 kW at 12 °C and 3330 kW against a 35 °C condenser.

The second correction concerns which temperature (8.7) and (8.8) are evaluated at. Both use $T_{chw}$, the temperature the machine is actually leaving water at, rather than $T_{chw,\mathrm{SP}}$, the setpoint it may be failing to reach, and that distinction decides whether a plant short of cooling recovers or diverges. Evaluated at the setpoint, a saturated chiller never recovers the capacity a warmer evaporator provides, so the shortfall grows, the hall warms, the server fans ramp and (8.3) carries it away. An early implementation evaluated at the setpoint ran the hall to 45 °C on a capacity fault instead of settling. Evaluated at the real leaving temperature the same fault settles, which Section 8.6 demonstrates on a chiller trip that stops at 19.1 °C.

### 8.3.4 Hybrid Cooling Tower and Make-Up Water

A hybrid tower has wet cells and dry cells, and the evaporative fraction $w$ decides how much of the rejection is which. Fully wet cells chase the wet bulb and fully dry cells chase the dry bulb, as shown in Equation (8.9).

$$\begin{aligned} T_{sink} &= w\,T_{wb} + (1-w)\,T_{db} \\ a &= \mathrm{clamp}\left(a_0 \left(\frac{\dot{Q}_{rej}}{\dot{Q}_{ref}}\right)^{0.6} f_{CT}^{-0.8},\ a_{min},\ 30\right) \\ \tau_{CT}\frac{\mathrm{d}T_{cw}}{\mathrm{d}t} &= T_{sink} + a - T_{cw} \end{aligned} \tag{8.9}$$

where $a$ is the approach above the sink, $a_0$ is the design approach, $f_{CT}$ is fan speed and $\dot{Q}_{rej} = \dot{Q}_{ch} + P_{ch} + \dot{Q}_{eco}$ is everything the tower has to reject. The gap between the dry bulb and the wet bulb is the wet-bulb depression, and at the design condition of 24 °C dry bulb and 17 °C wet bulb it is 7 °C. The evaporative fraction determines how much of that 7 °C depression the tower reaches.

Only the evaporative fraction spends water, and blowdown follows evaporation through the cycles of concentration, as in Equation (8.10).

$$\dot{V}_w = 60\left(1 + \frac{1}{N-1}\right)\frac{f_{lat}\,w\,\dot{Q}_{rej}}{h_{fg}} \tag{8.10}$$

where $N$ is the cycles of concentration, $f_{lat}$ is the latent share of wet-cell rejection, $h_{fg}$ is the latent heat and the factor 60 converts kg/s to L/min. Water chemistry is not modeled, because basin level and cycles as a controlled variable belong to Volume 2, Chapter 5, and here the tower is treated as a heat sink with a metered water consumption. At the design point it rejects 3736 kW, of which the economizer contributes 331 kW, and makes up 88.5 L/min, a water use effectiveness of 1.77 L/kWh of IT energy.

### 8.3.5 Objective Function and Constraints

The model reduces to one scalar objective, given in Equation (8.11), and five inequalities, given in Equation (8.12).

$$\begin{aligned} J &= c_e P_{cool} + 0.06\,p_w \dot{V}_w \\ P_{cool} &= P_f + P_{CRAH} + P_{ch} + P_{pump} + P_{cw} + P_{CT} \end{aligned} \tag{8.11}$$

where $P_{ch} = \dot{Q}_{ch}/COP$, $P_{pump} = P_{pump,0}(0.12 + 0.88u^3)$, $P_{CT} = P_{CT,0} f_{CT}^3 + P_{spray,0} w$, $P_{CRAH} = P_{CRAH,0} f_{CRAH}^3$, $P_{cw}$ is the constant condenser water pump, $c_e$ and $p_w$ are the electricity and water prices in \$/kWh and \$/m$^3$, and the factor 0.06 converts L/min to m$^3$/h. IT power is excluded from (8.11) because it is the same number whatever the plant does. The five limits are Equation (8.12).

$$\begin{aligned} T_c &\le T_{c,max}, & \dot{Q}_{req} &\le \dot{Q}_{cap} \\ T_{cw} &\le T_{cw,max}, & \dot{V}_w &\le \dot{V}_{max} \end{aligned} \tag{8.12}$$

**Table 8.2** Model Parameters. Apart From $c_p$, $c_w$ and $h_{fg}$, All Constants Are Adjustable in the Application, and the Four Limits Decide Where the Optimizer Stops.

| Symbol | Meaning | Value | Unit |
|---|---|---|---|
| *Hall and air side* | | | |
| $\dot{Q}_{IT,0}$ | design IT load | 3000 | kW |
| $\dot{m}_{s,0}$ | CRAH airflow at full fan | 260 | kg/s |
| $\dot{m}_{IT,0}$ | server airflow at base speed | 210 | kg/s |
| $C_c$ | cold-aisle capacity | 60 | kW min/°C |
| $C_h$ | hot-aisle capacity | 100 | kW min/°C |
| $P_{f,0}$ | server fan power at base speed | 180 | kW |
| $T_{knee}$ | fan ramp knee | 25 | °C |
| $k_f$ | fan speed rise per degree | 0.08 | 1/°C |
| $s_{max}$ | maximum server fan speed | 2.2 | — |
| $P_{CRAH,0}$ | CRAH fan power at full speed | 120 | kW |
| $c_p$ | specific heat of air | 1.005 | kJ/kg/°C |
| *Coil and chilled water* | | | |
| $\varepsilon_{max}$ | coil effectiveness at full flow | 0.72 | — |
| $k_{coil}$ | coil NTU shape factor | 2.6 | — |
| $\tau_{coil}$ | coil and duct lag | 0.4 | min |
| $\dot{m}_{w,max}$ | chilled water flow at full valve | 150 | kg/s |
| $C_{chw}$ | chilled water loop capacity | 600 | kW min/°C |
| $P_{pump,0}$ | chilled water pump at full flow | 140 | kW |
| $c_w$ | specific heat of water | 4.186 | kJ/kg/°C |
| *Chiller* | | | |
| $\dot{Q}_{ch,0}$ | capacity at 7 °C/30 °C | 3600 | kW |
| $\eta_C$ | fraction of Carnot achieved | 0.63 | — |
| $COP_\infty$ | lift-independent COP ceiling | 21 | — |
| $\Delta T_{cd}$ | condenser approach | 5 | °C |
| $k_{ev}$ | capacity gain per degree of CHW | 0.020 | 1/°C |
| $k_{cd}$ | capacity loss per degree of CW | 0.015 | 1/°C |
| $\tau_{ch}$ | capacity control lag | 2.0 | min |

| Symbol | Meaning | Value | Unit |
|---|---|---|---|
| *Tower and economizer* | | | |
| $a_0$ | approach at design load, full fan | 4.0 | °C |
| $a_{min}$ | minimum achievable approach | 1.5 | °C |
| $\dot{Q}_{ref}$ | reference heat rejection | 3900 | kW |
| $\tau_{CT}$ | condenser loop time constant | 3 | min |
| $P_{CT,0}$ | tower fan power at full speed | 90 | kW |
| $P_{spray,0}$ | spray pump power fully wet | 15 | kW |
| $P_{cw}$ | condenser water pump | 60 | kW |
| $N$ | cycles of concentration | 4 | — |
| $f_{lat}$ | latent share of wet rejection | 0.85 | — |
| $h_{fg}$ | latent heat of vaporization | 2440 | kJ/kg |
| $\Delta T_{HX}$ | free-cooling exchanger approach | 1.5 | °C |
| $\varepsilon_{eco}$ | economizer effectiveness | 0.90 | — |
| *Limits and prices* | | | |
| $T_{c,max}$ | recommended inlet limit | 27 | °C |
| $T_{c,allow}$ | allowable inlet limit | 32 | °C |
| $T_{cw,max}$ | chiller high-head limit | 36 | °C |
| $\dot{V}_{max}$ | water budget | 95 | L/min |
| $c_e$ | electricity price | 0.10 | $/kWh |
| $p_w$ | water price | 2.00 | $/m$^3$ |
| *Actuators, instruments, supervisor* | | | |
| $\tau_v$ | valve stroke time | 0.25 | min |
| $\tau_w$ | spray staging time | 1.0 | min |
| $\tau_f$ | tower fan ramp time | 0.5 | min |
| $\tau_{eco}$ | economizer stroke time | 0.75 | min |
| $\tau_{Tc}$ | cold-aisle sensor filter | 0.5 | min |
| $\tau_{Ts}$ | supply air sensor filter | 0.10 | min |
| $T_{opt}$ | optimizer execution interval | 3 | min |
| — | chilled water setpoint rate | 0.2 | °C/min |
| — | spray fraction rate | 10 | %/min |

together with a fifth that is structural rather than numerical, namely that the coil must be able to hold the cold-aisle setpoint at all. Every limit in (8.12) came from outside the control system, since the inlet limit is an industry guideline, the capacity a nameplate, the condenser limit a protection setting and the water budget a supply agreement.

🕮 **Pause.** Try Exercise 8.3 before reading on.

### 8.3.6 Parameters, Initialization and Integration

Constants are collected in Table 8.2. Apart from the three physical properties $c_p$, $c_w$ and $h_{fg}$, every constant is displayed in the app and is adjustable, a deliberate difference from the identification apps elsewhere in this series. No constant is withheld, because the task here is not to recover a hidden parameter from a trend but to search a cost surface whose shape has been fully declared.

No scenario begins with a transient left over from its own set-up. The initializer evaluates the same steady-state function that the supervisory layer searches and that generates the displayed cost map, which is the only place the steady-state algebra is written down, so the cost map, the optimizer and the plant cannot disagree. A manual scenario is started with the operator's own valve position held fixed instead.

The engine of Section 1.6.1 carries the states forward on a 0.005 min (0.3 s) fixed fourth-order Runge-Kutta step, with the minute as the time base. Controllers execute once per 0.05 min (3 s) and ten samples per simulated minute reach the historian. Only scenario 10 carries sensor noise, drawn from its own fixed seed of 90210 (the application default is 20260829), so every run repeats exactly.

Manual operation and five automatic architectures are provided, and a finished run exports at the historian rate with the true and measured inlet, the valve command and position, the six power terms, the water flow and the lowest feasible cost with its binding constraint on every row (a zero stand-in and the flag *no legal point* while no grid point is feasible), so a run can be reanalyzed offline. Twenty-nine internal checks cover the steady state, the coil energy balance, the sign of each actuator, the recirculation term, the capacity relation (8.8), the free cooling threshold, the water arithmetic, the agreement between the supervisor and a brute-force scan of its own grid, the reporting of an undefined optimum and, in thirteen checks, the predictive layer of Section 8.5.3.

### 8.3.7 Assumptions

Each simplification below fails somewhere, and the failure mode is stated beside it.

*Two well-mixed air nodes with one recirculation term.* A real hall has a temperature field, and the number a control system uses is an average over a handful of rack-inlet sensors. Hot spots, rack-level maldistribution and the underfloor pressure profile cannot appear in (8.1), and they are the reason a real hall runs a lower average setpoint than its worst rack could tolerate.

*Airflow is commanded as a mass flow.* There are no fan curves, no plenum pressure and no filter pressure drop, so moving the CRAH fan here is easier and more linear than in a raised-floor hall with a fixed grille layout. The cube-law fan power in (8.3) is the one piece of that physics kept.

*The coil is sensible only.* Equation (8.4) has no latent cooling, no dehumidification and no condensate, so nothing stops the supply air setpoint going below the dew point except the lower output limit set on the outer loop. A real coil driven below the dew point spends duty on condensation and the hall must then be rehumidified.

*The chiller is one machine with one lag.* Real machines have a minimum load, surge limits at low lift, staging across several units and part-load curves that are not as smooth as (8.7). Chilled water flow is proportional to valve position above a minimum-flow floor, so pump staging and low delta-T syndrome are outside the model.

*The tower is a lumped approach to a linear blend.* Equation (8.9) has no fill performance curve, no plume, no freeze protection and no basin, and the linear blend between wet bulb and dry bulb is a modeling convenience rather than a psychrometric result. Cycles of concentration in (8.10) are a constant rather than a controlled variable. The weather driving it is two numbers with an optional sine, with no relative humidity, no barometric pressure and no solar gain on the envelope, and the wet bulb clamped half a degree below the dry bulb rather than derived from a psychrometric relation.

*The economics are two constants.* Real electricity is priced by time of use with demand charges, and real water is often a stepped tariff or a hard allocation. A time-of-use price would repaint the map hour by hour and would make (8.11) worth solving over a horizon rather than instantaneously. Limits annunciate and no equipment trips, so a transient remains observable past the point where equipment protection would have ended it.

*Two disturbances are stylized.* A cold-aisle sensor that acquires exactly 2 °C of bias at $t = 0$ is not the behavior of a real transmitter, but it isolates the one failure mode that optimization does not detect. The diurnal weather is a symmetric sine, which real weather is not, so that the optimum trajectory of Table 8.6 is traversed within one simulated day.

## 8.4 Selection of the Supervisory Decision Variables

The supervisory layer is given exactly two decisions, and the choice of which two is the most consequential modeling decision in the application. Both were selected by sweeping the steady-state model rather than by argument.

### 8.4.1 Chilled Water Setpoint Sweep

The chilled water setpoint trades electricity against inlet margin, capacity and pump power at the same time. Warmer water means less lift and a better *COP* in (8.7), but the coil in (8.4) is less effective, so the valve opens further, the pump works harder as the cube of its flow, and eventually the coil cannot hold the cold aisle at all. Measured at design weather with the tower 90 % wet, the valve goes 23.6, 26.7, 33.0, 39.1, 58.8, 73.9 and 92.0 % at chilled water setpoints of 5, 7, 10, 12, 15.6, 17 and 18 °C, and at 19 °C it is wide open with the hall at 24.70 °C and the setpoint lost. Cost over the same sweep runs \$100.38, \$97.17, \$92.50, \$89.61, \$86.18, \$87.28 and \$91.23 per hour. The cost is unimodal with an interior minimum, and each term in it belongs to a different piece of equipment.

### 8.4.2 Tower Fan Speed Sweep

Tower fan speed is the obvious candidate for the second decision, and the sweep rules it out. More fan means a colder condenser, a better *COP* and less heat to reject in (8.10), which means *less* water rather than more. Measured at 7 °C chilled water and 85 % spray at design weather, taking the fan from 20 % to 100 % moves the condenser water from 32.71 to 21.93 °C, the chiller from 721 to 450 kW, and the water ledger from 94.2 down to 87.7 L/min. Both objectives point the same way. The only opposing term is the fan's own cube-law power, which puts an interior minimum in cost at 80 % of \$98.00 per hour against \$100.83 at 60 % and \$99.11 at 100 %, as shown in the upper panel of Figure 8.2. The fan speed is therefore a single-objective tuning problem with no conflict, so it remains an operator control.

### 8.4.3 Evaporative Fraction Sweep

The evaporative fraction behaves differently. It is the only variable in the plant that spends water and also the only one that moves the tower sink in (8.9) toward the wet bulb, so water against electricity is a genuine trade set by the wet-bulb depression. Measured at 15.6 °C chilled water and the fan at 80 %, taking spray from 0 to 100 % moves the water ledger from 0 to 99.0 L/min and the cost through \$87.95, \$88.24, \$88.49, \$88.67, \$86.18 and \$83.58 per hour, as shown in the lower panel of Figure 8.2. Cost *rises* over the first three quarters of the range and then falls sharply, because the economizer only comes into reach at the wet end. Free cooling carries 0 % of the load at 75 % spray, 7 % at 90 % and 14 % at 100 %.

Two conclusions follow from that one sweep. The surface is not unimodal in the spray fraction, so a gradient method started at the dry end would stop at \$87.95 and report convergence while a \$86.18 point existed elsewhere in the range. And the global minimum of the sweep, \$83.58 at a fully wet tower, is not available at all, because 99.0 L/min exceeds the 95 L/min budget in (8.12). The optimum is decided by the constraint and not by the gradient, which is the justification for the exhaustive search of Section 8.5.

🕮 **Pause.** Try Exercise 8.1 before reading on.

## 8.5 Controller Design

### 8.5.1 Regulatory Layer

Every loop runs the velocity-form algorithm of Section 1.6.2, with the derivative on the measurement, written here per scan as Equations (8.13) and (8.14).

$$e_k = \alpha\,(\mathrm{SP}_k - \mathrm{PV}_k) \tag{8.13}$$

$$\Delta u_k^P = K_c\,(e_k - e_{k-1})$$

$$\Delta u_k^I = \frac{K_c}{\tau_I}\,e_k\,\Delta t_c$$

$$\Delta u_k^D = -\,K_c\,\tau_D\,\alpha\,\frac{\mathrm{PV}_k^f - 2\mathrm{PV}_{k-1}^f + \mathrm{PV}_{k-2}^f}{\Delta t_c} \tag{8.14}$$

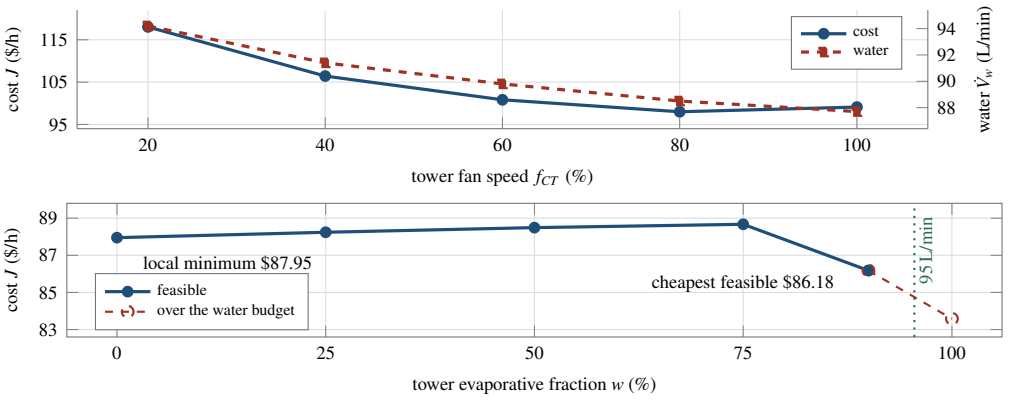



**Figure 8.2** Steady-state sweeps of the tower fan speed (upper) and the evaporative fraction (lower) at design weather. In the upper panel cost is the solid line with filled circles and water the dashed line with squares. In the lower panel feasible points are filled circles on a solid line, the points over the water budget open circles on a dashed line, and the budget a dotted vertical line. The fan sweep at 7 °C chilled water and 85 % spray lowers water monotonically while cost passes an interior minimum at 80 %, so the two objectives do not conflict. The spray sweep at 15.6 °C raises cost to $88.67 before it falls to $86.18, so the surface is not unimodal, and the lowest point on the curve, $83.58 at a fully wet tower, is infeasible because 99.0 L/min exceeds the budget.

where $\alpha = \pm 1$ carries the sign convention. The increments are combined and the demand is clipped to the actuator limits, in Equation (8.15).

$$\begin{aligned} s_k &= s_{k-1} + \Delta u_k^P + \Delta u_k^D \\ c_k &= u_{bias} + s_k + I_{k-1} \\ u_k &= \mathrm{clamp}(c_k,\ u_{min},\ u_{max}) \\ I_k &= I_{k-1} + \Delta u_k^I + \frac{\Delta t_c}{\tau_t}\,(u_k - c_k) \end{aligned} \tag{8.15}$$

In the symbols of Chapter 1, $\alpha$ is the action $a$, $s_k$ the running sum $S_k$, $c_k$ the output before limiting $v_k$ and $u_{bias}$ the output bias $u_b$. Anti-windup is back-calculation with tracking time $\tau_t = \tau_I$. At $\Delta t_c = 0.05$ min each scan removes $1/60$ of the clipped excess on the single loop ($\tau_I = 3$ min), $1/100$ on the outer cascade loop and $1/20$ on the inner one. The derivative reads a measurement smoothed by the recursion $\mathrm{PV}_k^f = \beta \mathrm{PV}_{k-1}^f + (1-\beta)\mathrm{PV}_k$, $\beta = \exp(-\Delta t_c/\tau_f)$, although every shipped tuning has $\tau_D = 0$. No controller output carries a rate limit. The valve is itself a state with a 0.25 min stroke, which is what separates the commanded position from the reached position.

Four architectures share that algorithm and the same manipulated variable. *Single* drives the valve from cold-aisle temperature with $K_c = 8\,\%/^\circ\mathrm{C}$ of valve, $\tau_I = 3$ min, $\tau_D = 0$, limits of 0 and 100 % and $\alpha = -1$, because opening the valve makes the hall colder. *Plus feedforward* adds the branch-circuit IT load measurement as the output bias $u_{bias} = K_{ff}(\dot{Q}_{IT}/\dot{Q}_{IT,0} - 1)$ in (8.15) with $K_{ff} = 0.3$, which opens the valve while the cold aisle is still on setpoint. *Cascade* makes the cold-aisle loop output a supply air setpoint with $K_c = 2.0\,^\circ\mathrm{C}/^\circ\mathrm{C}$, $\tau_I = 5$ min, limits of 12 and 26 °C and $\alpha = +1$, because a hall that is too warm needs a *lower* supply air setpoint, and a fast inner loop holds that setpoint with the valve at $K_c = 10\,\%/^\circ\mathrm{C}$ and $\tau_I = 1$ min. The feedforward is then added to the outer output, with its sign flipped and scaled by the 14 °C setpoint span, because more load asks for colder supply air rather than a larger number, and the sum is clipped again to 12 and 26 °C. That second clip lies outside the outer controller, whose back-calculation sees only its own limits. *Optimizer* leaves the cascade unchanged and puts the supervisory layer above it.

What the outer controller's output represents changes with the architecture, a valve fraction in two of them and a temperature in the other two, so carrying an integral term across that change would drive the valve closed or command a supply air setpoint of zero. The switch is handled as a transfer instead, in which every controller tracks the valve position the plant already has for one execution, which is the same seeding of $I_k$ in (8.15) that makes the transfer from manual bumpless (Section 1.6.2).

### 8.5.2 Supervisory Layer

The layer is a transparent constrained grid search rather than a model predictive controller, and it runs the same steady-state evaluator that generates the cost map. It solves the steady-state optimization of Section 1.4.7 by enumeration. Every $T_{opt} = 3\,\text{min}$ it evaluates Equation (8.16) over a $22 \times 15$ grid of cell centers.

$$T_{chw,\text{SP}}^{(i)} = 5 + \frac{15(i+\frac{1}{2})}{22}, \qquad w^{(j)} = \frac{j+\frac{1}{2}}{15} \tag{8.16}$$

for $i = 0 \ldots 21$ and $j = 0 \ldots 14$, which is 330 points. The bounds in (8.16) are the declared operating ranges of the two supervisory decisions, so the search cannot command a value outside them. Every point violating (8.12) is discarded, the lowest-cost feasible point is selected, and the four grid neighbors of that point are then probed. A neighbor that is lower in cost but infeasible identifies the constraint that limits further reduction, and that constraint is reported as *binding*. A lower-cost direction that falls outside a decision variable's own range is reported as *range*, because an optimum pressed against the first or last cell of the grid is on a bound rather than interior. A neighbor that is feasible and costlier means the optimum is interior and nothing binds. The two setpoints move toward the selected point at 0.2 °C/min and 10 %/min, so the chiller is never commanded a step change in lift.

Three design choices matter. The search runs once at reset through the engine's priming call rather than waiting for its own schedule, so the reported lowest feasible cost is correct from the first logged sample. It also runs in manual mode and in every architecture, because the identity of the binding constraint is informative even when no controller acts on it. And the layer is deliberately slow, since the plant underneath settles a setpoint step in three minutes on the single loop and eight on the cascade, and a faster supervisor would respond to its own transients rather than to the weather. The economizer valve is opened rather than optimized, because free cooling costs nothing once the tower is already running.

Exhaustive enumeration is chosen over a gradient method because of the measurement in Figure 8.2 rather than for convenience. At 330 evaluations of a fixed-point solve every three simulated minutes the search is inexpensive enough to run in the browser at the fastest simulation speed, and it cannot be trapped by the local minimum at the dry end of the spray range.

### 8.5.3 Predictive Supervisory Layer

A fifth architecture, the default only of its own scenario, replaces the grid search with a model predictive controller [22, 23] above the same cascade. Every $T_s = 1\,\text{min}$ it plans the same two decisions, $u = (T_{chw,\text{SP}}, w)$, over 30 min in ten 3 min blocks against the grid search's own cost, which makes it an economic predictive controller [24] in the sense of Section 1.4.8. A first design that also planned the supply air setpoint failed the challenge's 0.5 °C band at 0.874 to 0.905 °C over its settings, through its own overshoot of 0.90 °C after the 25 % load drop, and reached 0.763 °C with the shipped feedforward added, so the cold aisle stays with the cascade.

The plan's dynamics are the lags of its own moves, identified at the design point from step responses with the cascade closed, which give two equal lags of $\tau_{chw} = 2.063\,\text{min}$ from the chilled water setpoint to the supply (residual 4.2 % of the response), and $\tau_a = 1.061\,\text{min}$ from the spray command to its position followed by $\tau_{ct} = 3.169\,\text{min}$ to the condenser water (0.45 %). They run open loop on the applied commands, because no shipped controller measures either water loop. The steady map is the grid search's evaluator itself at the applied setpoints and the present weather, load

and parameters, whose central differences of 0.25 °C and 0.05 of spray give the cost's slope $g$ and diagonal curvature $H$ and the slopes of the five constraint quantities in the lagged states. The plan solves Equation (8.17)

$$\min_{u,s} \tfrac{T_s}{60} \sum_{k=1}^{30} \left( g^\top \delta_k + \tfrac{1}{2} \delta_k^\top \bar{H} \delta_k \right) + \sum_j \Delta u_j^\top S \Delta u_j + \sum_i \left( \rho_i s_i^2 + \sigma_i s_i \right) \tag{8.17}$$

subject to the lag model $x_{k+1} = Ax_k + Bu_k$, the rows $Cx_k \le y_{max} + s$ with $s \ge 0$, the ranges of (8.16), and $|u_1 - u_0| \le rT_s$ and $|\Delta u_j| \le 3rT_s$. Here $\delta_k = u_k - u_0$ is the departure from the applied setpoints $u_0$, $\bar{H}$ is $H$ with its diagonal floored at 2 and 80 \$/h per unit squared so that a flat stretch is still a bowl, $S = \mathrm{diag}(0.002, 0.02)$ prices each block boundary move $\Delta u_j$, $x$ holds the four lags and a constant, and $r$ is the supervisor's own rate limit of 0.2 °C/min and 10 %/min, so both layers walk the map at one speed.

The five rows are the constraints of (8.12), each with one slack over the horizon priced $(\rho_i, \sigma_i)$ per unit squared and per unit, the two prices of Equation (1.7). The full-valve cold aisle below $T_{c,max}$ and the condenser water below $T_{cw,max}$, the two limits the application raises as alarms, carry $(10^4, 10^3)$ per °C, the make-up water carries $(10^3, 10^2)$ per 10 L/min held 1 L/min inside $\dot{V}_{max}$, the full-valve cold aisle 0.25 °C below the setpoint, which keeps the cascade's authority, carries $(10^2, 10)$, and the capacity margin $\dot{Q}_{cap} - \dot{Q}_{req} \ge 0$ carries $(1, 1)$ per 100 kW. A slack for every third step was not taken, because it cost four to eight times the solve time, and on the condenser and water rows, at the first condenser price, it traded 159.1 for 20.9 min over 36 °C on the 45/30 °C day of Section 8.6.8 against 1.5 for 35.6 min over the water budget in scenario 8.

The plan tracks no output and re-evaluates its rows on the plant's steady model every minute, so it carries no disturbance estimator, and it carries no forecast because the application has none, holding the weather and the load at their present values as the grid search does. It reads only the cold-aisle setpoint, the applied commands, the rate limits and the evaluator's output. A plan takes effect when its status is optimal, its moves lie in their boxes with the first within one decision's travel, and every number is finite. Otherwise the grid search, which keeps running, takes both setpoints from where they are without a step until the next valid plan, which no run reported here needed. In Chrome the 20 decision variables and five slacks solve in a mean of 540 µs over scenario 13 (largest 6.0 ms), 0.054 % of a core at real time and 3.24 % at the 60-fold top speed.

## 8.6 Results

The fourteen scenarios of Table 8.3 are ordered by concept rather than by equipment. Each fixes an opening mode, an architecture, the three tuning constants, the disturbances in force and the events scheduled during the run, and each poses a question to predict and provides a measured answer. Every number below was measured on the running application at the prices of Table 8.2 unless the scenario sets its own.

### 8.6.1 Open-Loop Behavior and the Regulatory Baseline

Scenario 1 runs the hall with the valve parked at 27 % and no controller, so it opens at 23.83 °C. The IT load steps 8 % at $t = 5$ min and the cold aisle settles by itself at 25.09 °C, because the coil in (8.4) removes duty in proportion to how far the return air is above the chilled water, so a warmer hall is cooled more strongly. The required valve position is precise, because near the steady position of 28.9 % the inlet moves about 0.5 °C per percent of valve, landing at 22.5, 21.6 and 20.9 °C for valve positions of 32, 34 and 36 %.

The self-regulation runs out at 111 % of design load. With the valve back at 27 % the inlet creeps upward, crosses 27 °C 20.3 min after the load change and 32 °C four minutes later, peaks at 47.2 °C and comes to rest at 47.0 °C only when the server fans reach their speed limit, while at 110 % it settles at 25.5 °C. Both terms in (8.3) are responsible. Fan power rises as the cube of speed and lands

**Table 8.3** Scenario Sequence and the Mechanism Each One Exercises

| # | Scenario | What it exercises |
|---|---|---|
| 1 | Manual operation | self-regulation and its load limit |
| 2 | Single-loop regulation | regulation against a priced setpoint |
| 3 | IT load step | load step, gain sweep, oscillation |
| 4 | Measured-load feedforward | feedforward and its optimum gain |
| 5 | 7 °C to 15 °C chilled water | a disturbance the feedforward misses |
| 6 | Supervisory layer against a fixed setpoint | the four-objective trade under a budget |
| 7 | Hot humid afternoon | economizer shutoff at high wet bulb |
| 8 | Chiller capacity loss to 60 % | an infeasible search region |
| 9 | Coils fouled 25 % | equipment moves the setpoint axis |
| 10 | Cold-aisle sensor bias +2 °C | sensor bias propagates to every layer |
| 11 | Water at \$6.00 per $m^3$ | the price ratio as a policy choice |
| 12 | Twenty-four hours of weather | the diurnal optimum trajectory |
| 13 | The same cost as a 30 min plan | grid search against a predictive layer |
| 14 | Graded day, three load changes | four performance measures at once |

back in the hall as heat, and the ramping fans eventually pull more air than the CRAH delivers, at which point $R$ in (8.2) becomes positive and the inlet rises with the supply air unchanged. The limit between 110 and 111 % is set by the airflow rather than by the chiller, so the containment airflow margin, not the chiller, sets the overload capacity of the hall.

Scenario 2 closes the loop and the cold aisle holds 24.00 °C exactly. The plant spends \$98.00 per hour where the cheapest feasible point costs \$86.18, so the configuration as shipped spends \$11.81 per hour more than necessary, about \$103,000 per year on one hall. The regulatory loop is not at fault. The chilled water setpoint beneath it is.

### 8.6.2 Load Steps and Setpoint Moves Across Architectures

Table 8.4 carries the 18 % IT load step of scenarios 3 and 4, read over the pair of steps at $t = 5$ and $t = 35$ min and a run to $t = 60$ min. Half a minute of sensor filter, four tenths of a minute of coil lag and two minutes of chiller capacity control separate a load change from its appearance at the measurement, so the single loop responds late at any gain.

**Table 8.4** Cold-Aisle Response to an 18 % IT Load Increase at $t = 5$ min and Its Removal at $t = 35$ min, Read Over Both Events. Peak Is the Largest Inlet Deviation, IAE the Scorecard Value From 0 to 60 min

| Configuration | Peak (°C) | IAE (°C min) | Note |
|---|---|---|---|
| $K_c = 4,\ \tau_I = 8$ | 1.05 | 18.18 | detuned |
| $K_c = 8,\ \tau_I = 3$ | 0.78 | 5.39 | as shipped |
| $K_c = 12$ | 0.66 | 3.88 | best of the sweep |
| $K_c = 16$ | 0.59 | 8.89 | rings |
| $K_c = 20$ | 7.64 | 100.98 | unstable |
| $K_c = 8,\ K_{ff} = 0.15$ | 0.44 | 3.65 | half the job |
| $K_c = 8,\ K_{ff} = 0.30$ | 0.27 | 2.59 | as shipped |
| $K_c = 8,\ K_{ff} = 0.45$ | 0.56 | 3.33 | sign reversed |
| $K_c = 8,\ K_{ff} = 0.60$ | 1.02 | 4.98 | sign reversed |

Raising the proportional gain improves the peak error up to a limit. At 16 %/°C the peak is smaller than at 12 %/°C, 0.59 against 0.66 °C, while the integrated error more than doubles from 3.88 to 8.89 °C min because the response rings and decays only slowly, and at 20 %/°C the loop is unstable at 7.64 °C and spends 7.1 min above the recommended inlet limit. The feedforward reaches 0.27 °C for one additional measurement and no additional actuator, a 65 % reduction from 0.78 °C, and its gain

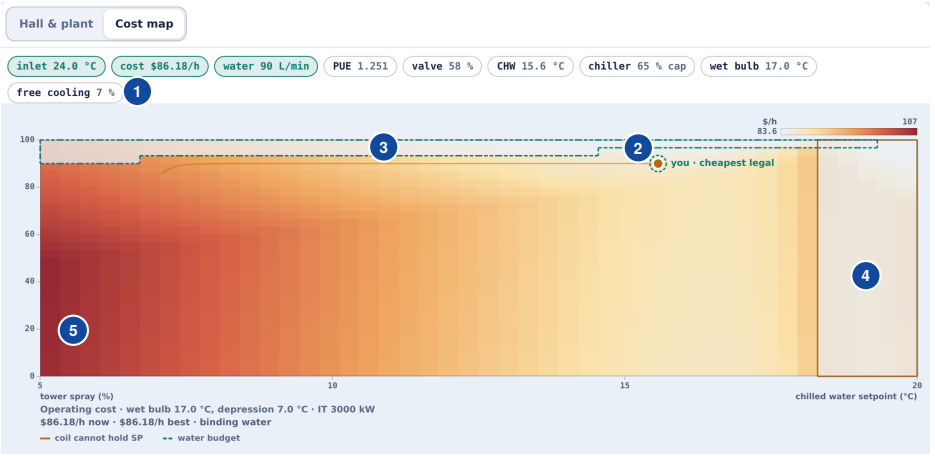



**Figure 8.3** Scenario 6 after its two-hour run, showing the cost map over the plane of the two supervisory decisions, chilled water setpoint across and tower spray fraction up. Color is operating cost. (1) The readout chips: inlet, cost rate, make-up water, valve, chilled water and free cooling. (2) The operating point. The supervisory layer has moved the plant from the shipped 7 °C chilled water setpoint to 15.6 °C with the tower 90 % wet, reached at 43 min, which is the lowest-cost feasible point at design weather (Table 8.7). (3) The water budget. The infeasible region beyond it is lower in cost than the operating point and is shown dimmed rather than removed. (4) The coil capability limit, traced on the right. (5) The cold and dry corner, where the color is darkest. Over the run the cold aisle stays within 0.05 °C of 24 °C while the cost falls from $98.00 to $86.18 per hour, make-up water settles at 90 L/min against the 95 L/min budget (Table 8.2), and the run totals are 1531 kWh and 10.73 $m^3$ with no time above 27 °C (Table 8.5). The readout under the map reports water as the binding constraint.

has an interior optimum. Up to the optimum the hall warms when the job starts and cools when it ends, as it does with no feedforward. Above it the signs reverse, because the valve moves further than the load needs in both directions. At $K_{ff} = 0.6$ the hall is overcooled by 0.79 °C at the start of the job and runs 1.02 °C warm one minute after it ends, when the feedforward has closed the valve too far. A feedforward gain twice its optimum value gives a larger peak error than no feedforward at all.

Scenario 5 separates the two structures. The chilled water setpoint is raised from 7 to 15 °C at $t = 6$ min to reduce chiller power, and the IT load is held constant. Read over a run to 30 min, the single loop peaks at 1.18 °C with an integrated error of 11.50 °C min, and the feedforward gives exactly the same two values, because the disturbance does not appear in the measured IT load the feedforward acts on. The cascade gives 0.17 and 1.59 °C min, about seven times better on both, because its inner loop measures the supply air, which responds within seconds of the coil losing duty, and it moves the valve from 27 to 54 % before the cold aisle deviates. A disturbance that enters inside the inner loop is the one a cascade rejects, the structural argument of Volume 3, Chapter 2. The same move lowers the operating cost from $98.00 to $87.69 per hour as the *COP* rises from 6.1 to 8.5, so the cascade structure is what makes an economically motivated setpoint move safe to execute.

🕮 **Pause.** Try Exercise 8.2 before reading on.

### 8.6.3 Supervisory Optimization Against a Fixed Setpoint

Scenario 6 assigns the two setpoints to the supervisory layer, which moves the plant across the cost map until it reaches a constraint, as shown in Figure 8.3. The result against the identical hours at a fixed 7 °C setpoint is given in Table 8.5. Over two hours at design weather the optimizer spends 1531 kWh against 1748 kWh, a saving of 217 kWh, and 10 732 L of water against 10 619 L, which is 113 L more water.

**Table 8.5** The Supervisory Layer Against a Fixed 7 °C Chilled Water Setpoint With the Tower 85 % Wet, Over Three Runs. PUE Is Total Facility Energy Divided by IT Energy Over the Run, Peak Error the Largest Cold-Aisle Deviation, and a Dash Indicates a Quantity Not Defined for That Run.

| | Two hours, design weather | | | Twenty-four hours, diurnal | | | Graded day, three load changes | | |
|---|---|---|---|---|---|---|---|---|---|
| | fixed | optimizer | change | fixed | optimizer | change | fixed | optimizer | change |
| Cooling energy (kWh) | 1748 | 1531 | −12.4 % | 15 959 | 12 344 | −22.7 % | 19 824 | 16 574 | −16.4 % |
| Water (L) | 10 619 | 10 732 | +1.1 % | 120 311 | 90 728 | −24.6 % | 130 817 | 111 124 | −15.1 % |
| Cost ($) | 195.99 | 174.57 | −10.9 % | 1836.55 | 1415.88 | −22.9 % | 2244.07 | 1879.60 | −16.2 % |
| Cost rate ($/h) | 98.00 | 86.18 | −$11.81 | — | — | — | — | — | — |
| PUE | 1.291 | 1.255 | −0.036 | 1.222 | 1.171 | −0.050 | 1.262 | 1.219 | −0.043 |
| Peak error (°C) | 0.00 | 0.05 | — | 0.00 | 0.06 | — | 1.73 | 0.42 | −76 % |

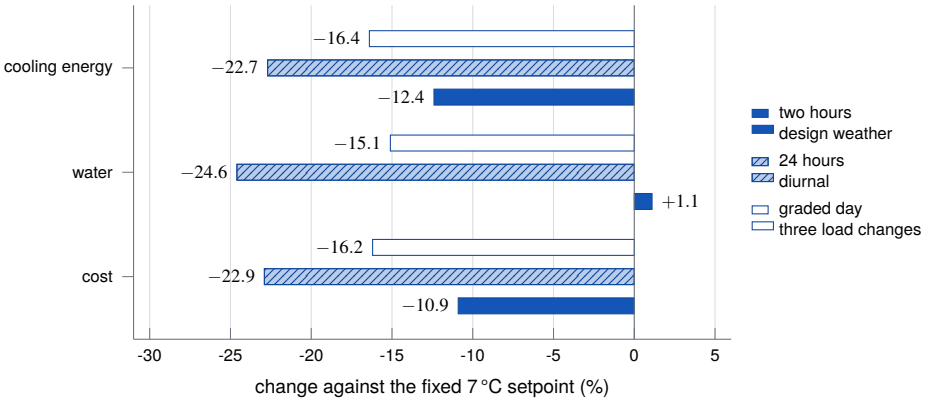



**Figure 8.4** The supervisory layer against the fixed 7 °C chilled water setpoint, redrawn from the change columns of Table 8.5: dark bars for the two hours at design weather, hatched bars for the diurnal day and open bars for the graded day. Notice that the optimizer spends 1.1 % more water over two hours at design weather but 24.6 % less over a full day, from the same objective.

> **TRY IT IN THE APP** **6 · The Supervisory Layer Against a Fixed Setpoint**
>
> Load the scenario, open the Cost map and press Run. Watch the marker walk from the shipped 7 °C chilled water setpoint to 15.6 °C with the tower 90 % wet and stop on the water budget. The Binding tile then reads water, and the cost rate has fallen from $98.00 to $86.18 per hour (Table 8.5).

The additional water is a property of the objective rather than an artifact. At that weather the direction of decreasing cost is toward a wetter tower, and (8.11) spends more of the resource that is cheaper per unit of cooling delivered. The water budget limits how far that substitution goes, and the optimum sits on it.

Over a full day the sign of the water change reverses, and Table 8.5 reports both cases for that reason. The fixed baseline sprays the tower at 85 % through the night, when evaporative operation delivers no benefit, and the optimizer does not. Measured at 6 °C dry bulb and 3 °C wet bulb with 16 °C chilled water, a 3 % wet tower costs $41.01 per hour and 2.7 L/min while a fully wet one costs $53.02 per hour and 90.7 L/min, with free cooling carrying 100 % of the load in both cases. Night spraying is charged twice, in spray pump power and in water, and returns nothing. Both results come from the same objective function evaluated at different weather. Figure 8.4 draws the three change columns of Table 8.5 side by side.

The supervisory layer also changes the process the regulatory loop was tuned against, which is a

gain-scheduling problem stated by measurement. At 7 °C the valve sits at 26.7 %, on the steep part of (8.4), where a percent of valve delivers a large change in duty. At the 15.6 °C optimum it sits at 58.5 %, on the saturating part, where a percent of valve delivers much less. Tested with a 0.5 °C step in the cold-aisle setpoint after the plant has settled, the single loop settles at 15 %/°C and sustains an oscillation at 15.5 %/°C with 7 °C chilled water, while at 15.5 °C it settles at 53 %/°C and oscillates only from 54 %/°C, a usable gain about 3.5 times larger. The outer cascade loop moves its limit from between 2.5 and 2.6 °C/°C at 7 °C to between 5 and 6 °C/°C at 15.5 °C, about twice as large. The direction of the change is the operating hazard, because a loop commissioned at the warm end and later run at the cold end during a period of coil cleaning carries a tuning that was never validated there.

### 8.6.4 Diurnal Weather

Scenario 12 runs twenty-four hours from midnight with the dry bulb swinging between 9 and 27 °C and the wet bulb between 6 and 16 °C, peaking at three in the afternoon, and with the IT load held constant so that the weather is the only input that moves. The lowest-cost feasible point through that day is given in Table 8.6, evaluated on the same steady-state function the supervisor searches, which the plant then trails by a few minutes through its rate limits.

**Table 8.6** The Lowest-Cost Feasible Point Through the Day of Scenario 12, Evaluated on the Same Steady-State Function the Supervisor Searches. The IT Load Is 3000 kW and the Cold-Aisle Setpoint 24 °C Throughout, and the Weather Profile Is Symmetric About Its 1500 Peak.

| Time | $T_{wb}$ (°C) | $T^{\star}_{chw,\mathrm{SP}}$ (°C) | $w^{\star}$ (%) | free cool. (%) | $J^{\star}$ ($/h) |
|---|---|---|---|---|---|
| 03:00 | 6.0 | 15.6 | 3 | 99 | 40.76 |
| 06:00 | 7.5 | 16.9 | 17 | 92 | 46.13 |
| 09:00 | 11.0 | 16.9 | 97 | 89 | 56.98 |
| 12:00 | 14.5 | 16.9 | 97 | 41 | 71.75 |
| 15:00 | 16.0 | 16.3 | 97 | 21 | 79.51 |
| span | 10.0 | 1.3 | 94 | 78 | 38.75 |

The result of interest is which decision variable did not move. Across the whole day the chilled water setpoint ran between 15.6 and 16.9 °C, a span of 1.3 °C, while the evaporative fraction ran from 3 to 97 % and back. In terms of the search grid in (8.16), the setpoint took three of the twenty-two available values in twenty-four hours, all three interior, while the evaporative fraction swept all fifteen from the first to the last. The weather is carried almost entirely on the water axis, so monitoring only the chilled water setpoint would indicate that the supervisory layer had not acted at all.

The mechanism follows from (8.7) and (8.6). On the cold night the economizer carries the whole load and the chiller is off, so the lift the chilled water setpoint controls no longer contributes to cost and the only remaining term on that axis is pump power, which is minimized at the lowest setpoint that keeps the valve away from its upper limit. On the hot afternoon the chiller carries three quarters of the load, warmer water improves the *COP* substantially, and the setpoint rises. The two conditions give optimum setpoints within 1.3 °C of each other. The evaporative fraction has almost no effect on cost at night and a large effect in the afternoon, because it alone determines whether the free-cooling threshold in (8.6) is crossed.

Which constraint binds is a property of the operating condition rather than of the plant, and scenario 12 is a case where no process constraint binds. Sampled every ten minutes across the day, no process constraint is reported at any of the 144 samples, and the closest approach is at the 1500 peak, where the optimum draws 94.6 L/min against the 95 L/min budget. What is reported instead is *range* at 116 of those samples against *none* at 28, so for 81 % of the day the lowest-cost feasible point lies at an end of the spray axis, fully dry or fully wet, rather than on a plant constraint. The

chilled water optimum is interior at every sample. Over most of the day the limit on operating cost is therefore the travel of the evaporative fraction itself, between a fully dry and a fully wet tower, which no wider search can extend. The design condition of 24 °C dry bulb and 17 °C wet bulb is more humid than anything this profile reaches, and at that condition the water budget binds. Table 8.7 collects the conditions that move the optimum and reports the constraint that limits each one.

**Table 8.7** The Lowest-Cost Feasible Operating Point Under Seven Conditions, With the Binding Constraint Reported. Range Marks an Optimum at the Fully Dry or Fully Wet End of the Spray Axis. The Regulatory Loop and Its Tuning Are Identical in Every Row.

| Condition | $T^{\star}_{chw,\mathrm{SP}}$ (°C) | $w^{\star}$ (%) | $J^{\star}$ ($/h) | Binding |
|---|---|---|---|---|
| Design, 17 °C wet bulb | 15.6 | 90 | 86.18 | water |
| Hot humid, 26 °C wet bulb | 16.3 | 90 | 105.87 | water |
| Coils fouled 25 % | 9.4 | 3 | 101.93 | range |
| Chiller at 60 % | — | — | — | infeasible |
| and inlet SP 26 °C | 18.3 | 97 | 81.11 | range |
| Mild weather, water $2.00 | 16.9 | 97 | 60.26 | range |
| Mild weather, water $6.00 | 15.6 | 3 | 75.84 | range |
| Mild weather, budget 40 L/min | 16.3 | 37 | 68.61 | water |

Coil fouling moves the setpoint axis further in one step than a whole day of weather does. A 25 % loss of coil effectiveness takes the optimum from 15.6 to 9.4 °C and the cost from $86.18 to $101.93 per hour, and it takes the tower from 90 % wet to 3 % and the water consumption from 90 to 4 L/min. The mechanism is the free-cooling threshold. Colder chilled water means a colder chilled water return, which puts the threshold in (8.6) out of reach, so the only remaining benefit of a wet tower is the *COP* improvement. At that operating point a 90 % wet tower saves $9.94 per hour of electricity and costs $10.82 per hour of water, a net penalty of $0.88 per hour, so the optimum is a dry tower. A change in coil condition therefore changes the optimal water policy, which is an argument against a fixed table of seasonal setpoints.

### 8.6.5 Chiller Capacity Loss

Scenario 8 removes chiller capacity and leaves 60 % in service. The direction of decreasing cost on this map is warmer chilled water, and (8.8) states that capacity increases in the same direction, so the two appear compatible. They are not. With 60 % of the machine and a 24 °C cold-aisle setpoint no feasible point exists anywhere on the grid, because the capacity constraint excludes the cold end, the coil cannot hold 24 °C at the warm end, and the two infeasible regions meet. The optimizer has walked the chilled water setpoint from 7 to 9 °C when the trip arrives at 10 min, finds no feasible point after it and stops moving the setpoint, and the chilled water itself drifts to 19.1 °C and settles there. The valve pins fully open at 34.7 min and is at its upper limit for 71 % of the first two hours, and the hall settles 0.78 °C above setpoint at 24.78 °C. The drift terminating is the demonstration that evaluating (8.7) and (8.8) at the actual leaving water temperature gives a stable response, as Section 8.3.3 states. Holding the setpoints while no grid point is feasible, 168 of the first 180 min, is deliberate policy, and the optimum is then reported as undefined. The gap and binding readouts show *no legal point*, and the export logs zero for the best cost and setpoints, the cost as the gap and that flag, where an earlier version logged a zero gap. The correction moved only those five columns, in this scenario and scenario 13.

The margin the plant requires is in a variable the supervisor is not permitted to move. Raising the cold-aisle setpoint to 26 °C when the valve pins restores feasibility at 18.3 °C chilled water and a 97 % wet tower, $81.11 per hour at the steady-state optimum and below the $98.00 the plant started at, after an excursion to 26.73 °C inside the recommended band. The plant reaches that point but does not settle on it. A 26 °C inlet ramps the server fans in (8.3) past the 221 kg/s the CRAH delivers once

the inlet exceeds 25.66 °C, so the recirculation $R$ in (8.2) switches on and off around the setpoint, and the hall cycles between 25.5 and 26.5 °C with a period of 5.0 min while the cost averages $81.09 per hour. Capacity is recovered by spending inlet margin, at a rate given by $k_{ev}$ in (8.8), where one degree of chilled water is worth 2 % of capacity.

### 8.6.6 Sensor Bias

Scenario 10 applies +2 °C of bias to the cold-aisle transmitter and changes nothing else. The controller closes the loop on the biased measurement, so it holds the hall two degrees colder than intended while the trend shows no error, with the measurement reading 23.90 °C at three hours while the true inlet is 21.98 °C. The additional cooling is not free, because colder supply air requires a more open valve and colder chilled water, and colder chilled water lowers the $COP$ in (8.7). The cost is $103.89 per hour against $98.21 with an unbiased sensor, which is 2781 kWh over three hours against 2622 kWh, a 6 % electricity penalty for a two-degree instrument error.

Enabling the supervisory layer from the start does not recover the loss. The same bias costs $93.81 per hour against $86.40 at three hours. The search never reads the measurement, because it evaluates the steady-state model at the 24 °C setpoint, so it finds the same $86.18 optimum at 15.6 °C and 90 % spray as with an honest sensor while the loop beneath it holds the hall at 22 °C. The application reports the result as a $7.63 per hour gap between the plant and that optimum, against $0.22 with an honest sensor. The optimizer solves the correct problem for a hall that is not the one being cooled, and nothing in the control system acts on the persistent gap that is the only symptom of the bias. The countermeasure is a redundant measurement at a second rack or a calibration check with a portable instrument, which belongs to the commissioning plan rather than to the control configuration.

> ✦ DID YOU KNOW? **A Bias the Single Loop Never Sees**
>
> From the simulator: a supply air sensor that reads 3 °C high leaves the single loop with a worst true error of 0.00 °C, because the single loop never reads that sensor. The cascade, whose inner loop holds on it, reaches 2.23 °C, and the supervisory layers built on the cascade reach 1.46 to 2.10 °C (Table 8.9, case W9).

### 8.6.7 Water Price and Water Budget

Scenario 11 runs mild weather at 18 °C dry bulb and 12 °C wet bulb, and changes only the water price in (8.11). At $2.00 per cubic meter the optimizer runs 16.9 °C chilled water with the tower 97 % wet, drawing 496 kW and 89 L/min at $60.26 per hour. At $6.00 it runs 15.6 °C with the tower 3 % wet, drawing 746 kW and 3 L/min at $75.84 per hour. Over two hours that is 1052 kWh and 10 673 L against 1488 kWh and 795 L, so tripling the water price saves 9878 L at a cost of 436 kWh. No part of the control system changed. One constant in the objective changed.

A hard budget behaves differently from a price. Returning the price to $2.00 and capping water at 40 L/min places the optimum between the two cases, at 37 % spray and 643 kW at $68.61 per hour, drawing 35.6 L/min and reporting water as the binding constraint where the same weather at the same price had reported only the wet end of the spray range. A price shifts the optimum continuously, while a budget makes the optimum sit on a boundary.

The graded scenario reports the same distinction in one set of numbers. Twenty-four hours of weather and three load changes run first on a detuned single loop at 4 %/°C with eight minutes of reset, where the peak cold-aisle error is 1.73 °C and the day costs $2244.07. Tuning the loop to 8 %/°C takes the peak to 1.21 °C and adding the load feedforward takes it to 0.41 °C, at $2243.84 and $2243.89. Eighteen cents separate the worst regulatory performance in the set from the best. The supervisory layer takes the same day to $1879.60 at a peak error of 0.42 °C, within 0.01 °C of the best fixed-setpoint run, as recorded in Table 8.5. Against the challenge's four limits, a ±0.5 °C band, no time above 27 °C, $2000 and 120 $m^3$ of water, the day as delivered fails three, at 1.73 °C, $2244 and 130.8 $m^3$, although the inlet never exceeds 27 °C. The supervisory layer passes all four, at 0.42 °C, $1880 and 111.1 $m^3$, with margins of 0.08 °C, $120 and 8.9 $m^3$ and no time above 27 °C.

Regulatory performance and operating cost are therefore separate measures, and improving one does not improve the other.

### 8.6.8 Predictive Layer Against the Grid Search

Scenario 13 repeats the design weather of scenario 6 and trips one chiller to 60 % at 60 min. Both layers walk from 7 °C at the same rates, arriving at 42.6 and 46.0 min, and settle on the water budget, the grid at its nearest legal cell of 15.57 °C and 90 % (90.1 L/min) and the plan on its own row at 15.96 °C and 94.5 % (94.0 L/min), at $84.96 per hour against $86.18 over 70 to 120 min of scenario 6. After the trip no grid point is feasible for 120 of the 180 min, the grid search holds its setpoints, and the hall settles at 24.78 °C, as in scenario 8. The plan raises the spray to 95.8 % and the hall settles at 24.38 °C for $0.73 per hour more (Figure 8.5). Its chilled water setpoint of 18.53 °C changes nothing, since the chiller at capacity delivers 18.69 °C regardless, and with either 15.57 or 18.5 °C held the hall settles at 24.379 °C at 95.8 % spray and 24.778 °C at 90 %. Holding the setpoints fixed in this way assigns both differences to the spray, the $1.22 per hour the plan saves before the trip and the 0.40 °C it gains after it. Over 180 min the plan spends $265.65, 2322 kWh and 16.71 $m^3$ against $265.29, 2335 kWh and 15.88 $m^3$.

> **TRY IT IN THE APP** **13 · The Same Cost as a 30-Minute Plan (MPC)**
>
> Load the scenario and run three hours on MPC with the Cost map open, through the chiller trip at 60 min. The plan raises the spray to 95.8 % and the hall settles at 24.38 °C. Press Reset, set Control architecture to Optimizer and run again: the grid search holds 90 % spray and the hall settles at 24.78 °C (Figure 8.5).

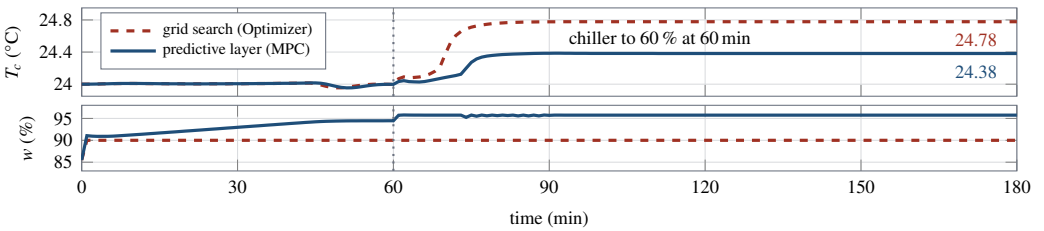



**Figure 8.5** Scenario 13, true cold aisle (upper) and spray command (lower), under the grid search (dashed) and the predictive layer (solid). The dotted vertical line marks the chiller trip. After the chiller trip the grid search holds 90 % spray with no feasible point, and the plan's 95.8 % holds the hall 0.40 °C cooler.

**Table 8.8** Cost ($) and Minutes Over the Water Budget of the Supervisory Layers. The Last Three Are Scripts, the Lowest Cost in Bold.

| Case | Grid search | Grid, retuned | MPC | Grid 88×60 | Cont. | Cont., margins |
|---|---|---|---|---|---|---|
| 13, 180 min | 265.29 | **262.87** | 265.65 | 265.38 | 264.91 | 265.13 |
| min over | 0 | 0.6 | 0 | 12.5 | 12.4 | 6.5 |
| 12, day | 1415.88 | 1411.08 | 1409.40 | 1407.54 | **1405.04** | 1409.49 |
| min over | 3.5 | 3.7 | 0 | 89.9 | 180.4 | 16.1 |
| 14, graded | 1879.60 | 1874.59 | 1873.74 | 1873.53 | **1871.95** | 1873.63 |
| min over | 21.7 | 24.6 | 19.0 | 82.1 | 305.5 | 33.3 |
| 3, job, 90 min | 139.33 | **135.91** | 137.84 | 138.22 | 138.16 | 138.38 |
| min over | 0.8 | 0 | 3.9 | 24.8 | 26.2 | 11.4 |
| 13, $T_c$ after trip | 24.78 | 24.78 | 24.39 | 24.38 | 24.39 | 24.47 |

Table 8.8 sets the plan against the shipped layer and four stronger forms of it, with nothing retuned per case and no fallback in the plan's 90 to 1440 solves per case. The retuned column is the grid

**Table 8.9** Nine Worst Cases Under Every Structure, 180 min Unless Noted. Peak $T_c$, $T_{cw}$ (°C) and $\dot{V}_w$ (L/min) With Minutes Over 27 °C, 36 °C or the Budget in Parentheses, Worst True Error $|e|$ (°C), and Minutes at Capacity (cap) or Saturated (sat). W2, W4, W5, W7, W9 Start From Scenario 6, Events at 10 min (W6 at 30 min). Feedforward Equals Single Except W5 ($|e| = 0.372$).

| Case: measures | Single | Cascade | Grid search | Grid, retuned | MPC | Grid 88×60 | Cont., margins |
|---|---|---|---|---|---|---|---|
| W1 scenario 8: $T_c$, $\dot{V}_w$, sat | 25.14, 88.5 (0), 127 | 25.14, 88.5 (0), 150 | 24.78, 92.4 (0), 145 | 24.78, 91.0 (0), 158 | 24.39, 95.3 (1.5), 123 | 24.38, 97.9 (6.2), 122 | 24.47, 96.5 (4.0), 132 |
| W2 pump to 55 %: $\dot{V}_w$ | 116.1 (2.0) | 116.1 (1.9) | 117.4 (2.1) | 101.6 (1.1) | 118.7 (3.1) | 123.6 (19.0) | 122.2 (7.2) |
| W3 scenario 7: $T_{cw}$, $\dot{V}_w$ | 32.48, 95.7 (4.8) | 32.48, 95.6 (4.5) | 31.79, 96.1 (9.0) | 31.79, 100.7 (7.4) | 32.07, 94.8 (0) | 31.88, 97.9 (9.6) | 31.85, 96.5 (8.1) |
| W4 weather to 45/30 °C: $T_{cw}$, $\dot{V}_w$, $|e|$ | 37.14 (162.4), 97.5 (166.1), 0.136, 163 cap | 37.14 (162.3), 97.4 (166.1), 0.032, 163 cap | 36.35 (158.9), 101.4 (166.5), 0.051 | 36.27 (158.0), 102.6 (165.8), 0.196 | 36.10 (147.5), 100.0 (142.9), 0.236, 5.0 sat | 35.46 (0), 107.9 (173.8), 0.057 | 35.67 (0), 106.3 (171.5), 0.055 |
| W5 load to 125 %, 120 min: $\dot{V}_w$, $|e|$ | 114.3 (109.2), 1.079, 2.0 cap | 114.9 (109.3), 0.460, 5.4 cap | 104.9 (3.5), 0.393 | 91.0 (0), 1.004, 5.4 cap, 5.8 sat | 95.7 (0.3), 0.394, 2.6 cap | 111.3 (10.6), 0.398 | 109.8 (8.3), 0.397 |
| W6 scenario 7, chiller to 60 %: $T_c$ | 50.80 (142.6) | 50.80 (139.4) | 50.26 (142.1) | 50.26 (144.7) | 49.18 (142.0) | 50.35 (142.1) | 50.32 (142.1) |
| W7 coils fouled 40 %: $T_c$ | 59.22 (169.5) | 59.22 (169.4) | 58.88 (169.4) | 58.88 (169.5) | 58.20 (169.4) | 58.49 (169.4) | 58.58 (169.4) |
| W8 scenario 10: $|e|$, $\dot{V}_w$ | 2.37, 93.2 (0) | 2.39, 100.6 (0.8) | 2.28, 104.3 (1.6) | 2.37, 103.2 (2.6) | 2.28, 105.0 (40.6) | 2.28, 107.5 (146.3) | 2.28, 106.9 (49.9) |
| W9 supply air bias +3 °C: $|e|$, $\dot{V}_w$ | 0.00, 88.5 (0) | 2.23, 102.5 (0.4) | 2.10, 100.4 (0.4) | 1.46, 97.7 (3.1) | 2.10, 101.9 (5.7) | 2.09, 106.8 (140.6) | 2.10, 105.3 (13.9) |

search under the tuning rule fixed for the plan before any run, which chose execution every minute and rates of 0.8 °C/min and 40 %/min. The rule tied the plan's own 27 weightings at $265.65 to the cent, and the water peak broke the tie afterward. The last three columns are scripts of the same cost, the grid on $88 \times 60$ points, a pattern search from the grid's best cell, and that search held to the plan's 1 L/min and 0.25 °C margins.

The plan is the cheapest on none of the four cases. Its gain over the shipped layer, $6.48 on the ambient day and $5.86 on the graded day, comes from continuous decisions and not from the horizon, because its only dynamics are its own lags and the rate limits set the walk. A grid sixteen times denser matches or beats it on every case but the job step, and the continuous search held to its margins comes within $0.54 of it on every case with more minutes over the budget (6.5, 16.1, 33.3 and 11.4 against 0, 0, 19.0 and 3.9). Free of the margins it is the cheapest on both days but rides the budget for 180.4 and 305.5 min, so the margins and the lag model buy water discipline, not cost. The retuned grid search beats the plan on scenario 13 and the job step, because the walk from 7 °C is the largest cost of the morning, and loses the days by $1.68 and $0.85. On the job step the plan is a local method on a cost with two basins in the spray at 118 % load, staying near 77.9 % mean spray over the 30 min job where the grid search goes to 21.6 %, and pays $51.52 against $50.92 before recovering $2.09 over the next 55 min. Both layers pass the challenge, at worst errors of 0.411 and 0.416 °C.

In Table 8.9 only W6 and W7 pass 27 °C, under every structure, through the server fan feedback of (8.3). The plan is worse than the shipped layer on six levels. In W4 the valve saturates for 5.0 min against none and the worst error is 0.236 against 0.051 °C, because the lags were identified at the design valve of 27 %. Near a full valve the chilled water loop is second order with a damping ratio of 0.36 and trails a ramp by 1.0 min instead of 3.5, so the plan's lag state trailed the plant by 0.45 to 0.62 °C and the chilled water overshot to 18.43 °C when the plan reversed. At the first condenser price, judging the coil row on the commands, or every row on both, cut the saturation from 4.7 to 1.1 and 2.1 min but put scenario 8 17.7 and 24.3 min over the water budget, so the transient stands as a limitation. The other four levels are the water of W1, W2, W8 and W9. The W2 peak comes 0.1 min after the pump trip, before any decision, and scales with the spray in place, 91.2 against 90.0 %. In W8 and W9 the steady map assumes the cascade holds the true cold aisle on its setpoint, the biased sensor holds it 2 °C colder, and the plant's water sits 0.91 L/min above the plan's row (94.90 against 94.00), nearly the whole margin, so the premise that a re-evaluated steady model leaves nothing

to estimate fails under sensor bias. The grid search hides the same error (90.74 against 90.0) in its 5 L/min cell. A 2 L/min margin clears W8 to 1.6 min but costs $3.98 and $1.87 on the two days and raises the trip error of scenario 13 from 0.386 to 0.454 °C. The plan is better on the cold aisle of W1, the water of W3 and W5, the peaks of W6 and W7 (by 1.08 and 0.68 °C) and both broken limits of W4, the last only since a review priced the condenser row with the inlet row, as the application ranks its alarms. At the water row's price the plan had bought 5.4 L/min of water with 0.04 °C of head, peaking at 36.43 °C for 159.1 min, and the repricing moved no other case by a cent.

## 8.7 Technology in Practice

### What Measures It

Performance monitoring of a data hall puts temperature and humidity sensors at the air inlets of the IT equipment, and wireless networks of sensors across the rack inlets are becoming more common. Supply air temperature and humidity are monitored at each room air handler, and the facility's energy monitoring and control system trends every measured value and archives it for at least a year. Linking the servers' own on-board temperature sensors to that system is a developing technology [25]. In this chapter the cold-aisle transmitter and the supply air transmitter of Figure 8.1 stand for those two measurements, each a first-order filter on the true temperature.

> **TECHNOLOGY Sensor: The Rack Inlet Thermometer**
>
> The temperature a data hall is held to is read where air enters the IT equipment, at heights set by the industry thermal guidelines, by wired or wireless sensors outside the rack or by the servers' own on-board sensors read over the network. Guidance asks that supply temperature and airflow be controlled from these intake readings and not from the return air [25]. An electrical sensor of this kind is a resistance detector, a metal element whose resistance rises almost linearly with temperature, or a thermistor, a semiconductor of high sensitivity over a narrow, strongly nonlinear range [26]. Accuracy must fit the use: a pair of sensors good to $\pm 1.5$ °F says nothing about a chilled water difference that can fall to 5 °F [25]. The chapter's cold-aisle sensor has a 0.5 min filter and an optional bias $b_c$ (Tables 8.1 and 8.2).

### What Moves It

Variable-speed drives on the room air handlers vary the airflow with the load. On the plant side, variable-flow pumping and chillers with variable-speed compressors keep the plant efficient at low load. A water-side economizer is a heat exchanger between the chilled water and the cooling tower water. Placed upstream of the chillers, it cools the returning chilled water first: when it can reach the chilled water setpoint the chillers are bypassed, and otherwise they run at reduced load to finish the job. Larger towers with variable-speed fans approach the wet bulb more closely [25]. The chapter's plate exchanger in series (Equation (8.6)), its chiller with a 2 min capacity lag and its tower approach model are this arrangement in lumped form.

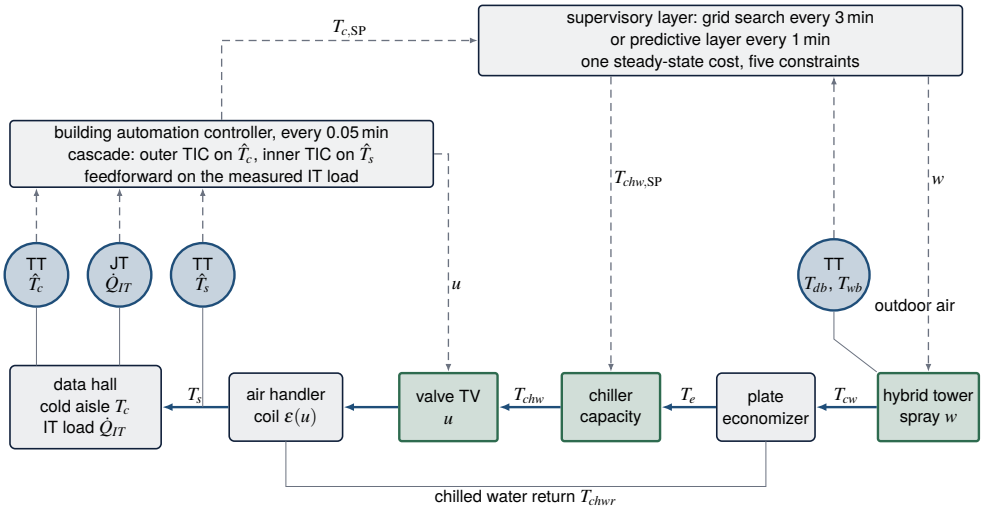



**Figure 8.6** The chapter's loops on field hardware, drawn from Sections 8.3 and 8.5. Green boxes are the final elements (the chilled water valve TV, the chiller's capacity control and the tower spray), blue circles the sensors (cold aisle, supply air, IT load, outdoor dry and wet bulb), dashed arrows signals. Notice that the supervisory layer moves only the chiller and the tower, while the valve, the one manipulated variable of every architecture, stays with the regulatory loops.

> **TECHNOLOGY** **Final Element: The Chilled Water Valve**
>
> A control valve throttles the chilled water to each air handler's coil. In a globe valve a plug moves against a fixed ring seat, and the shape of the plug gives a quick-opening, linear or equal-percentage characteristic, the last two the best for throttling. Hysteresis makes the flow depend on the direction the valve last moved, and when it grows large it can hold a loop in oscillation around its setpoint [27]. Guidance for data centers asks for two-way valves rather than three-way, so that variable-speed pumps, not a bypass, set the chilled water flow [25]. The chapter's valve is a 0.25 min stroke with a minimum-flow floor and no stuck-valve fault (Section 8.8).

### Where the Controller Runs

A building automation system controls heating, ventilation and air conditioning among a building's other systems, and its sensors reach building controllers over wired or wireless paths [28]. Supervisory loops and the loops below them run with cycle times from milliseconds to minutes [28]. Guidance asks that the control system be programmed to maximize the efficiency of the cooling under changing weather and IT load, with all air handlers communicating and working to the same temperature [25]. In the chapter the loops execute every 0.05 min, the grid search every 3 min and the predictive layer every 1 min, as Figure 8.6 draws them.

> **IN PRACTICE** **Raise the Setpoints, a Little at a Time**
>
> A federal best-practices guide asks operators to raise the supply air temperature until the most demanding equipment intake sits as close to 80 °F as possible and, leaving room for error, suggests a setpoint of 77 to 79 °F, reached in small steps and only after air management is in place. Supply air and chilled water temperatures are to be set as high as the needed cooling capacity allows. Chilled water at 55 °F or above helps the chiller and adds economizer hours, and the chilled water supply may be reset higher at low load [25]. The supervisory layer of this chapter makes that reset against a 24 °C inlet setpoint.

### What Is Modernizing It

Three developments stand out. A hybrid of dry and wet heat rejection, priced as this chapter prices it, has run at one computing center. A dry cooler that moves heat in a passive refrigerant cycle was placed upstream of the cooling towers, and its fan controls watch the difference between the entering water and the dry bulb and the ratio of the water price to the electricity price, adding dry cooling as either rises. In its first year the site's water use effectiveness was 0.70 L/kWh against 1.27 L/kWh with towers alone, a saving of 4400 $m^3$ [29]. Direct liquid cooling moves heat into a fluid at or near the chip, and its warmer chilled water pairs well with a water-side economizer [25]. The data centers with the lowest PUEs are those that cool their artificial intelligence servers with liquid [20]. Supervisory optimization is the third: the chilled water reset that guidance gives as a rule, chosen by a search against prices and limits.

## 8.8 Limitations

The supervisory layer searches a fixed grid and moves at a rate limit. It has no prediction horizon, no model of the transient its own move causes, and no mechanism to accept a higher cost now for a lower cost later, and it does not account for chiller starts, staging or minimum run times. It also solves against exactly the same steady-state model the plant is built from, which no real optimizer has available. Plant and model mismatch, and the parameter estimation or bias updating that a real-time optimization layer requires to handle it, are the largest omission in the application.

Only two decisions are optimized. The cold-aisle setpoint, the tower fan and the CRAH fan remain operator controls, and scenario 8 demonstrates that the margin the plant requires is in a variable the supervisor may not move. The control scope is PID with a feedforward, a cascade, a grid-search supervisor and the linear predictive layer of Section 8.5.3, with no nonlinear model predictive control and no explicit gain scheduling on the chilled water setpoint even though Section 8.6 measures a change of about a factor of 3.5 in usable proportional gain across the range the supervisor moves through. Instruments are exact apart from the bias, the noise and the filters set by the user, so drift, a stuck valve and an inlet sensor mounted at an unrepresentative rack are absent, and the last of those is the most consequential real failure the application does not represent.

The predictive layer models only its own lags at the design valve, with no cascade transient, cross curvature, forecast, water measurement or estimator, so it corrects neither the 45/30 °C day's saturation nor the 0.91 L/min error under sensor bias. Moving horizon estimation on the same model could estimate the unmeasured coil fouling $\phi$ and the cold-aisle sensor bias $b_c$ from the valve position, the supply air temperature and the chiller duty, which is the soft sensing whose absence scenarios 9 and 10 demonstrate.

Nothing here supports a quantitative claim about a real hall. With no airflow field, no psychrometrics, no latent cooling, a lumped tower and two constant prices, the model cannot predict a PUE, a water consumption or the setpoint a particular site should run. What the application does establish is the structure of the decision, which is that four objectives conflict, that the optimum lies on a constraint, and that the constraint moves with the uncontrolled inputs. Where each assumption fails is set out in Section 8.3.7.

**OPEN CHALLENGE** **Demand That Moves in Time**

The 2025 update of United States data center electricity use asks for research that characterizes how data center demand varies in time, so that the grid or on-site generation can serve it, and that explores ways of shifting the timing of loads, called demand flexibility. It also asks how the sector's growth may be constrained by water resources [20]. This chapter prices electricity and water as two constants and holds the weather and load at their present values, so its predictive layer gains from continuous decisions and not from its horizon. Progress would be a plant objective with time-varying prices and forecasts, over which a plan's horizon has something to buy.

## 8.9 Conclusion

This chapter has set out a data hall and its cooling plant as an interactive laboratory for supervisory optimization above a regulatory layer. The simulator integrates ten states covering two air nodes with a recirculation term, a coil saturating in water flow, a chilled water loop whose capacity clamp is the constraint, a condenser loop and four actuator positions, with a chiller written as a fraction of Carnot in series with a ceiling of 21 and evaluated at the temperature the machine is actually leaving water at. The regulatory loop holds 24.00 °C wherever the plant has the capacity, and at the shipped 7 °C chilled water setpoint it costs $98.00 per hour against $86.18 for the cheapest point that satisfies all five constraints. The supervisory layer is a $22 \times 15$ constrained grid search, chosen over a gradient method because an evaporative fraction sweep at 15.6 °C rises from $87.95 to $88.67 before falling to $86.18. Over two hours at design weather it saves 217 kWh and spends 113 L more water, and over a diurnal day it saves 22.7 % of the electricity and 24.6 % of the water, while the chilled water setpoint moves 1.3 °C and the evaporative fraction moves from 3 to 97 %. A cold-aisle sensor two degrees high produces a trend with no visible error, a hall two degrees colder than intended and a 6 % electricity penalty that the optimizer propagates rather than detects. A predictive layer on the same cost saves $6.48 and $5.86 on the two days through continuous decisions rather than its horizon. A fuller model would add latent cooling with a psychrometric model, a time-of-use electricity price, plant and model mismatch with a bias-updating layer, moving horizon estimation and nonlinear predictive control on the same model, and a gain schedule on the chilled water setpoint. In its present form the simulator separates regulatory performance from plant operating cost in a browser, with nothing to install.

**KEY POINTS**

- A loop that holds the server inlet on its setpoint says nothing about the cost, which the chilled water setpoint and the tower spray decide.
- The cost is not unimodal in the evaporative fraction, so the supervisory layer searches a grid rather than following a gradient.
- Over a day the chilled water optimum moves 1.3 °C while the evaporative fraction sweeps from 3 to 97 %.
- A predictive layer on the same cost saves $6.48 and $5.86 on the two days through continuous decisions rather than its horizon.

**Where else this applies.** An office building's supervisory layer (Chapter 9) also gains little from a plan's horizon: on the design day the 24-hour plan's bill is $557.61 against $552.74 for its supervisor retuned once (Table 9.5).

## References

[1] E. Masanet, A. Shehabi, N. Lei, S. Smith, and J. Koomey, "Recalibrating global data center energy-use estimates," *Science*, vol. 367, no. 6481, pp. 984–986, 2020.

[2] M. A. B. Siddik, A. Shehabi, and L. Marston, "The environmental footprint of data centers in the United States," *Environmental Research Letters*, vol. 16, no. 6, p. 064017, 2021.

[3] B. Ristic, K. Madani, and Z. Makuch, "The water footprint of data centers," *Sustainability*, vol. 7, no. 8, pp. 11 260–11 284, 2015.

[4] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner, "United states data center energy usage report," Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-1005775, 2016.

[5] S. Skogestad, "Plantwide control: the search for the self-optimizing control structure," *Journal of Process Control*, vol. 10, no. 5, pp. 487–507, 2000.

[6] S. Engell, "Feedback control for optimal process operation," *Journal of Process Control*, vol. 17, no. 3, pp. 203–219, 2007.

[7] M. L. Darby, M. Nikolaou, J. Jones, and D. Nicholson, "RTO: An overview and assessment of current practice," *Journal of Process Control*, vol. 21, no. 6, pp. 874–884, 2011.

[8] S. Wang and Z. Ma, "Supervisory and optimal control of building HVAC systems: A review," *HVAC&R Research*, vol. 14, no. 1, pp. 3–32, 2008.

[9] Z. Ma and S. Wang, "Supervisory and optimal control of central chiller plants using simplified adaptive models and genetic algorithm," *Applied Energy*, vol. 88, no. 1, pp. 198–211, 2011.

[10] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[11] B. W. Bequette, *Process Control: Modeling, Design, and Simulation*. Upper Saddle River, NJ: Prentice Hall, 2003.

[12] K. J. Åström and R. M. Murray, *Feedback Systems: An Introduction for Scientists and Engineers*, 2nd ed. Princeton, NJ: Princeton University Press, 2021.

[13] S. Huang, W. Zuo, and M. D. Sohn, "Improved cooling tower control of legacy chiller plants by optimizing the condenser water set point," *Building and Environment*, vol. 111, pp. 33–46, 2017.

[14] C. D. Patel, C. E. Bash, R. Sharma, M. Beitelmal, and R. Friedrich, "Smart cooling of data centers," in *2003 International Electronic Packaging Technical Conference and Exhibition, Volume 2*. ASME, 2003, pp. 129–137.

[15] R. R. Schmidt, E. E. Cruz, and M. Iyengar, "Challenges of data center thermal management," *IBM Journal of Research and Development*, vol. 49, no. 4.5, pp. 709–723, 2005.

[16] K. Ebrahimi, G. F. Jones, and A. S. Fleischer, "A review of data center cooling technology, operating conditions and the corresponding low-grade waste heat recovery opportunities," *Renewable and Sustainable Energy Reviews*, vol. 31, pp. 622–638, 2014.

[17] L. Parolini, B. Sinopoli, B. H. Krogh, and Z. Wang, "A cyber-physical systems approach to data center modeling and control for energy efficiency," *Proceedings of the IEEE*, vol. 100, no. 1, pp. 254–268, 2012.

[18] M. K. Patterson, "The effect of data center temperature on energy efficiency," in *2008 11th Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronic Systems*. IEEE, 2008, pp. 1167–1174.

[19] H. Zhang, S. Shao, H. Xu, H. Zou, and C. Tian, "Free cooling of data centers: A review," *Renewable and Sustainable Energy Reviews*, vol. 35, pp. 171–182, 2014.

[20] S. J. Smith, A. Hubbard, A. Newkirk, M. Ganeshalingam, B. Holecek, D. Sartor, M. Mills, and A. Shehabi, "United states data center energy usage report: 2025 update," Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-2001758, 2026.

[21] J. M. Gordon and K. C. Ng, "Thermodynamic modeling of reciprocating chillers," *Journal of Applied Physics*, vol. 75, no. 6, pp. 2769–2774, 1994.

[22] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[23] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[24] M. Ellis, H. Durand, and P. D. Christofides, "A tutorial review of economic model predictive control methods," *Journal of Process Control*, vol. 24, no. 8, pp. 1156–1178, 2014.

[25] O. Van Geet and D. Sickinger, "Best practices guide for energy-efficient data center design," National Renewable Energy Laboratory for the U.S. Department of Energy Federal Energy Management Program, Golden, CO, Tech. Rep. DOE/GO-102024-6283, 2024.

[26] A. Boghossian, J. Brown, and S. Zak, "3.2: Temperature sensors," in *Chemical Process Dynamics and Controls*, P. Woolf, Ed. University of Michigan open textbook, LibreTexts, 2023.

[27] D. Katzman, J. Moreno, J. Noelanders, and M. Winston-Galant, "Valves: Types, kinds, and selection," in *Chemical Process Dynamics and Controls*, P. Woolf, Ed. University of Michigan open textbook, LibreTexts, 2023, ch. 3.9.

[28] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, Special Publication 800-82r3, 2023.

[29] D. Sickinger, O. Van Geet, S. Belmont, T. Carter, and D. Martinez, "Thermosyphon cooler hybrid system for water savings in an energy-efficient HPC data center: Results from 24 months and impact on water usage effectiveness," National Renewable Energy Laboratory, Golden, CO, Tech. Rep. NREL/TP-2C00-72196, 2018.

## 8.10 Exercises

**8.1** Concept check · 2 min

The supervisory layer prices all 330 points of its grid every three minutes instead of following the slope of the cost downhill. Which measurement in Section 8.4 makes that a modeling decision rather than a convenience?

(a) Cost falls steadily toward a wetter tower, so a slope step would overshoot the water budget

(b) The chilled water sweep has two minima, one near 7 °C and one near 15.6 °C

(c) The tower fan sweep has an interior minimum at 80 %, which a slope method cannot find

(d) Cost rises over most of the spray range before it falls, so a search started dry stops early

**8.2** Concept check · 1 min

In Scenario 5 the chilled water setpoint rises from 7 to 15 °C at 6 min while the IT load stays constant. The single loop and the feedforward both peak at 1.18 °C, the cascade at 0.17 °C. Why does the cascade do better?

(a) Its inner loop measures the supply air, which answers within seconds, and opens the valve early

(b) Its feedforward enters the outer output, so it sees the setpoint change before the cold aisle does

(c) It moves the CRAH fan as well as the valve, so it has a second handle on the supply air

(d) Its outer loop has a larger gain than the single loop, so it reacts harder to the same error

**8.3** Concept check · 2 min

What does the supervisory layer minimize every three minutes?

(a) The deviation of the cold aisle from its 24 °C setpoint, weighted by the price of electricity

(b) The cost of cooling electricity and make-up water, subject to five limits that include holding the setpoint

(c) The power usage effectiveness of the hall, subject to the water budget and the chiller capacity

(d) The make-up water the tower draws, subject to the inlet limit and the condenser high-head limit

**8.4** Theory and calculation · 6 min

At the design point the tower runs 85 % wet and rejects 3736 kW, the cooling equipment draws 874 kW and the IT load is 3000 kW. Use Equations (8.10) and (8.11) and Table 8.2.

(a) Compute the make-up water $\dot{V}_w$ in L/min. *Checkpoint:* 88.5 L/min.

(b) Compute the water use effectiveness, liters of make-up per kilowatt-hour of IT energy. *Checkpoint:* 1.77 L/kWh.

(c) Compute the cost rate $J$, split into electricity and water, and the PUE. *Checkpoint:* \$87.40 and \$10.62, \$98.02 per hour, PUE 1.291.

**8.5** Theory and calculation · 6 min

At the design point the chiller makes 7 °C water against condenser water at 22.71 °C. Use Equation (8.7) with $\eta_C = 0.63$, $COP_\infty = 21$ and $\Delta T_{cd} = 5\,^\circ\mathrm{C}$.

(a) Compute the Carnot term alone, $\eta_C(T_{chw} + 273.15)/(T_{cond} - T_{chw})$. *Checkpoint:* 8.52.

(b) Compute the combined $COP$ and the chiller's electric power for a duty of 2923 kW. *Checkpoint:* 6.06 and 482 kW.

(c) Raise the chilled water to 8 °C and find the percentage gain in $COP$. Explain why the error of the Carnot term alone grows at the operating points the supervisory layer selects. *Checkpoint:* 6.29, a gain of 3.8 %.

**8.6** Theory and calculation · 7 min

The server fans follow Equation (8.3) with $T_{knee} = 25\,^\circ\mathrm{C}$, $k_f = 0.08/^\circ\mathrm{C}$, $\dot{m}_{IT,0} = 210\,\mathrm{kg/s}$ and $P_{f,0} = 180\,\mathrm{kW}$. The CRAH units deliver $\dot{m}_{s,0} f_{CRAH}$, with $\dot{m}_{s,0} = 260\,\mathrm{kg/s}$, and their fans draw $P_{CRAH,0} f_{CRAH}^3$ with $P_{CRAH,0} = 120\,\mathrm{kW}$.

(a) Find the fan speed, server airflow and server fan power at a 27 °C inlet. *Checkpoint:* 1.16, 243.6 kg/s, 281 kW.

(b) At the shipped $f_{CRAH} = 85\,\%$, find the inlet temperature above which $R$ in (8.2) is positive. Why does a hall held at a 26 °C setpoint cycle? *Checkpoint:* 221 kg/s, 25.65 °C.

(c) Repeat (b) at $f_{CRAH} = 90\,\%$ and find what the extra CRAH fan power costs per hour at \$0.10 per kWh. *Checkpoint:* 234 kg/s, 26.43 °C, 13.8 kW or \$1.38 per hour.

**8.7** App challenge · 8 min

Load Scenario 4, *The load is known before the temperature moves*, at the default seed and switch to the Advanced level. Make three runs to 60 min at a speed of 5 (five simulated minutes per second), each with *IT load feedforward gain* set before Run (Reset between runs keeps the new value): 0.3 as shipped, 0, and 0.6. At the end of each run, pausing between 60 and 65 min, read the IAE tile and the *Cost now* tile. Export the CSV (its first lines start with #, and the time column reads in hours) and find the largest positive and the most negative value of `inlet_error_C`, which is the cold aisle minus its setpoint.

(a) Give the IAE and the two extreme errors of the shipped gain and of gain 0, to 0.05 °C min and 0.02 °C. By what percentage does the feedforward cut the larger error? *Checkpoint:* 0.3: 2.59, +0.27 and −0.26. 0: 5.39, +0.78 and −0.77. About 65 %.

(b) Give the same three numbers for gain 0.6 and say when each extreme occurs. Explain the signs. *Checkpoint:* 4.98, +1.02 at about 36 min and −0.79 at about 6 min.

(c) Compare *Cost now* across the three runs and explain. *Checkpoint:* \$98.08 to \$98.10 per hour at 60 min and \$98.01 to \$98.02 at 65 min, the three runs within \$0.02 of one another at the same minute.

**8.8** App challenge · 9 min

Load Scenario 8, *A chiller trips and capacity becomes the constraint*, at the default seed. Make three runs to 180 min at a speed of 20, with every setting made before Run and Reset between runs, which keeps them: (1) as loaded, (2) with *Cold-aisle setpoint* at 26 °C, and (3) with the Advanced level and *CRAH fan speed* also at 90 %. For each run read the *Binding*, *Cost now* and *Over 27 °C* tiles at the end, and from the exported CSV the lowest and highest `inlet_C` between 2.5 and 3.0 h.

(a) Give where run 1's hall settles and what the Binding tile reads. *Checkpoint:* 24.78 °C, no legal point.

(b) Give run 2's band of `inlet_C` and the range of `cost_usd_h` over the same window, to 0.05 °C and $0.10. *Checkpoint:* 25.46 to 26.54 °C, $79.38 to $84.08 per hour.

(c) Give run 3's cold aisle and *Cost now*, and explain the difference from run 2. *Checkpoint:* 26.00 °C, $78.63 per hour.

**8.9** App challenge · 9 min

Load Scenario 13, *The same cost as a 30-minute plan (MPC)*, at the default seed and run it as loaded to 180 min at a speed of 20. Press Reset, set *Control architecture* to Optimizer before Run, and run the same 180 min. For each run export the CSV and read, in the row at 3.00 h, `energy_kWh`, `water_liters`, `cost_usd`, `inlet_C` and `spray_pct`.

(a) Give the energy, water and cost of each run, to 2 kWh, 20 L and $0.05. Which layer is cheaper over the three hours? *Checkpoint:* MPC 2322 kWh, 16 706 L, $265.65. Optimizer 2335 kWh, 15 878 L, $265.29.

(b) Give the cold aisle and the spray of each run after the trip, and the difference in cold aisle. *Checkpoint:* MPC 24.38 at 95.8 %, Optimizer 24.78 °C at 90 %, 0.40 °C.

(c) The plan is cheaper per hour before the trip. Why is it not cheaper over the run, and what does the Optimizer do after the trip that the plan does not?

**8.10** Application area · 4 min

A site on a $2.00 per $m^3$ water tariff is asked to cut make-up water on mild days. Use the Scenario 11 results of Section 8.6.

(a) Compare pricing water at $6.00 per $m^3$ inside the objective with keeping $2.00 and capping water at 40 L/min, in cooling power and water.

(b) Name one fact from the chapter's Introduction that argues against judging the trade on the tower's water alone.

**8.11** Application area · 3 min

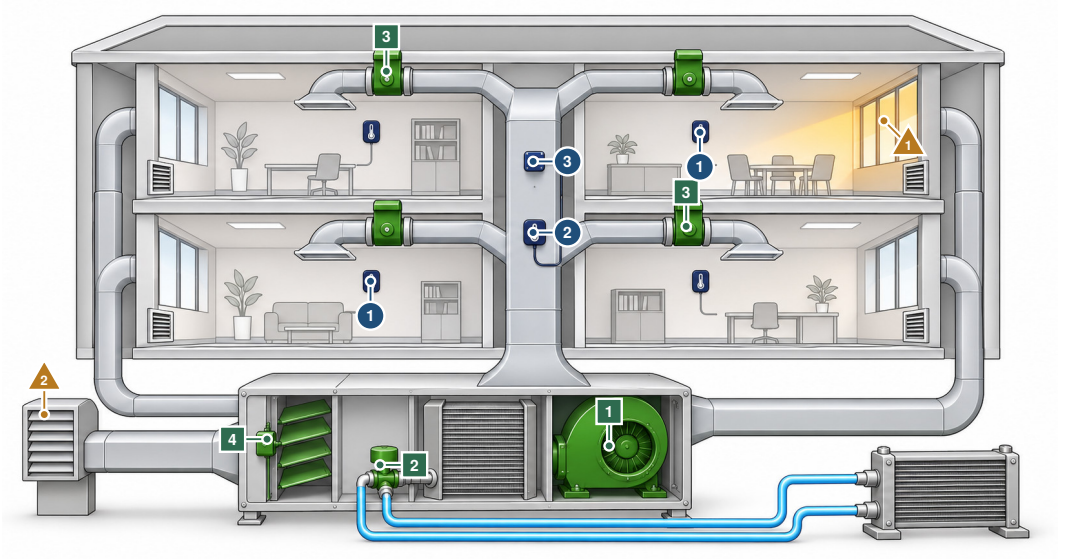
A commissioning engineer sees a cold-aisle trend sitting on its 24 °C setpoint with no visible error, while the optimizer's *Above the optimum* tile holds at about $7.6 per hour. What does the chapter say this pattern indicates, what does it cost, and what check would find the cause?

**8.12** Review the agent · 5 min

An AI assistant reviews the graded challenge and writes: *"The day fails because the single loop is detuned at 4 %/°C. Switch to + Feedforward with* $K_c = 8\,\%/^\circ\mathrm{C}$, $\tau_I = 3\,\mathrm{min}$ *and* $K_{ff} = 0.3$. *The worst error falls from 1.73 to 0.41 °C, and the tighter control will also bring the day under $2000 and 120* $m^3$ *of water."* Find the flaw and say how you would check it in the app.

CHAPTER 9

# A Building HVAC Simulator With a Comfort Band Instead of a Setpoint



**The controller moves** 1 supply fan and its speed drive 2 chilled water valve at the coil 3 four terminal dampers, one per room 4 outdoor air damper

**The controller reads** 1 room temperatures, one per room 2 supply air temperature 3 duct static pressure

**What disturbs it** 1 sun on the glass 2 outdoor air temperature

*One fan and one coil serve all four rooms while the sunlit room asks for the most air and the concrete floors hold the cooling bought early. Illustration, not to scale.*

IN PLAIN TERMS **Spending the Freedom in a Comfort Range**

**The main idea.** Office workers accept any room temperature inside a range, not one exact number. The concrete floors soak up cooling bought early and give it back later. So a controller can cool hard on cheap, cool mornings and coast through the costly afternoon.

**The challenge.** The concrete takes hours to cool, so the choice is made long before anyone can tell whether it was right. A planner that looks too short a time ahead cools in a rush. It then sets a new and costly peak in power use.

**Why it matters.** On this chapter's main summer test day, using the range cuts the bill by 8.8 percent while every room stays comfortable.

THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 9.2, the facts in the Introduction and in Section 9.6.6, the try-it segments in Section 9.6, and the key points |
| **Study** | 70 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 25 min | the first page, the results (Section 9.6), Technology in Practice (Section 9.7), Limitations (Section 9.8) and the open challenges |

**The control question** An office is judged against a comfort band, so the setpoint inside it is free and the concrete behind it stores cooling. How should a controller spend that freedom against a tariff, a forecast and a curtailment, and what does a day-ahead plan add to a priced rule?

**Learning objectives**

- Compute the cooling a building's fabric stores across a comfort band.
- Explain how the tariff and the chiller curve set the cost of cooling and the supervisor's charge demand.
- Explain why a horizon shorter than the time to charge the store costs more than none.
- Measure what notice of a curtailment and the comfort band are each worth on the same event.
- Compare a receding-horizon plan with the priced rule, against a retuned baseline.

**Prerequisites** Sections 1.6.2, 1.4.3, 1.4.5, 1.4.6, 1.4.8 and 1.6.4, and Chapter 8 for a supervisor that chooses a cooling plant's setpoints.

**The app** https://apmonitor.com/apps/building-hvac/

**See also** Volume 17, Chapter 10 for an inventory dispatched inside its working band, Volume 17, Chapter 6 for a setpoint chosen rather than defended, and Volume 17, Chapter 11 for storage run by an inventory policy under a forecast bias.



## 9.1 Introduction

Nothing on this air handler is difficult to hold. The duct static pressure reaches 63 % of an 80 Pa step in 0.55 min, the supply air temperature reaches 63 % of a one degree step in 1.00 min, and the west zone reaches 63 % of a half degree step in 11.6 min, all three under ordinary proportional-integral control. What is hard is deciding what the zone number ought to be, when anywhere between 21.0 and 24.0 °C is a pass, when the concrete behind that number holds 856 kWh of cooling nobody is using, and when the decision has to be taken five hours before anybody can tell whether it was right.

That gap is the misconception the application corrects. A control course teaches a loop to hold a number, and a student who has only ever seen a block diagram concludes that the number was handed down by somebody else. In a commercial building it was not. The comfort specification is an interval. Comfort research states it as a range rather than a point [1], and the lease that turns that research into an obligation is written the same way. The width of the interval is the cheapest storage most buildings will ever have. Reading a comfort band as a setpoint plus a tolerance throws it away. Reading it as an interval turns the setpoint into a manipulated variable for a slower loop, and the plant then has four loops rather than three.

The class of problem is control against a specification that is a range, with the remaining freedom spent on an economic objective, and it is not confined to buildings. Industrial model predictive controllers routinely control to a zone rather than to a setpoint and give the leftover degrees of freedom to a steady-state economic optimization, which is the standard commercial architecture rather than a special case [2]. A blend property held between specification limits, a product moisture held above a giveaway floor and below a quality ceiling, a surge drum level held anywhere between its alarms, and a battery state of charge held inside a warranty window all have this shape, because each has an inventory whose charge and discharge cost different amounts at different times and a tolerance that pays for the arbitrage. The building version is the largest by installed capacity, because the fabric of the existing stock is storage already built and already paid for [3]. Volume 17, Chapter 10 dispatches the battery version against a feeder's net-load curve, and Volume 17, Chapter 6 chooses a solar field's outlet setpoint rather than defending a fixed one.

**DID YOU KNOW? Buildings Draw Most of the Grid's Power**

Buildings account for more than 70 percent of United States electricity use and a comparable share of peak demand [4]. In United States office buildings, heating, cooling and ventilation make up 51 percent of total energy use [5]. The modeled office puts about 40 % of its electricity into the HVAC plant (Section 9.8).

Two measurements make the building case quantitative. Four zones of slab and partition at 250 000 kJ/K each, plus their air, hold 855.8 kWh across a 3 K band, and 28.5 kWh for every tenth of a degree the band is widened. And cooling bought in the morning is cheaper than cooling bought in the afternoon before the tariff is read at all, because an air-cooled chiller rejects heat into the afternoon it is cooling against, so the machine modeled here runs at a coefficient of performance of 5.43 at five in the morning and 3.10 at four. A kilowatt-hour of cooling on the design day therefore costs 1.4 cents at five in the morning and 10.4 cents at three in the afternoon, a factor of 7.5 that is the tariff's 4.3 times the chiller's 1.8.

The building control literature reaches this decision through optimization formulations a first course cannot run [6–8], and the interactive material available to a student is either a thermal model with no controller or a controller with no economics. This chapter works with an interactive office building and its air handler, given in full so that either can be rebuilt, and prices a comfort band against a weather forecast on the same plant. The plant has seventeen states, and its rest condition is solved by one routine that the initializer, the manual balancing aid and the self-test all call. The four loops above it are transcribed from the executing code. The two-node zone model, the coil effectiveness relation and the precooling idea are standard. The central measurement is that on a three-hour utility curtailment two hours of notice reduces the energy drawn above the cap from 19.4 to 16.1 kWh while removing the comfort band raises it to 152.3 kWh, and that eight hours of notice still leaves 5.7 kWh because a ventilation minimum of 4.17 $m^3/s$ puts a floor of 133 to 160 kW under the building. The graded challenge adds a smaller lesson: one of the four improvements it asks for turns out to be worth exactly zero dollars.

Section 9.2 traces how a building came to have a band to spend. Section 9.3 gives the seventeen-state model, its constants and its assumptions, and Section 9.4 its two time constants, its three process gains and the fan the zones share. Section 9.5 sets out the three regulatory loops, the supervisor that prices a kilowatt-hour of cooling, and the predictive controller of Section 9.5.4 that plans a day of cooling. Section 9.6 measures every day-long scenario: a band floated and a band used, the forecast horizon, a curtailment run three ways, a plant that runs short, the graded challenge, and the plan against the priced rule in Section 9.6.6. Section 9.7 describes the sensors, terminal boxes and controllers of real buildings. Sections 9.8 and 9.9 state what the model cannot settle and what the curtailment comparison means, and the exercises in Section 9.11 run the cases.

## 9.2 How a Building Came to Have a Band to Spend

The comfort band predates any decision to use it. A pneumatic room thermostat is a proportional device whose throttling range spans two or three degrees between a shut damper and an open one, so the first generation of automatic buildings held a range because the hardware could hold nothing narrower. Constant volume systems with terminal reheat met the worst zone by supplying cold air everywhere and burning it back where it was not wanted, and the cost of doing so is what ended the arrangement. Variable air volume replaced the reheat coil with a damper and made supply airflow the manipulated variable at the zone, which put a variable speed fan and a duct static pressure loop into a building that had previously had neither, and made resetting that pressure setpoint against the most open damper the standard efficiency measure on the resulting plant [9].

Direct digital control arrived next and quietly removed the band. A pneumatic thermostat has a throttling range and a digital one has a setpoint in a database, so the number a building is held to became a single value entered once at commissioning. The band survived in the specification and disappeared from the control system, which is the state most of the existing stock is still in.

Commercial buildings account for a large and well documented share of electricity use, and cooling is the part of it that peaks with the grid [10].

The first answers to the resulting peak were open loop. Night setback and optimum start recovered energy from the unoccupied hours, and demand limiting was duty cycling of equipment, which reduces the meter reading by taking cooling away from whichever zone is unlucky. Chilled water and ice tanks were the capital answer, and they work, at the cost of a plant room and a maintenance contract. The measurement that changed the argument was that the building itself is already such a tank. Preconditioning an office at night was shown to shift peak load by an amount set by the mass of the structure rather than by the size of the plant [11], and a sequence of studies then established what an optimal charge and discharge trajectory looks like, how much of the ideal a simple strategy recovers, and what the trajectory costs when the building is occupied throughout the discharge [12–14]. Demand limiting was reformulated on the same basis, as a model-based choice of a zone temperature trajectory rather than as equipment shedding [15], which is the form the strategy in Section 9.5 takes.

Three developments then made the strategy ordinary. Grey box models identified from a building's own operating data gave the resistance and capacitance parameters a controller needs and a design calculation never produces [16]. Weather forecasts reached the control system, and a controlled comparison on an occupied office building showed a predictive controller using them beating the rule-based sequence it replaced [6]. Utility programs made the economics explicit, because a demand charge and a curtailment penalty price a fifteen-minute average rather than a kilowatt-hour and that is what a thermal store arbitrages [17]. Predictive control of building cooling with storage then moved from simulation to demonstration on real plant [7], and the field is now reviewed as a mature subject with its own catalog of failure modes [8]. What remains unusual is not the algorithm. It is the willingness to write the comfort specification down as an interval and let a controller use it, because that is a term in a lease rather than a line in a control sequence.

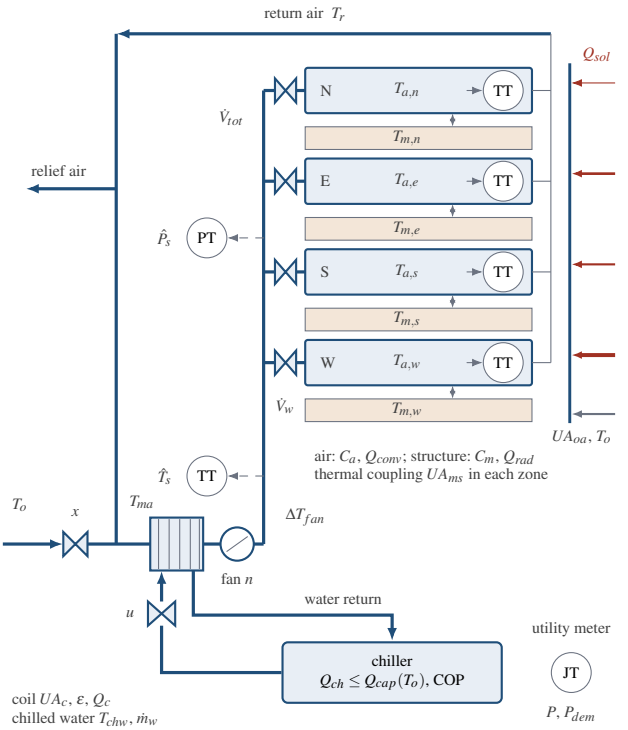



**Figure 9.1** The floor plate in section with its air handler. Four perimeter zones carry a fast air node $T_{a,z}$ of capacitance $C_a$ over a slow structural node $T_{m,z}$ of capacitance $C_m$, coupled by the surface conductance $UA_{ms}$, with convective gains $Q_{conv}$ reaching the air and radiant gains $Q_{rad}$ reaching the structure. Transmitted solar $Q_{sol}$ enters through glazed area $A_g$ and differs by orientation, shown by the weight of the four arrows at the design hour. Outdoor air enters at fraction $x$ and mixes to $T_{ma}$ ahead of the coil, and one coil, one fan and four pressure independent terminals serve all four zones with $\dot{V}_z = \min(\dot{V}_{z,\mathrm{SP}}, K_{box}\sqrt{\hat{P}_s})$. Three instruments close loops, the supply air thermistor $\hat{T}_s$, the duct static transmitter $\hat{P}_s$ and the zone sensors $\hat{T}_{a,z}$, and the utility meter reports the fifteen-minute demand $P_{dem}$.

> **HISTORY 2003: Precooling an Occupied Office**
>
> In October 2003 a national laboratory and a university tested precooling in an occupied office building of about 80,000 square feet with a medium-weight concrete structure. Its zones were normally held at a constant 72 °F. In eleven one-day tests the zone setpoints were held low through the morning, at 70 °F from 5 am, and then raised to 78 °F at 2 pm so that the rooms could float through the afternoon. From 2 to 5 pm the chiller power fell by 80 to 100 percent, without a thermal comfort complaint [18]. The report observed that conventional controls treat the building's thermal mass as an obstacle rather than an asset [18]. The supervisor of this chapter makes the same move with a price and a forecast in place of a clock.

## 9.3 Process Model

The building is a 5000 $\mathrm{m}^2$ office floor plate on four perimeter orientations, occupied by 417 people at design, with 40 kW of lighting and 50 kW of equipment at full output. One air handler serves all four zones through a chilled water coil on a modulating valve, a variable speed supply fan holding duct static pressure, four pressure independent variable air volume terminals and a dry-bulb economizer. An air-cooled chiller makes the water. Every quantity the equations below refer to is labeled on Figure 9.1.

**Table 9.1** Classification of Variables. Nominal Values Are Those of the Design Point, 15:00 on the Design Day, at Which the Bench Scenarios Start.

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $T_{a,z}$ (4) | CV, measured | 23.00 °C | 21.0–24.0 °C band |
| $T_s$ | CV, measured | 13.000 °C | 10–18 °C SP |
| $P_s$ | CV, measured | 300.0 Pa | 100–500 Pa SP |
| $E$ | CV, inferred | 285 kWh | 0–856 kWh |
| $\dot{V}_{z,\mathrm{SP}}$ (4) | MV | 3.80–6.25 m³/s | 0–9.0 m³/s |
| $u$ valve | MV | 52.5 % | 0–100 % |
| $n$ fan | MV | 86.2 % | 25–100 % |
| $T_{z,\mathrm{SP}}$ | MV of the supervisor | 23.0 °C | $T_{lo}$–$(T_{hi}-0.3)$ |
| $T_o$ | DV, measured | 35.0 °C | ±10 °C shift |
| $Q_{sol}$ | DV, measured | 6.5–36.0 kW | 0.15–1.15× |
| occupancy | DV, measured | 1× | 0.3–1.4× |
| $p(t)$ tariff | DV, known | 0.320 $/kWh | 0.5–3× |
| $P_{cap}$ | DV, announced | none | 0–400 kW |
| coil fouling | DV, unmeasured | 1× | 0.45–1.05× |
| filter loading | DV, unmeasured | 1× | 1–2.4× |
| chiller capacity | DV, unmeasured | 1× | 0.4–1.05× |
| $T_{a,z}$ (4) | state | 23.00 °C | 5–45 °C |
| $T_{m,z}$ (4) | state | 24.19–25.84 °C | 5–45 °C |
| $\dot{V}_z$ (4) | state | 3.80–6.25 m³/s | 0–9.0 m³/s |
| $T_s$ | state | 13.000 °C | 2–40 °C |
| $T_{chw}$ | state | 6.00 °C | 2–30 °C |
| $n$, $u$ | states | 86.2, 52.5 % | 0–100 % |
| $P_{dem}$ | state | 217.8 kW | 0–5000 kW |

Table 9.1 sorts the variables by the part they play. The clock runs in minutes, because a day is 1440 of them and the structural time constant is 333. Every duty is in kW, every capacitance in kJ/K, and the factor of 60 between them is applied in one place rather than scattered through the balances. Airflow is volumetric in m³/s throughout against one volumetric heat capacity $\rho c_p$, so no mass flow of air appears anywhere.

### 9.3.1 Two Thermal Masses per Zone

Each zone carries an air node and a structural node. The air node holds the air and the light furnishings, sees the supply air through the volumetric heat capacity of whatever flow its terminal is passing, and takes the convective share of the internal and solar gains, as shown in Equation (9.1).

$$\frac{C_a}{60}\frac{\mathrm{d}T_{a,z}}{\mathrm{d}t} = \dot{V}_z \rho c_p (T_s - T_{a,z}) + UA_{ms}(T_{m,z} - T_{a,z}) + UA_{oa}(T_o - T_{a,z}) + Q_{conv,z} \quad (9.1)$$

where $C_a$ is in kJ/K, $t$ is in minutes, every duty is in kW and the 60 converts between them. The structural node holds the slab, the columns and the partitions, exchanges with the air through the same surface conductance, and takes the radiant share of the gains, as shown in Equation (9.2).

$$\frac{C_m}{60}\frac{\mathrm{d}T_{m,z}}{\mathrm{d}t} = UA_{ms}(T_{a,z} - T_{m,z}) + Q_{rad,z} \quad (9.2)$$

Nothing but $UA_{ms}$ connects the structure to the rest of the plant, which is what makes the store slow and the only route into and out of it. Two consequences are used later. The structural node has one time constant, $C_m/(60\,UA_{ms}) = 333\,\mathrm{min}$, that no actuator changes, and at rest (9.2) gives $T_{m,z} = T_{a,z} + Q_{rad,z}/UA_{ms}$, so the slab always sits above its zone air by the radiant gains over the surface conductance, 2.84 K in the west zone at the design point and 1.19 K in the north.

Gains are split between the two nodes by fixed radiant fractions, as shown in Equation (9.3).

$$\begin{aligned} Q_{rad,z} &= f_{sol} q_{s,z} + f_p q_p + f_l q_l + f_e q_e \\ Q_{conv,z} &= (1-f_{sol}) q_{s,z} + (1-f_p) q_p \\ &\quad + (1-f_l) q_l + (1-f_e) q_e \end{aligned} \quad (9.3)$$

where $q_{s,z} = A_g\,\mathrm{SHGC}\,I_{v,z}(1-f_b)/1000$ is the transmitted solar in kW for glazed area $A_g$ in $\mathrm{m}^2$ and vertical irradiance $I_{v,z}$ in $\mathrm{W/m}^2$ with a blind fraction $f_b$, and $q_p$, $q_l$ and $q_e$ are the per-zone people, lighting and equipment gains, each the whole-building value times its schedule divided by four. Vertical irradiance comes from the solar geometry at latitude 38° and declination 20°, with beam, half the sky diffuse and half the ground-reflected component on each face. At the design point the four faces take 114, 114, 224 and 636 $\mathrm{W/m}^2$, which is 6.5, 6.5, 12.7 and 36.0 kW of transmitted gain, and 70 % of each goes straight into concrete.

### 9.3.2 Terminals, Fan and Duct

A pressure independent terminal is a flow controller with a damper, so it delivers what it is asked for until its damper is wide open and then delivers whatever the static pressure will pass, as shown in Equation (9.4).

$$\tau_{box}\frac{\mathrm{d}\dot{V}_z}{\mathrm{d}t} = \min\left(\dot{V}_{z,\mathrm{SP}},\, K_{box}\sqrt{P_s}\right) - \dot{V}_z \tag{9.4}$$

The fan curve and the duct upstream of the pressure sensor are collapsed into one quadratic, as shown in Equation (9.5).

$$P_s = \max\left(0,\; P_0 s^2 - (R_f + R_{up} f_{filt})\dot{V}_{tot}^2\right), \qquad s = \frac{n}{100} \tag{9.5}$$

where $P_0$ is the shut-off pressure at full speed, $R_f$ is the fan curve droop, $R_{up}$ is the resistance of filter, coil and main duct, $f_{filt}$ is a filter loading multiplier and $\dot{V}_{tot} = \sum_z \dot{V}_z$. Duct pressure dynamics are seconds rather than minutes, so $P_s$ is algebraic and every dynamic in the pressure loop comes from the drive ramp and the dampers. Fan power and the fan heat that follows it are given by Equation (9.6).

$$P_{fan} = \frac{\dot{V}_{tot}\max(0, P_0 s^2 - R_f \dot{V}_{tot}^2)}{1000\,\eta_{fan}}, \qquad \Delta T_{fan} = \frac{P_{fan}}{\dot{V}_{tot}\rho c_p} \tag{9.6}$$

The motor sits in the airstream, so every watt of fan power arrives at the terminals as a temperature rise the coil has to remove again. At the design point that is 1.012 K for 22.19 kW of fan into 18.18 $\mathrm{m}^3/\mathrm{s}$ of air.

### 9.3.3 Mixing, Coil and Chiller

Return air is the flow-weighted mixture of the four zones, $T_r = \sum_z \dot{V}_z T_{a,z}/\dot{V}_{tot}$. Outdoor air enters at fraction $x$, which is the ventilation minimum unless the economizer opens it further, as shown in Equation (9.7).

$$\begin{aligned} \dot{V}_{oa} &= \dot{V}_{oa,des}\,(0.25 + 0.75\,\phi_{occ}), \qquad x_{min} = \frac{\dot{V}_{oa}}{\dot{V}_{tot}} \\ x &= \begin{cases} \mathrm{clamp}\left(\dfrac{T_r - T_c^*}{\max(0.5,\, T_r - T_o)}, x_{min}, 1\right) & \text{eco} \\ x_{min} & \text{otherwise} \end{cases} \end{aligned} \tag{9.7}$$

where $\phi_{occ}$ is the occupied fraction, $T_c^* = T_{s,\mathrm{SP}} - \Delta T_{fan}$ is the coil-leaving target, and the economizer branch is taken only when the economizer is enabled, the outdoor air is at least 0.5 K below the return, and $T_o$ is below the dry-bulb high limit of 18 °C. Opening only as far as the coil-leaving target requires keeps the coil from being asked to reheat air the damper over-cooled. The mixed air is $T_{ma} = xT_o + (1-x)T_r$, and $x_{min}$ is floored at 5 % and clamped to 1.

The coil is a counter-flow effectiveness model whose conductance scales with both flows, as shown in Equation (9.8).

$$\begin{aligned} UA_c &= UA_{c,0}\left(\frac{\dot{V}_{tot}}{\dot{V}_0}\right)^{0.5}\left(\frac{\dot{m}_w}{\dot{m}_{w,ref}}\right)^{0.8} f_{foul} \\ \mathrm{NTU} &= \frac{UA_c}{C_{min}}, \qquad C_r = \frac{C_{min}}{C_{max}} \\ \varepsilon &= \frac{1-\mathrm{e}^{-\mathrm{NTU}(1-C_r)}}{1-C_r\mathrm{e}^{-\mathrm{NTU}(1-C_r)}} \\ Q_c &= \max\left(0,\ \varepsilon\, C_{min}\,(T_{ma}-T_{chw})\right) \\ T_s^{\dagger} &= T_{ma} - \frac{Q_c}{\dot{V}_{tot}\rho c_p} + \Delta T_{fan} \end{aligned} \tag{9.8}$$

where $C_{min}$ and $C_{max}$ order the air capacity rate $\dot{V}_{tot}\rho c_p$ and the water capacity rate $\dot{m}_w c_{pw}$, with $\dot{m}_w = \dot{m}_{w,\max}u/100$, and $T_s^{\dagger}$ is the supply air the coil is currently making. The removable singularity at $C_r \to 1$ uses $\varepsilon = \mathrm{NTU}/(1+\mathrm{NTU})$. Clipping $Q_c$ at zero matters, because a cooling coil cannot heat. At the design point the coil is 69.7 % effective and removes 301.7 kW with 9.18 kg/s of water through a valve 52.5 % open.

Supply air, chilled water and the two actuators are first-order lags, as shown in Equation (9.9).

$$\begin{aligned} \tau_{sup}\frac{\mathrm{d}T_s}{\mathrm{d}t} &= T_s^{\dagger} - T_s, & \frac{C_w}{60}\frac{\mathrm{d}T_{chw}}{\mathrm{d}t} &= Q_c - Q_{ch} \\ \tau_{fan}\frac{\mathrm{d}n}{\mathrm{d}t} &= n_{cmd} - n, & \tau_{val}\frac{\mathrm{d}u}{\mathrm{d}t} &= u_{cmd} - u \\ \tau_{dem}\frac{\mathrm{d}P_{dem}}{\mathrm{d}t} &= P - P_{dem} \end{aligned} \tag{9.9}$$

The chilled water state is the one that reports a short plant. The machine takes the coil duty plus whatever is needed to pull the loop back to setpoint, up to its capacity, as shown in Equation (9.10).

$$\begin{aligned} Q_{want} &= Q_c + \frac{C_w}{60}\frac{T_{chw}-T_{chw,\mathrm{SP}}}{\tau_{chw}} \\ Q_{ch} &= \mathrm{clamp}(Q_{want},\,0,\,Q_{cap}), \quad \mathrm{PLR} = \frac{Q_{ch}}{\max(1, Q_{cap})} \\ Q_{cap} &= \max\left(20,\ Q_{cap,0}\,(1-\beta_{cap}(T_o-30))\right) f_{chl} \end{aligned} \tag{9.10}$$

Above capacity the chilled water temperature walks up rather than the coil duty being clamped, which is what a short plant does and what takes authority away from every valve in the building at the same instant. The coefficient of performance is a fixed fraction of Carnot with a condenser approach that grows with load, as shown in Equation (9.11).

$$\begin{aligned} T_{cond} &= T_o + \Delta T_{c,a} + \Delta T_{c,b}\,\mathrm{PLR} \\ T_{evap} &= T_{chw} - \Delta T_{evap} \\ \mathrm{COP} &= \eta_C \frac{T_{evap}+273.15}{\max(3,\, T_{cond}-T_{evap})} f(\mathrm{PLR}) \\ f(\mathrm{PLR}) &= \mathrm{clamp}\big(1+0.28\nu-0.55\nu^2,\,0.55,\,1.12\big) \end{aligned} \tag{9.11}$$

with $\nu = 1-\mathrm{PLR}$. Equation (9.11) contains half of the case for precooling, because $T_{cond}$ follows the outdoor air. On the design day it gives 5.43 at five in the morning, 4.70 at eight, 3.94 at ten, 3.39

at noon and 3.10 at four in the afternoon, a loss of more than two fifths of the machine's efficiency to nothing but the time of day. Electrical demand is the sum in Equation (9.12).

$$P = \underbrace{\frac{Q_{ch}}{\mathrm{COP}} + P_{pmp,0}\left(\frac{\dot{m}_w}{\dot{m}_{w,\max}}\right)^3 + P_{pmp,b} + P_{fan}}_{\text{HVAC}} + \underbrace{P_{lgt} + P_{plg} + P_{base}}_{\text{house load}} \tag{9.12}$$

The chiller and pump terms are taken as zero when $Q_{ch} \le 2\,\mathrm{kW}$. At the design point the four terms are 98.5, 3.6, 22.2 and 93.5 kW for 217.8 kW at the meter, and the 93.5 kW of house load is the part no control decision can move.

### 9.3.4 What Is Stored, and What the Day Costs

Stored cooling is measured against the building this building would be if it were floating at the top of its comfort band, as shown in Equation (9.13).

$$E = \sum_z \frac{C_m\left(T_{hi} + \dfrac{Q_{rad,z}}{UA_{ms}} - T_{m,z}\right) + C_a\left(T_{hi} - T_{a,z}\right)}{3600}$$
$$E_{\max} = \frac{4(C_m + C_a)(T_{hi} - T_{lo})}{3600} \tag{9.13}$$

The structural reference is the band top plus the radiant gains over the surface conductance, not the band top itself, because (9.2) puts the slab above the air at rest by exactly that amount. Referencing it to the air ceiling instead reports a well-run building as 468 kWh in debt, and a supervisor asked to hold zero charge would then cool hard all day chasing a state that does not exist. At the design point $E = 285\,\mathrm{kWh}$ of $E_{\max} = 855.8\,\mathrm{kWh}$, and $E_{\max}$ grows by 28.5 kWh for every tenth of a degree of band width.

The day is billed by Equation (9.14).

$$\mathrm{Bill} = \int p\,P\,\mathrm{d}t + r_d \max_t P_{dem} + c_{DR}\int_{event} (P - P_{cap})_+\,\mathrm{d}t \tag{9.14}$$

where $p(t)$ is the time-of-use energy price, $r_d$ is the demand charge on the highest fifteen-minute average the meter saw, and $c_{DR}$ is the penalty on energy drawn above a curtailment cap while an event is running. The middle term is the one a fixed setpoint cannot touch. On the fixed-setpoint design day it is \$182 against \$434 of energy, set by one quarter of an hour in the whole month.

### 9.3.5 Weather, Schedules and the Tariff

The outdoor dry bulb is two half-cosines rather than one sinusoid, with the minimum at 05:00 and the maximum at 15:00, as shown in Equation (9.15).

$$T_o(c) = \begin{cases} T_{lo}^o + \dfrac{\Delta T_o}{2}\left[1 - \cos\dfrac{\pi\,u_r}{\Delta c_r}\right], & c_{min} \le c \le c_{max} \\ T_{hi}^o - \dfrac{\Delta T_o}{2}\left[1 - \cos\dfrac{\pi\,u_f}{\Delta c_f}\right], & \text{otherwise} \end{cases} \tag{9.15}$$

where $c$ is the clock minute, $\Delta T_o = T_{hi}^o - T_{lo}^o$, the rising leg runs over $u_r = c - c_{min}$ and $\Delta c_r = c_{max} - c_{min}$, and the falling leg over $u_f = (c - c_{max}) \bmod 1440$ and $\Delta c_f = 1440 - \Delta c_r$. A single sinusoid would put the minimum twelve hours after the maximum and make the morning warm-up as slow as the evening cool-down, and the shape matters because the precooling

window is exactly the warming half. Occupancy, lighting and equipment follow smoothstep windows $w(c;a,b,e,f) = \min\left(S\left(\frac{c-a}{b-a}\right), 1 - S\left(\frac{c-e}{f-e}\right)\right)$ with $S(\xi) = \xi^2(3-2\xi)$ clamped to $[0,1]$, so every gain term is continuously differentiable and the fixed-step integrator sees no kink. Occupancy is $0.03 + 0.97\,w(c;400,540,1000,1140)$ reduced by a quarter over a lunch window $w(c;700,730,790,820)$, lighting is $0.08 + 0.87\,w(c;380,440,1080,1200)$ and equipment is $0.35 + 0.60\,w(c;400,500,1020,1140)$. The building counts as occupied when occupancy exceeds 0.5, and the band tightens $t_{lead}$ minutes before that, so a building floating to its 27 °C night setback is not still pulling down at half past nine.

Energy costs 7.5 cents a kilowatt-hour from 22:00 to 08:00, 11.0 cents from 08:00 to 14:00 and from 19:00 to 22:00, and 32.0 cents from 14:00 to 19:00. A curtailment is described by a cap $P_{cap}$, a start, an end, a notice period and a penalty rate. Three windows follow from those and they are deliberately not the same window. The event is active over $[c_{start}, c_{end})$, it is known to the controller only from $c_{start} - t_{notice}$, and it is priced as an event from $c_{start} - t_{DR}$ so that the plant has unloaded by the time the cap takes effect. Without that last lead the precool is still running flat out at the instant the cap bites and the building fails the first minutes of every event by having prepared for it too well.

### 9.3.6 Constants, Starting State and Numerics

Table 9.2 carries the constants. Twenty-two of the eighty-three never appear on screen, and those twenty-two are what a learner has to infer from a measurement rather than read off a panel. Post-step clipping holds the zone and structural temperatures between 5 and 45 °C, the terminal flows between 0 and $1.6\dot{V}_0$, the supply air between 2 and 40 °C, the chilled water between 2 and 30 °C and both actuators between 0 and 100 %.

Every scenario opens from a solved steady state. One routine evaluates the plant at rest for a given clock and a given set of actuator positions, and the initializer, the manual balancing aid and the verification suite all use it, so the design-point test and the operating point a student loads are the same point rather than two that look alike. It eliminates the structure with $T_{m,z} = T_{a,z} + Q_{rad,z}/UA_{ms}$, solves the terminal flows and duct static as a contracting fixed point, and closes the supply air loop by successive substitution whose loop gain is about 0.23, one minus the coil effectiveness times the recirculated fraction. Twelve day-long scenarios override the result with a measured overnight state, because a building whose plant has been off since six in the evening is at 23.3 to 23.9 °C of air over 23.9 to 24.7 °C of slab at four in the morning, and not at the free-float 28 °C it would reach if the sun never came up.

The engine of Section 1.6.1 advances the states under fourth-order Runge-Kutta at a frozen 0.05 min (3 s) step. Control executes on a 0.25 min (15 s) period, which is a building controller scan rather than a process control scan, and four samples of every logged column are kept per simulated minute. The sensor noise comes from a generator seeded at 20260914. The speeds are 1, 5, 20, 60 and 240 simulated minutes per second, and the twelve day-long scenarios need the top two.

### 9.3.7 Assumptions

Detail is spent where the behavior forces a controller design decision and withheld everywhere else. Each of the simplifications below fails somewhere.

*Sensible heat only.* There is no moisture anywhere in the model, so there is no latent coil load, no dehumidification, no reheat, and the economizer changeover is on dry bulb rather than on enthalpy. In a humid climate the latent load is often a third of the coil duty and a dry-bulb economizer can cost money rather than save it, so the coil sizing in (9.8) and the 18 °C limit in (9.7) belong to a dry climate and nowhere else.

*Two thermal nodes per zone.* A real slab has a temperature gradient through its depth and answers a step roughly as the square root of time rather than as the single exponential of (9.2). Two-node lumped models are the standard identifiable choice for control work and are adequate over a few

**Table 9.2** Model Parameters. Values Are the Shipped Defaults. Hidden Constants Are Never Displayed in the Application and Are the Quantities a Student Has to Infer. Every Other Parameter Is Shown and Adjustable in the Model Panel.

| Symbol | Meaning | Value | Unit | Visibility |
|---|---|---|---|---|
| *Building and gains* | | | | |
| $A$ | conditioned floor area | 5000 | $m^2$ | shown |
| $C_a$ | zone air capacitance | 6750 | kJ/K | shown |
| $C_m$ | zone structural capacitance | 250 000 | kJ/K | shown |
| $UA_{ms}$ | structure-to-air conductance | 12.5 | kW/K | shown |
| $UA_{oa}$ | envelope plus infiltration, per zone | 0.85 | kW/K | shown |
| $A_g$ | glazed area per zone | 177 | $m^2$ | shown |
| SHGC | solar heat gain coefficient | 0.32 | — | shown |
| — | design occupancy | 417 | people | shown |
| — | sensible heat per person | 0.075 | kW | shown |
| — | lighting at full output | 40 | kW | shown |
| — | equipment at full output | 50 | kW | shown |
| $P_{base}$ | lifts, hot water, other house load | 8 | kW | shown |
| $f_{sol}$ | solar fraction reaching the structure | 0.70 | — | shown |
| $f_p, f_l, f_e$ | radiant fractions, people/light/plug | 0.45, 0.40, 0.25 | — | hidden |
| *Weather* | | | | |
| $T^o_{lo}, T^o_{hi}$ | outdoor minimum and maximum | 17, 35 | °C | shown |
| $c_{min}, c_{max}$ | clock minute of each | 300, 900 | min | hidden |
| $\varphi, \delta$ | latitude and solar declination | 38, 20 | ° | shown |
| $I_{dn,0}$ | extinction-free beam normal | 950 | $W/m^2$ | hidden |
| $\tau_{atm}$ | clear-sky atmospheric extinction | 0.14 | — | hidden |
| $I_{dh,0}$ | horizontal diffuse at the zenith | 110 | $W/m^2$ | hidden |
| $\rho_g$ | ground reflectance | 0.20 | — | hidden |
| $c_0$ | clock minute the run starts | 240 | min | shown |
| — | freeze the clock, bench mode | 0 | — | hidden |
| *Air side* | | | | |
| $\dot{V}_0$ | design supply air flow | 22.5 | $m^3/s$ | shown |
| $\rho c_p$ | volumetric heat capacity of air | 1.206 | kJ/($m^3$ K) | hidden |
| $P_0$ | fan shut-off pressure at full speed | 1150 | Pa | shown |
| $R_f$ | fan curve droop coefficient | 0.296 | Pa $s^2/m^6$ | hidden |
| $R_{up}$ | resistance upstream of the sensor | 1.383 | Pa $s^2/m^6$ | shown |
| $\eta_{fan}$ | fan and drive total efficiency | 0.62 | — | shown |
| $K_{box}$ | terminal coefficient, damper open | 0.433 | $m^3/(s\ Pa^{0.5})$ | shown |
| — | terminal minimum, maximum flow | 0.30, 1.60 | of zone design | shown |
| $\dot{V}_{oa,des}$ | outdoor air at design occupancy | 4.17 | $m^3/s$ | shown |
| — | economizer dry-bulb high limit | 18 | °C | shown |

| Symbol | Meaning | Value | Unit | Visibility |
|---|---|---|---|---|
| *Coil and plant* | | | | |
| $UA_{c,0}$ | coil conductance at design flows | 46.4 | kW/K | shown |
| $\dot{m}_{w,\max}$ | chilled water flow, valve open | 17.5 | kg/s | shown |
| $\dot{m}_{w,ref}$ | flow the conductance is quoted at | 11.4 | kg/s | hidden |
| $c_{pw}$ | water heat capacity | 4.18 | kJ/(kg K) | hidden |
| — | coil exponents on air, water flow | 0.50, 0.80 | — | hidden |
| $Q_{cap,0}$ | chiller capacity at 30 °C | 420 | kW | shown |
| $\beta_{cap}$ | capacity lost per K above 30 °C | 0.0075 | 1/K | shown |
| $\eta_C$ | fraction of Carnot achieved | 0.45 | — | shown |
| $\Delta T_{c,a}$ | condensing approach at zero load | 4.2 | K | shown |
| $\Delta T_{c,b}$ | extra approach at full load | 7.8 | K | shown |
| $\Delta T_{evap}$ | evaporating below chilled water | 3.0 | K | shown |
| $C_w$ | chilled water loop capacitance | 12 540 | kJ/K | shown |
| $P_{pmp,0}$ | pump at full flow | 11 | kW | shown |
| $P_{pmp,b}$ | pump and controls standing load | 2 | kW | shown |
| *Dynamics and instruments* | | | | |
| $\tau_{box}$ | terminal damper travel | 0.35 | min | shown |
| $\tau_{fan}$ | supply fan drive ramp | 0.50 | min | shown |
| $\tau_{val}$ | chilled water valve travel | 0.50 | min | shown |
| $\tau_{sup}$ | coil and supply duct transport | 0.50 | min | shown |
| $\tau_{chw}$ | chiller capacity control response | 2.0 | min | shown |
| $\tau_{dem}$ | utility demand meter window | 15.0 | min | shown |
| $\tau_{T_z}$ | zone temperature sensor | 1.0 | min | shown |
| $\tau_{T_s}$ | supply air thermistor | 0.30 | min | shown |
| $\tau_{P_s}$ | duct static transmitter | 0.08 | min | shown |
| *Tariff, curtailment and schedules* | | | | |
| — | off, mid, on-peak energy price | 0.075, 0.110, 0.320 | $/kWh | shown |
| — | on-peak window | 840–1140 | min | hidden |
| $r_d$ | demand charge | 0.85 | $/kW day | shown |
| $c_{DR}$ | penalty above a curtailment cap | 1.20 | $/kWh | shown |
| $t_{notice}$ | notice the utility gives | 120 | min | shown |
| $t_{DR}$ | charging stops before the event | 30 | min | shown |
| $w_{DR}$ | share of penalty the supervisor prices | 0.50 | — | shown |
| — | occupancy schedule breakpoints | 400, 540, 1000, 1140 | min | hidden |
| $t_{lead}$ | optimum start, band tightens early | 60 | min | shown |
| $T_{set}$ | unoccupied upper limit | 27.0 | °C | shown |
| $\Delta_{sp}$ | setpoint margin below the ceiling | 0.30 | °C | shown |

hours [16], but this one reaches the last part of the store faster than a real building would, so every

storage figure reported here is optimistic by that amount.

*Four perimeter zones and no core.* A real 5000 $m^2$ floor plate has an interior zone with no solar gain and no envelope, which is easier to hold and would dilute the orientation spread the application exists to show.

*The plant is reduced.* The chiller of (9.11) is a Carnot fraction with a load curve rather than a compressor map, with no cycling at low load, no minimum unloading step, no staging and no condenser fouling, and below about ten percent load a real machine cycles and loses more efficiency than the curve shows. Equation (9.5) is a fan curve and a system curve collapsed into one quadratic, accurate near the design duty and increasingly approximate away from it, with no surge, no stall and no minimum flow. The demand meter in (9.9) is one first-order state where a utility meter is a fifteen-minute window whose boundaries are fixed by the clock, so a real peak that straddles two boundaries reads lower than this model shows.

*The tariff is known exactly and the forecast errs only in level.* The only weather forecast error available is a constant bias. Real precooling controllers spend most of their engineering effort on the forecast, and the interesting failures are the ones where it is wrong in shape. This model shows only the other kind, and Section 9.6 shows it to be worth about a dollar a day.

*Occupant behavior is a schedule.* Nobody opens a blind, complains at 24.1 °C or brings in a heater, and the comfort score counts zone minutes inside a band rather than anything a person would recognize as comfort. Adaptive comfort research puts the acceptable range wider and makes it depend on recent outdoor conditions [1], which would change the width of the store without changing what a controller does with it.

## 9.4 Dynamic Character

### 9.4.1 Two Time Constants in the Same Room

Timing one node at a time with the other pinned is what each time constant means, and the application measures both from its own derivative function rather than quoting them from the parameters. Stepping the supply air from 13 to 10 °C with the terminal flows and the structural nodes held, the west zone air reaches 63 % of its move in 5.45 min. Holding the zone air 3 K below where it was and releasing the structure, the west slab reaches 63 % in 335 min. The ratio is 61. Both agree with what (9.1) and (9.2) predict at the design duty, where the west terminal passes 6.25 $m^3/s$ for an airflow conductance of 7.54 kW/K against 12.5 kW/K of surface and 0.85 kW/K of envelope, giving 5.39 min, and where $C_m/(60\,UA_{ms})$ gives 333 min.

The coupled pair has neither of those as an eigenvalue. Linearizing the two equations together at the design duty gives modes of 5.33 min and 838 min, and the slow one is the least useful of the three because it describes the building giving up rather than the store charging. What the separation means operationally is one experiment. With the terminals pinned and the supply air dropped three degrees, the west zone air falls 0.75 K in the first ten minutes and 1.02 K in twenty and then very nearly stops, because it has run into the slab, and twelve hours later the air is at 21.0 °C while the west slab has come from 25.84 to 24.30 °C. The first ten minutes are a temperature and are gone the moment the plant stops. Everything after is energy in storage and is still there afterward. The store cannot be filled in a hurry, and every disappointing result in Section 9.6 is that sentence being ignored.

### 9.4.2 Three Process Gains and a Shared Actuator

The three regulatory gains are measured at the design point. More valve is colder supply air at −0.123 °C/%. More fan speed is more duct static at +20.5 Pa/%, and it is the only one of the four loops with a positive process gain. More airflow into a zone makes it colder, and that gain has no single value because it depends on how long the measurement waits. With the supply air and duct static loops in service and the west terminal stepped from 6.25 to 7.35 $m^3/s$, it is −0.54 °C per $m^3/s$ twenty minutes after the step, −0.59 at an hour while the slab has barely moved, and −1.24 once the slab has caught up, which takes the best part of two days. A loop tuned against the third number is

2.3 times too slow for the load it has to reject. With the valve and fan also held by hand the supply air warms as the flow rises, and the same kind of step reads only −0.27 at twenty minutes.

What the three share is one fan and one duct, and the sharing is what fails. Asking all four zones for 22.5 °C at the design hour works cleanly, and every zone lands on 22.49 °C with the fan peaking at 95.7 %. Asking all four for 22.0 °C does not. The terminals open together, total flow reaches 23.2 m³/s, the fan saturates and duct static falls to 168 Pa. North, east and south settle on 21.98 °C within two hours while the west zone is still at 22.57 °C at two hours and 22.15 °C at five. A shared resource does not run out equally, and the zone that runs out first is the one asking for the most, which here is the west terminal already passing 111.2 % of its 5.625 m³/s design flow against the north terminal's 67.6 %.

The same failure without the ambiguity is every terminal asked for 115 % at once. The fan reaches 100 %, static settles at 191 Pa rather than 300, each terminal passes 5.98 m³/s of the 6.47 m³/s it asked for, and 37.8 kW of fan is buying 23.91 m³/s. Lowering the static setpoint changes nothing, because the loop is already saturated in the direction that would need more pressure. Bringing the four demands back to 90 % holds 300 Pa at 92.7 % speed on 28.3 kW, a quarter less fan power for 15 % less air (20.25 against 23.91 m³/s), which is the argument for static pressure reset [9] in one comparison. Chapter 10 closes the same kind of reset as a loop, from the most-open air valve to the header pressure setpoint.

## 9.5 Control Structure as Implemented

### 9.5.1 The Shared PID Algorithm

One algorithm serves all four loops. It is the velocity form of Section 1.6.2, written here in this chapter's symbols in Equations (9.16) and (9.17).

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k), \quad P_k = K_c e_k, \quad D_k = -K_c \tau_D\, a\, \dot{\mathrm{PV}}^{f}_{k} \tag{9.16}$$

$$\begin{aligned} \Sigma_k &= \Sigma_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \\ u^{uns}_k &= u_{bias} + \Sigma_k + I_{k-1} \\ u_k &= \mathrm{clamp}(u^{uns}_k, u_{\min}, u_{\max}) \\ I_k &= I_{k-1} + \left[ \frac{K_c}{\tau_I} e_k + \frac{u_k - u^{uns}_k}{T_t} \right] \Delta t_c \end{aligned} \tag{9.17}$$

where $a = \pm 1$ is the action flag, $\Delta t_c = 0.25$ min is the scan period, $\Sigma_k$ is the running sum $S_k$ of Equation (1.15), $u^{uns}_k$ is the output before limiting ($v_k$ of Equation (1.16)) and the tracking constant $T_t$ is $\tau_I$ for all four loops. None of the four loops sets a rate in the shared algorithm, so $u_k$ is the clamped value. Section 1.6.2 explains why the sum is never clamped and why the integral alone is wound back at a limit. Under manual control $I$ is set so that $u_{bias} + \Sigma_k + I_k$ equals the operator's actuator position exactly, which makes the transfer bumpless in both directions. All four loops run with $\tau_D = 0$, so (9.16) contributes no derivative term and the derivative filter is inactive. The one rate limit in the application is applied outside the controller and in one direction only, for the reason given in Section 9.5.3.

### 9.5.2 The Three Regulatory Loops

Table 9.3 lists the four loops as they execute. The three regulatory loops are paired in the obvious way, and the signs are worth stating because three of the four have a negative process gain and one does not. The zone, supply air and storage loops are direct acting, $a = -1$, because more of the manipulated variable makes the measurement smaller. The duct static loop is reverse acting, $a = +1$.

Manual operation hands the operator the chilled water valve, the supply fan and all four terminal flows, with the setpoint sliders inert, so the three process gains can be measured one actuator at a time. The zone architecture setting is separate from the mode, and selecting Manual there in Auto leaves

**Table 9.3** The Four Loops as Executed. All Four Run the Algorithm of (9.16) and (9.17) With $\tau_D = 0$ at a 0.25 min Scan.

| Loop | CV → MV | $K_c$ | $\tau_I$ | Output limits |
|---|---|---|---|---|
| Duct static | $\hat{P}_s \to n$ | 0.045 | 0.60 min | 25–100 % |
| Supply air | $\hat{T}_s \to u$ | 8.0 | 2.5 min | 0–100 % |
| Zone (4) | $\hat{T}_{a,z} \to \dot{V}_{z,\mathrm{SP}}$ | 1.30 | 12 min | $\dot{V}_{z,\min}$–9.0 m³/s |
| Storage | $E/E_{\max} \to T_{z,\mathrm{SP}}$ | 0.045 | 60 min | $T_{lo}$–$(T_{hi} - \Delta_{sp})$ |

Units: $K_c$ in %/Pa, %/°C, m³/(s °C) and °C/%
Action: $a = +1$ on duct static, $a = -1$ on the other three

the four flow stations to the operator while the air handler keeps holding supply air temperature and duct static, which is the commissioning case rather than a mode.

Three details of the regulatory layer are not generic. The zone loop's lower output limit is the ventilation minimum $\dot{V}_{z,\min} = 0.30\dot{V}_0/4 = 1.69\,\mathrm{m}^3/\mathrm{s}$ while the building is occupied and zero when it is not, so a zone controller with nothing to do still passes outdoor air. The air handler runs whenever the building is occupied or expected to be within $t_{lead}$, or whenever any zone measurement is more than 0.2 °C above what the controller asks of it, and that second clause is what lets the supervisor start the plant at four in the morning without a schedule, with both air-side loops tracking zero while it is off. The duct static measurement comes from a consistent evaluation of (9.5) rather than from a state, because the pressure is algebraic, and its filter is seeded from the plant rather than from zero, because the priming scan runs at a full 15 s step and a 0.08 min transmitter reaches 96 % of the real pressure in one call, so every automatic scenario would otherwise step its fan on the first scan.

### 9.5.3 The Supervisory Loop

The supervisor decides where inside the comfort band the zones should sit. The band itself is not constant. Its floor is always $T_{lo}$, and its ceiling is $T_{hi}$ while the building is occupied or about to be, $T_{DR}$ during a curtailment, and the night setback $T_{set}$ otherwise. Cooling below $T_{lo}$ is not storage, because it is money spent filling a tank whose extra volume the occupants would reject the moment they arrived.

What the supervisor compares is the marginal cost of a kilowatt-hour of cooling rather than of electricity, as shown in Equation (9.18).

$$c(t+\tau) = \frac{p(t+\tau) + w_{DR} c_{DR} \mathbb{1}[\text{priced event}]}{\max\left(0.5,\ \mathrm{COP}\left(T_o(t+\tau) + b_f,\ T_{chw,\mathrm{SP}},\ 0.7\right)\right)} \tag{9.18}$$

where $b_f$ is the forecast bias, the part load ratio is fixed at 0.7 because the question is what the machine will be like in three hours rather than what it is doing this second, and the indicator is one only when the notice has arrived and the evaluated hour lies inside the priced event window. That single term is the entire difference between a controller that answers a curtailment and one that does not. Three models are in play at this point and they are not the same model. The seventeen-state plant of Section 9.3 is what the building does. The gains and time constants of Section 9.4 are what a loop is tuned against. And (9.18) is everything the supervisor itself knows, a price and a chiller curve with no thermal capacitance in it anywhere, which is why the storage loop has to close around a measured state of charge rather than predict one. The decision follows from the ratio of the worst cost inside the horizon to the cost now, as shown in Equation (9.19).

$$\begin{aligned} c_{worst} &= \max_{\tau = 0, 15, \ldots, H} c(t+\tau) \\ \sigma &= \mathrm{clamp}\left( K_{pre} \left( \frac{c_{worst}}{c_{now}} - 1 \right), 0, 1 \right) \end{aligned} \tag{9.19}$$

The charge demand $\sigma$ is the setpoint of the storage loop as a fraction of $E_{\max}$, and the storage PI controller in Table 9.3 drives the zone setpoint until the measured $E/E_{\max}$ from (9.13) matches it.

Writing the excess of the ratio over one rather than the fractional saving $1 - c_{now}/c_{worst}$ matters for one case and it is the important one, because the fractional saving is bounded above by $K_{pre}$, so a curtailment penalty making the afternoon eight times dearer would commit barely more of the band than a tariff peak making it three times dearer and the event would get the same half-hearted precool as a Tuesday.

Two things make this a controller rather than a schedule. It responds to the price and the weather in front of it, so abolishing the on-peak rate changes what it does with no rule rewritten. And it discharges on its own, because the moment $c_{now}$ catches up with $c_{worst}$ the ratio collapses and $\sigma$ goes to zero. When $\sigma \leq 0.02$ the loop is bypassed, the setpoint is placed at $T_{ceil} - \Delta_{sp}$, where $T_{ceil}$ is whichever of $T_{hi}$, $T_{DR}$ and $T_{set}$ is the ceiling in force, and the controller tracks, so the building floats rather than regulating against a target of zero.

The one rate limit in the application is applied to the supervisor's output, outside the controller, and downward only, as shown in Equation (9.20).

$$T_{z,\mathrm{SP}} = \mathrm{clamp}\Big( \max\Big( T^{-}_{z,\mathrm{SP}} - \tfrac{r\Delta t_c}{60}, u_{sup} \Big), \; T_{lo}, T_{ceil} - \Delta_{sp} \Big) \tag{9.20}$$

where $u_{sup}$ is the PI output, $T^{-}_{z,\mathrm{SP}}$ is the previous setpoint and $r = 1.0\,^{\circ}\mathrm{C/h}$ is the ramp limit, with $I$ back-calculated by the whole amount the pacing removed, in one scan, so the integral carries what was applied. The limit is one-sided on purpose. Charging has to be paced, because the charge demand steps the moment the horizon reaches the peak window and an unrestrained PI answers that step by taking every zone to the floor of the band at once, which opens four terminals together and sets a new demand peak two hours before the one it was avoiding. Releasing is ceasing to buy and must be instant. A symmetric limit of the same rate inside the controller would get this wrong twice. It would take 2.7 hours to let a building precooled to the 21.0 °C floor back up to its 23.7 °C ceiling, so the plant would keep cooling through most of a three-hour event it had prepared for. It would also stop the setpoint following the band, so the drop from the 26.7 °C night ceiling to 23.7 °C when the band tightens at 06:49 would not be complete until 09:49, two hours after the building is occupied.

### 9.5.4 Model Predictive Control

The zone control has a fourth setting, which no shipped scenario carries as its default and only scenario 15 pins, in which a model predictive controller (MPC) writes the four zone setpoints in place of the storage loop of Table 9.3. Every $T_p = 15\,\mathrm{min}$ (sixty scans) it plans the cooling each zone should take over the next 24 h, $N = 96$ steps, in seventeen move blocks that run quarter-hourly for the first half hour, half-hourly for the next, hourly to ten hours ahead, two-hourly to sixteen and four-hourly to the end (102 variables with the two load slacks per block), and only the first block reaches the plant before the next decision repeats the computation, the receding horizon of Section 1.4.3 [2, 19]. The regulatory loops are untouched, the storage PI tracks the applied setpoint so that a rejected plan hands it back without a bump, and where the supervisor writes one setpoint for four zones the plan writes one for each.

The controller's model is Equations (9.1) and (9.2) with the airflow term written as the cooling $Q_z = \dot{V}_z \rho c_p (T_{a,z} - T_s)$ the terminal delivers, in which the balance is exactly linear, plus an unmeasured load $d_z$ per zone, as shown in Equation (9.21).

$$\begin{aligned} \frac{C_a}{60}\frac{\mathrm{d}T_{a,z}}{\mathrm{d}t} &= -Q_z + UA_{ms}(T_{m,z} - T_{a,z}) + Q_{conv,z} + d_z \\ &\quad + UA_{oa}(T_o - T_{a,z}) \\ \frac{C_m}{60}\frac{\mathrm{d}T_{m,z}}{\mathrm{d}t} &= UA_{ms}(T_{a,z} - T_{m,z}) + Q_{rad,z} \end{aligned} \tag{9.21}$$

Every constant is the shipped default of Table 9.2 rather than the value in force on the run, so a learner who changes the building makes the model wrong in exactly that way, and $Q_{conv,z}$ and $Q_{rad,z}$

come from the schedules and the clear-sky geometry of Section 9.3 at their nominal values. The uncoupled time constants are 8.43 min for the air and 333 min for the slab. Stepped as the plant was in Section 9.4, the west terminal by 20 % with the air handler's loops closed, the model reads −0.541, −0.586 and −1.221 °C per m³/s at twenty minutes, an hour and two days against the plant's −0.537, −0.585 and −1.221, within 0.9 % with no fitted parameter.

The electrical side is a static map on the predicted cooling. At step $k$ one kilowatt of zone cooling costs $\alpha_k = f_k(1+a_1)/\mathrm{COP}_k + a_1$ electric kilowatts linearly and $\beta_k = a_2(1+f_k/\mathrm{COP}_k)$ per kilowatt squared, where $\mathrm{COP}_k$ is (9.11) at the forecast ambient and the 0.7 part load ratio that (9.18) also prices, $f_k$ is the share of the coil duty the chiller carries (one unless the economizer of (9.7) is open), and $a_1 = 0.0422$ and $a_2 = 3.09\times10^{-4}$/kW are the tangent at zero flow of the fan law of (9.6) at the static setpoint and the least-squares quadratic of its cube over 0 to 250 kW. The outdoor-air load is $\dot{V}_{oa}\rho c_p(T_o - \bar{T}_a)/\mathrm{COP}_k$ on the mean zone air the plan itself predicts, and the house load of (9.12) is added as it is. The map prices the energy of scenario 8's realized day at \$401.55 under the supervisor and \$402.62 under the plan, where the plant charged \$401.46 and \$402.47, and the flat-tariff day within 0.3 %.

A steady-state Kalman filter carries the estimate, on the twelve states $[T_{a,z}, T_{m,z}, d_z]$ sampled at the 0.25 min scan with $d_z$ a random walk, corrected from the five signals the shipped controllers already read, the four filtered zone sensors $\hat{T}_{a,z}$ and the stored cooling $E$ of (9.13), which fixes the slabs' sum. The per-scan variances are $10^{-4}$ and $10^{-5}$ K² on the air and the slab, 0.05 kW² on the load, 0.01 K² on each sensor and 25 kWh² on the stored cooling, and the gain is designed once at the declared defaults and frozen. One integrating load per measured zone makes the tracking offset-free, the input-disturbance estimate of Section 1.4.5 [20], and a 60 % plug-load step the model does not know lands each $\hat{d}_z$ within 30 % of the 7.5 kW it adds inside an hour.

The plan is the quadratic program of Equation (9.22), in dollars over the horizon, an economic objective in the sense of Section 1.4.8, with $Q_{z,k}$ the cooling of zone $z$ at step $k$ held over its block $b$, $p_k$ the tariff, $P_k$ the metered load the map predicts and $s$ the slacks.

$$
\begin{aligned}
\min\ & \sum_{k=1}^{N}\sum_{z} p_k\left(\alpha_k Q_{z,k} + \beta_k Q_{z,k}^2\right)\frac{T_p}{60} + r_d\, s_{pk} \\
& + c_{DR}\sum_{b} s_{DR,b}\frac{T_p}{60} + \sum_{z}\left(q_b s_z^2 + l_b s_z\right) \\
& + w_T\left(E_N - E_0\right)^2 + m\sum_{b}\sum_{z}\Delta Q_{z,b}^2 \\
\text{s.t. } & T_{lo} - s_z \le T_{a,z,k} \le T_{ceil,k} - \Delta_{sp} + s_z \\
& P_k \le P_{peak} + s_{pk}, \qquad P_k \le P_{cap} + s_{DR,b} \ \text{(notified)} \\
& \textstyle\sum_z f_k Q_{z,k} + Q_{vent,k} \le Q_{cap}(T_{o,k}) \\
& Q_{z,k} \ge \dot{V}_{z,\min}\rho c_p\left(T_{ceil,k} - T_s\right) \ \text{(occupied } k\text{)} \\
& T_{a,z,k} - T_{a,z,k'} \ge -\tfrac{r}{60}T_p\,(k-k'), \quad 0 \le Q_{z,k} \le Q_{box}
\end{aligned}
\tag{9.22}
$$

where $r_d$ and $c_{DR}$ are the demand charge and the penalty of (9.14), so the second term prices every kilowatt of the predicted peak above a ceiling $P_{peak}$ (zero by default) and the third every kilowatt-hour above a notified cap, $q_b = 40\,\$/\mathrm{K}^2$ and $l_b = 4\,\$/\mathrm{K}$ price each zone's worst excursion from the band over the horizon, $w_T = 2\times10^{-3}\,\$/\mathrm{kWh}^2$ holds the heat content of the fabric at the horizon's end, $E_N$, to its value now, and $m = 10^{-3}\,\$/\mathrm{kW}^2$ times the move setting damps the moves. The band, the demand ceiling and the cap are soft, each with its slack, as in Section 1.4.6. Hard are the box between zero (the coil cannot heat) and $Q_{box}$, the ventilation minimum $\dot{V}_{z,\min}$ while the building is occupied, the chiller's declared capacity of (9.10) at the forecast ambient over every block with $Q_{vent,k}$ the outdoor-air duty, and the ramp limit $r$ of (9.20) written on each zone's predicted air. The load slacks are decision variables in electric kilowatts, because a kilowatt of cooling costs a different number

**Table 9.4** Measured Outcome of Every Day-Long Scenario, Read From the Application's Scorecard at $t =$ 1440 min. Comfort Counts Occupied Zone Minutes Inside the Band, Per Zone.

| # | Day | kWh | Peak (kW) | Total ($) | Comf. (%) | Over cap (kWh) |
|---|---|---|---|---|---|---|
| 5 | design, fixed 23 °C | 2388 | 214 | 615 | 100.0 | — |
| 6 | mild, economizer at 09:00 | 1887 | 158 | 471 | 100.0 | — |
| 7 | design, band, $H = 0$ | 2269 | 202 | 585 | 100.0 | — |
| 8 | design, band, $H = 480$ min | 2302 | 188 | 561 | 100.0 | — |
| 9 | design, band, $H = 90$ min | 2286 | 222 | 602 | 100.0 | — |
| 10 | event, band, 120 min notice | 2266 | 209 | 431 | 97.8 | 16.1 |
| 11 | event, band, no notice | 2258 | 213 | 437 | 97.7 | 19.4 |
| 12 | event, fixed 23 °C | 2388 | 214 | 614 | 100.0 | 152.3 |
| 13 | heat wave, band, short plant | 2800 | 232 | 684 | 100.0 | — |
| 13 | heat wave, fixed 23 °C | 2817 | 247 | 723 | 89.8 | — |
| 14 | flat daytime tariff, band | 2280 | 192 | 397 | 100.0 | — |
| 15 | design, MPC, 24-hour plan | 2356 | 183 | 558 | 100.0 | — |
| 16 | challenge, as handed over | 2508 | 247 | 953 | 36.1 | 220.9 |
| 16 | challenge, suggested line-up | 2631 | 235 | 621 | 100.0 | 0.2 |

of electric kilowatts at every hour, and a zone's ceiling is relaxed by a funnel with a 20 min time constant after every drop of the band, so that the one slack the solver carries per output over the horizon is not set by an unavoidable recovery and then licensed everywhere.

The setpoint written to each zone loop is the air the plan predicts at the end of its first step, clamped into the band in force and paced downward by the ramp limit of (9.20) exactly as the supervisor's command is, and a zone the plan leaves at its ventilation minimum floats at the ceiling. The preview is what the supervisor already reads and nothing more: the tariff, the forecast ambient with its bias $b_f$, the band schedule, and the curtailment cap from the minute the notice arrives, through the same test (9.18) uses, and never before it: with two hours of notice the first cap row appears in the 09:00 plan. With the preview switched off the present price, ambient and event state are held over the horizon, and that setting is a row of every comparison in Section 9.6.6.

Five conditions admit a plan: an optimal status from the solver, a first move equal to the plan's first entry, every planned cooling inside its box, no negative slack, and every prediction, the cost and the four setpoints finite. Anything else is rejected, the historian is marked, and the band supervisor, whose storage loop has been tracking the mean of the four setpoints at every scan, takes the zone setpoint until a plan succeeds. Five forced rejections in the verification suite each moved the supervisor's command by 0.0006 °C against a ramp step of 0.0042 °C, and the next valid plan took the setpoint back fifteen minutes later. A plan takes 15.6 ms on average in the browser (the largest 41 ms), which is 0.10 % of one core at real time and 25 % at the 240 times speed, so a frame guard stops the simulation's stepping after 25 ms of wall time in a frame while the MPC is chosen. At a three-hour horizon a plan takes 0.3 ms.

Four models are now in play rather than three: the plant of Section 9.3, the loop gains of Section 9.4, the price and chiller curve of (9.18) the supervisor knows, and Equation (9.21) with the load map the plan knows, with no coil, duct, sensor filter or chilled water loop in it.

## 9.6 Results

Sixteen scenarios run in order. The first four are frozen at 15:00 on the design day so that nothing drifts underneath a measurement: the four actuators by hand, every loop closed on the bench, the supply air step with the terminals pinned, and every terminal asked for 115 %. The other twelve run a full 1440-minute day. A scenario pins the opening mode, the zone architecture, the four tunings, the disturbance sliders and any timed events, and the completion rules and the challenge work as Section 1.6.4 describes. The challenge is judged on the case as given, so changing a model parameter or a disturbance voids the attempt. Table 9.4 collects the scorecard each day-long run ends on.




**Figure 9.2** Scenario 8 at $t = 1440$ min, the end of its day, on the design weather with the supervisor given an eight-hour horizon, cropped to the application's plant view. During the day, as the trends recorded (they are not in the crop), the supervisor walked the west setpoint down to the 21.0 °C floor of the band by 07:42 and released it to the 26.7 °C night ceiling at 17:52, and the fabric filled to 68.4 % of its band, which is 585 kWh, by 08:44, and both regulatory loops held their setpoints whenever the fan ran. The frame catches the next morning's charge beginning at 04:00. (1) The chips: the day closes at 100 % comfort and $561. (2) The four zone strips, air over slab, with the band in force and the setpoint at 22.70 °C. (3) The cost inset: a kilowatt-hour of cooling now against the worst inside the horizon, the only pair of curves the supervisor reads, with the day's 188 kW peak below it. (4) The charge demand of 0.27. (5) The air handler strip with the fan restarting. At this minute the supply air was at 13.88 °C against 13.0 °C and duct static at 31 Pa against 300 Pa. The day used 2302 kWh (Table 9.4).

### 9.6.1 Floating a Band Is Not Using One

Scenario 5 is the building most of the stock actually is, a fixed 23 °C setpoint with night setback and optimum start, run competently. It holds every zone in band for 100 % of the occupied zone minutes and comes to 2388 kWh, $434 of energy and a 214 kW peak set by the quarter hour ending 15:27, which is $182 of demand charge for $615 all in. The orientation spread shows in the terminals rather than in the temperatures. East and west glass both peak at 705 W/m$^2$, while the east terminal tops out at 4.68 m$^3$/s at 10:33 and the west at 5.51 m$^3$/s at 16:56, because a zone integrates and because only one of those hours coincides with a 35 °C ambient, a full building and the on-peak rate.

Scenario 7 gives the same building a comfort band and a zero horizon, so the supervisor never sees a dearer hour and simply parks the zones just under the ceiling. The day costs $585 against $615, a saving of 4.9 %, and the peak falls only from 214 to 202 kW. What it is not is storage. The charge demand of (9.19) never leaves zero, the stored cooling of (9.13) peaks at 286 kWh at 08:40 on nothing but the slab's overnight coolth meeting a tightened band, and the afternoon load is met the moment it arrives. Widening a band saves energy, because a building rejects roughly in proportion to how far inside is from outside and seven tenths of a degree is not far. Using a band saves money, and the two are different operations.

🕮 **Pause.** Try Exercise 9.1 before reading on.

Scenario 8 gives the same supervisor an eight-hour horizon. It reaches the 21.0 °C floor by 07:42, the fabric is at its fullest at 585 kWh at 08:44, and the day costs $561 with every zone in band all day. Against the fixed setpoint that is 8.8 % off the bill. Two features of that saving are worth separating. Twelve dollars of it is energy and twelve is demand charge, because a 188 kW peak is fourteen kilowatts below a 202 kW one, so exactly half of the win is a term the fixed-setpoint building

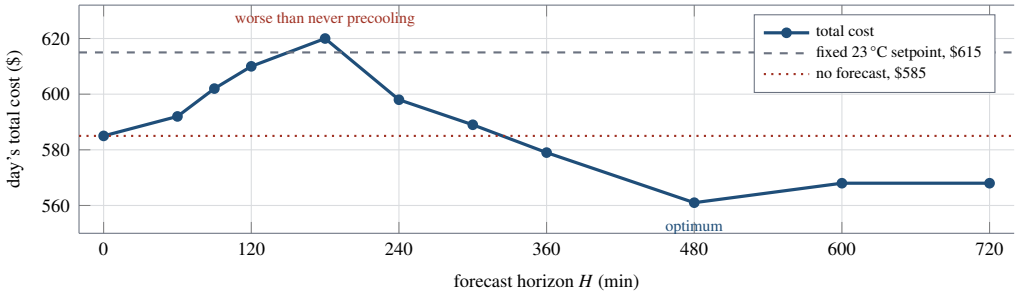



**Figure 9.3** The day's bill against the supervisor's forecast horizon on scenario 8's weather, with everything else unchanged. Between about one and five hours a forecast is strictly worse than no forecast, and at three hours the building pays $620 against the $615 of a fixed 23 °C setpoint. Comfort is 100 % at every point on the curve, so the whole of the variation is money.

cannot reach. And the day uses more electricity than the floating case, 2302 kWh against 2269 kWh. Precooling is not an efficiency measure. It buys the same cooling at a better hour and accepts a little extra loss for doing it.

> **TRY IT IN THE APP** **8 · The Band, With Eight Hours of Forecast**
>
> Load the scenario and press Run in Auto. Watch the inset of what a kilowatt-hour of cooling costs, now against the worst ahead, and then the zone setpoint walk down to the 21.0 °C floor by 07:42. The stored cooling is fullest, at 585 kWh, at 08:44, and the day closes at $561 with every zone in band (Table 9.4).

The supervisor has no schedule, no clock and no rule naming a peak. Scenario 14 removes the on-peak rate, so electricity costs $0.110 per kilowatt-hour from 08:00 to 22:00 and the $0.075 overnight rate is the only step left in the tariff, and the building still precools. Floating costs $409 and the supervisor costs $397 with the peak down from 202 to 192 kW. Nearly two thirds of that saving is cooling bought before eight at the overnight rate. Raising the overnight rate to $0.110 as well, which leaves only the chiller curve of (9.11) and the demand meter, cuts the saving to $4.39 ($417.14 against $421.53) and the peak reduction to 202 against 198 kW. The same scenario measures the forecast, because an error of −6 °C gives $396 and +6 °C gives $398 against $397 with none. The forecast is worth about a dollar a day here, less than the horizon setting and very much less than the width of the band.

### 9.6.2 A Forecast Horizon That Fails at Both Ends

Scenario 9 cuts the horizon to ninety minutes and costs $602 against the $585 of a building that never precooled. Every cent of the difference is demand charge, because the energy bill is $413 either way while the peak goes from 202 to 222 kW. Sweeping the horizon over the same day gives Figure 9.3, over which the energy bill moves from 401 to $414 while the demand charge moves from 160 to $206, so the horizon is a demand-charge knob and very little else.

The shape has one rule behind it. The horizon has to cover the time it takes to charge the store, not the time until the peak. The floating building draws 137 to 163 kW between nine and eleven against a 202 kW afternoon peak, so a charge has 39 to 66 kW of headroom before it sets a new peak, and the slab takes up what it is given on a 335 min time constant. At eight hours the supervisor starts charging at four in the morning and the store is not at its fullest until 08:44. At three hours it first sees the 14:00 price rise at 11:00, takes the setpoint from 23.7 to 22.3 °C in the next two hours and sets the day's 242 kW peak at 14:00, the minute the on-peak rate begins. A horizon shorter than the charging time guarantees the charging happens in a rush at the worst possible hour. Comfort is 100 %

at every horizon in the sweep, which is what makes the failure invisible from the temperature trends and visible only on the demand meter.

🕮 **Pause.** Try Exercise 9.2 before reading on.

Raising the aggressiveness $K_{pre}$ at a long horizon produces the same failure by a different route, because the charge demand saturates the moment the peak comes into view, the setpoint dives and four terminals open together. At $K_{pre} = 0.8$ and $H = 480$ min the day costs \$613 at a 243.5 kW peak and 98.5 % comfort, against \$561 and 188 kW at the shipped 0.20. The ramp limit of (9.20) is what stands between a storage controller and its own demand charge.

### 9.6.3 A Curtailment, Run Three Ways

Scenarios 10 to 12 are the same day three times. The tariff is flat from 08:00 to 22:00, so the only price signal in the working day is the event itself, a 140 kW cap on the whole building from eleven until two with \$1.20 per kilowatt-hour charged on everything above it, against 93.5 kW of lights, equipment and house load.

With two hours of notice nothing in the controller changes except one term of (9.18). At nine o'clock the forecast cost of a kilowatt-hour of cooling inside the event window goes from 3.6 to 22.6 cents, the ratio in (9.19) goes from 1.40 to 8.85 and saturates the charge demand, and the setpoint walks down at its 1 °C/h limit. The plant runs to 190 kW by half past ten and stops charging half an hour before eleven so the chiller has unloaded when the cap takes effect. Through the event the zones drift up to 25.3 °C, under the 25.5 °C ceiling the tenants agreed to, and the building draws 16.1 kWh above the cap, \$19 of penalty on a \$431 day.

Taking the notice away gives 19.4 kWh and \$437. That is the honest size of two hours of warning on this plant, a sixth of the overage and six dollars, and it is much smaller than intuition suggests. The reason is arithmetic rather than tuning. Two hours of charging on a store whose time constant is 335 min moves the west slab from 24.47 to 24.30 °C, a difference of 0.17 K, and puts 327 kWh in the fabric at eleven o'clock instead of 278 kWh. Eight hours of notice instead of two still leaves 5.7 kWh above the cap with 686 kWh stored.

Running the same event on a building holding 23.0 °C because that is what it was told to hold gives 152.3 kWh above the cap, \$183 of penalty and \$614 for a day the band building did for \$431. It is also the most comfortable of the three on paper, at 100 % of occupied zone minutes inside the band against 97.8 % and 97.7 %, because it never let a zone move, and it paid for that at \$1.20 per kilowatt-hour. The storage was present in all three buildings, because it is the same slab and the same 856 kWh. What the first two had and the third did not was permission to use it.

Where the remaining 16.1 kWh sits explains why more notice cannot reach zero, and it is at the end of the event rather than the start, with 0.1, 2.5 and 13.4 kWh drawn in its three hours. From the first scan of the event every terminal is on its ventilation minimum, because the relaxed ceiling lifts the setpoint to 25.2 °C at once, total flow falls from 11.6 to 6.75 m³/s, and (9.7) still asks for up to 4.17 m³/s of outdoor air, so the outdoor air fraction is 0.62 as the event begins and still 0.46 over lunch at one o'clock, and the coil cools most of a summer afternoon from scratch. Whole-building demand bottoms out at 133 kW at 11:02 and climbs to 160 kW by two o'clock as the outdoor air warms, and at one o'clock the 146 kW on the meter is 93.5 kW of house load, 46.1 kW of chiller, 4.5 kW of fan and 2.3 kW of pump. Once the terminals are on their minimum the store buys comfort rather than kilowatts. Precooling buys the first hour of a curtailment and the ventilation rate decides the third, and the ventilation rate is not a control setting.

🕮 **Pause.** Try Exercise 9.3 before reading on.

The precool does not set a new peak here. The draw jumps to 211 kW at three minutes past two as the plant restarts, but the day's 209 kW demand peak is set at 15:38 by the ordinary afternoon load, and it is below the 214 kW of a building that did nothing. The event also does not end when the event ends, because the zones are at the relaxed ceiling at two o'clock and the ordinary 24 °C band returns the same minute, so every one of the lost comfort minutes falls after the curtailment finished.

### 9.6.4 A Plant That Runs Short

Scenario 13 raises the whole diurnal curve by 6 °C and removes a fifth of the chiller. On a fixed 23 °C setpoint the machine gives out first and takes every other loop with it, because the chilled water walks to 14.3 °C, the supply air setpoint becomes unreachable and every coil valve loses authority at the same instant, so each zone spends 61 min above the band for 89.8 % comfort and the day costs $723 at a 247 kW peak. The same plant under the band architecture never leaves setpoint. Chilled water holds 6.00 °C all day, comfort is 100 %, the peak is 232 kW and the day costs $684, on a plant that is no larger. What precooling buys on a day like this is capacity that did not have to be installed.

### 9.6.5 The Graded Challenge

The challenge is a day three degrees above design with ten percent more people and a three-hour 165 kW curtailment from three o'clock notified at one, handed over on a fixed setpoint with soft tuning and the economizer switched off. The day passes when every zone is in band for at least 99 % of the occupied zone minutes, less than 5 kWh is drawn above the cap and the bill is under $750 all in. As handed over it costs $953, holds the band for 36.1 % of the occupied zone minutes and draws 220.9 kWh above the cap. The suggested configuration does it for $621 with 100.0 % comfort and 0.2 kWh above the cap, margins of 1.0 percentage point, 4.8 kWh and $129, whether the settings are applied before Run or carried through Reset. It departs from the tuning the other scenarios ship in one supervisory value, a precool aggressiveness $K_{pre}$ of 0.30 rather than 0.20. The supervisor is in service from the priming scan at 04:00 and charges from the 23.0 °C operator setpoint, and at 0.20 it commits too little of the band, holds 342 kWh at three o'clock and lets the day draw 6.9 kWh above the cap, which fails. At 0.30 it holds 539 kWh.

Disabling one layer at a time from that configuration ranks the improvements the scenario asks for, and the ranking is the result. Restoring the handed-over zone loop tuning of $K_c = 0.30$ and $\tau_I = 55$ min alone takes the day to $840 and 112 kWh above the cap, and restoring the handed-over duct static tuning alone drops comfort to 95.3 %. The handed-over supply air tuning of $K_c = 1.5$ and $\tau_I = 18$ min costs $20 and the day still passes. Returning to a fixed setpoint takes the day to $909 and 188.9 kWh, and keeping the band on a 30 min horizon takes it to $681 and 31.2 kWh, because the supervisor does not see the curtailment until half past two. The economizer is worth exactly nothing. Switched on or off the day costs $621 either way to the dollar and draws 0.2 kWh above the cap either way, because on a day three degrees above design the outdoor air never falls below 20.0 °C and so never gets under the 18 °C high limit of (9.7). The damper never opens past its ventilation minimum. The scenario's briefing asks for the economizer to be switched on, and it was worth checking rather than assuming.

The band edges are the part with no tuning rule at all. Every tenth of a degree of width is 28.5 kWh of storage by (9.13), so narrowing the band to 22 to 24 °C takes the same day from 0.2 to 5.9 kWh above the cap and dropping the event ceiling from 25.5 to 24.5 °C takes it to 28.0 kWh and $665. Both fail, and neither is an engineering decision.

One thing the suggested answer does not fix is worth stating rather than hiding. The day's 235 kW demand peak is set at 14:00 by the charge before the event. It is below the 247 kW the untouched building reached, so against the day as handed over the energy bill falls $57, the penalty $265 and the demand charge $11, but Equation (9.18) prices energy and a penalty and knows nothing about the demand charge, and a supervisor carrying that term would ramp more gently and start earlier. The predictive controller of Section 9.5.4 carries it, and Section 9.6.6 measures what that is worth on this day.

Table 9.4 carries one caution. Every figure in it is a whole-day total over the same 1440-minute window, so the comparisons between rows are sound, and none of them is an integrated error that would keep accruing after a transient ended. The comfort column is the one that misleads on its own. Row 12 is the most comfortable and by $183 the most expensive of the three curtailment days, and rows 10 and 11 lose their comfort minutes after the curtailment they survived had already finished.

**Table 9.5** The Plan Against the Band Supervisor Over Full Days With the Same Seed, Instruments and Forecast. Excursion Is the Band Overshoot Integrated Over Occupied Zone Minutes and Travel the Supervisor's Setpoint Travel.

| Case | Structure | Bill ($) | Energy ($) | Demand ($) | Penalty ($) | Peak (kW) | Over cap (kWh) | Comfort (%) | Excursion (°C min) | Travel (°C) |
|---|---|---|---|---|---|---|---|---|---|---|
| 8, design day | supervisor, $H = 480$ min (shipped) | 561.19 | 401.46 | 159.73 | 0 | 187.9 | — | 100.0 | 0.0 | 13.29 |
| | supervisor retuned once, $K_{pre} = 0.25$ | 552.74 | 398.96 | 153.78 | 0 | 180.9 | — | 100.0 | 0.0 | 12.59 |
| | MPC, 24 h (scenario 15) | 557.61 | 402.47 | 155.14 | 0 | 182.5 | — | 100.0 | 0.0 | 9.34 |
| | MPC, 3 h | 578.65 | 408.85 | 169.80 | 0 | 199.8 | — | 100.0 | 0.0 | 10.58 |
| | MPC, 3 h, demand charge not priced | 650.89 | 436.93 | 213.96 | 0 | 251.7 | — | 100.0 | 0.0 | 16.30 |
| | MPC, 24 h, demand charge not priced | 600.28 | 407.78 | 192.50 | 0 | 226.5 | — | 100.0 | 0.5 | 10.35 |
| | MPC, 24 h, preview off | 566.96 | 403.53 | 163.43 | 0 | 192.3 | — | 100.0 | 0.0 | 7.12 |
| 10, curtailment, 120 min notice | supervisor (shipped) | 431.16 | 233.99 | 177.90 | 19.27 | 209.3 | 16.1 | 97.8 | 25.1 | 15.55 |
| | supervisor retuned once | 427.83 | 233.23 | 176.76 | 17.83 | 208.0 | 14.9 | 97.8 | 24.7 | 15.79 |
| | MPC, 24 h | 421.43 | 237.37 | 164.46 | 19.60 | 193.5 | 16.3 | 100.0 | 0.0 | 12.86 |
| | MPC, 3 h | 449.36 | 235.88 | 164.93 | 48.55 | 194.0 | 40.5 | 100.0 | 0.1 | 16.50 |
| | MPC, 24 h, preview off | 422.20 | 229.99 | 175.53 | 16.68 | 206.5 | 13.9 | 97.8 | 24.6 | 9.66 |
| 11, curtailment, no notice | supervisor (shipped) | 437.05 | 233.08 | 180.65 | 23.31 | 212.5 | 19.4 | 97.7 | 25.6 | 12.55 |
| | supervisor retuned once | 433.50 | 232.42 | 179.53 | 21.56 | 211.2 | 18.0 | 97.8 | 25.4 | 12.79 |
| | MPC, 24 h | 421.69 | 236.38 | 164.82 | 20.50 | 193.9 | 17.1 | 100.0 | 0.0 | 12.39 |
| | MPC, 3 h | 451.19 | 236.44 | 163.91 | 50.84 | 192.8 | 42.4 | 100.0 | 0.1 | 15.74 |
| 13, heat wave, chiller at 0.8 | supervisor (shipped) | 684.30 | 487.09 | 197.21 | 0 | 232.0 | — | 100.0 | 0.0 | 13.74 |
| | supervisor retuned once | 668.77 | 480.54 | 188.22 | 0 | 221.4 | — | 100.0 | 0.0 | 12.94 |
| | MPC, 24 h | 682.12 | 487.09 | 195.03 | 0 | 229.4 | — | 100.0 | 0.0 | 8.70 |
| | MPC, 24 h, preview off | 693.38 | 491.42 | 201.95 | 0 | 237.6 | — | 100.0 | 0.0 | 8.06 |
| 14, flat daytime tariff | supervisor, $K_{pre} = 0.5$ (shipped) | 396.65 | 233.22 | 163.43 | 0 | 192.3 | — | 100.0 | 0.0 | 11.76 |
| | supervisor retuned once, $K_{pre} = 0.25$ | 403.18 | 235.18 | 168.00 | 0 | 197.7 | — | 100.0 | 0.0 | 9.79 |
| | MPC, 24 h | 398.41 | 243.21 | 155.20 | 0 | 182.6 | — | 100.0 | 0.0 | 8.78 |
| | MPC, 24 h, preview off | 398.66 | 235.05 | 163.61 | 0 | 192.5 | — | 100.0 | 0.0 | 7.45 |
| 16, challenge | as handed over (fixed setpoint, soft loops) | 952.77 | 477.34 | 210.37 | 265.07 | 247.5 | 220.9 | 36.1 | 937.1 | 8.00 |
| | suggested (supervisor, $K_{pre} = 0.30$, tuned loops) | 620.53 | 420.71 | 199.59 | 0.23 | 234.8 | 0.2 | 100.0 | 0.0 | 14.25 |
| | suggested loops, supervisor retuned once | 616.52 | 415.19 | 200.94 | 0.38 | 236.4 | 0.3 | 100.0 | 0.0 | 14.82 |
| | suggested loops, MPC, 24 h | 589.68 | 414.28 | 173.71 | 1.69 | 204.4 | 1.4 | 100.0 | 0.0 | 8.40 |
| | suggested loops, MPC, preview off | 612.13 | 412.71 | 181.89 | 17.53 | 214.0 | 14.6 | 100.0 | 0.0 | 5.51 |
| | handed loops, MPC, 24 h | 761.27 | 440.45 | 185.34 | 135.47 | 218.0 | 112.9 | 28.4 | 1633.6 | 6.03 |

Every MPC row makes 96 plans in a day, one every fifteen minutes from the priming scan, none rejected, at 15 ms a plan (0.3 ms at three hours). The supervisor retuned once is $K_{pre} = 0.25$ at $H = 480$ min on every case. On scenario 11 the preview-off plan gives scenario 10's preview-off row, because a plan that holds the present event state learns of the curtailment only when it starts.

### 9.6.6 A Receding-Horizon Plan Against the Priced Rule

Scenario 15 runs scenario 8's day under the MPC of Section 9.5.4 with the preview on, a 24-hour horizon, the demand ceiling at zero and the move setting at one, and those settings are carried to every other case with no retuning. Table 9.5 sets the plan beside the band supervisor at its pinned tuning on the day-long cases that exercise the supervisor and on the challenge, each a full day with the same seed, instruments, noise, tariff, forecast and notice for both. The third row of each case is the supervisor retuned once, at $K_{pre} = 0.25$ and the eight-hour horizon, the single best setting over scenarios 8, 10 and 14 together on a grid of 72, and the remaining rows are the plan at three hours, without the demand charge in its objective, and with its preview off. Every number in Table 9.4 is the same run as before.

> **TRY IT IN THE APP** **15 · The Same Day, Planned: A 24 h MPC Against the Priced Rule**
>
> Load the scenario and run the day on MPC at speed 240. Watch the MPC plan inset and the four zone setpoints: the plan never takes a zone to the floor, and at noon the west sits at 21.91 °C against the north's 22.91 °C. Then read Today and Peak demand and set them beside $557.61 and 182.5 kW (Table 9.5).

On the design day the plan closes at $557.61 and a 182.5 kW peak against the supervisor's $561.19 and 187.9 kW, both at 100 % comfort. The $3.58 is $4.59 less demand charge and $1.00 more energy, because both buy their cooling before eight at the overnight rate, and the plan then shapes the afternoon under the peak it has already set, as shown in Figure 9.5: its 182.5 kW peak is set at 10:15 by the charge itself and the meter stays under 181 kW through the on-peak window, where the supervisor's day peaks at 16:13. The plan never takes a zone to the 21.0 °C floor, and it releases the north first, to 22.91 °C by noon against the west's 21.91 °C, which one setpoint for four zones

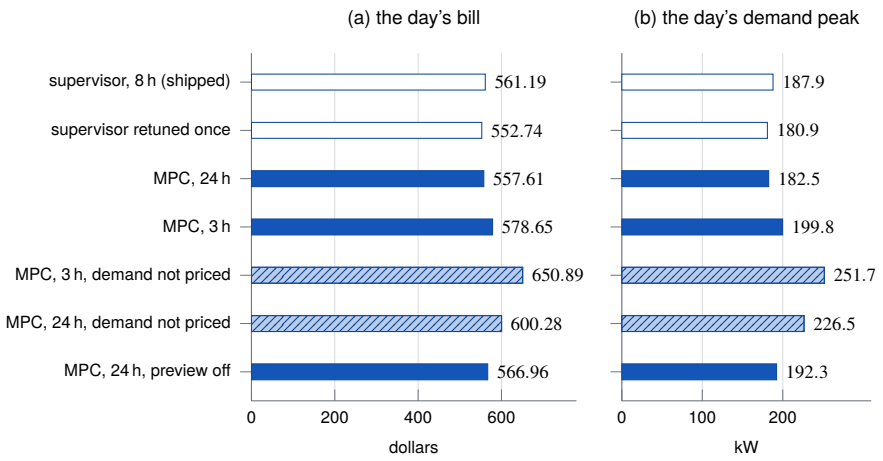



**Figure 9.4** The design day of Table 9.5 (case 8) under the band supervisor (open bars), the plan (dark bars) and the plan with the demand charge left out of its objective (hatched bars): (a) the bill and (b) the demand peak. Notice that the hatched bars are the two longest in both panels, at either horizon, while the supervisor retuned once has the shortest.

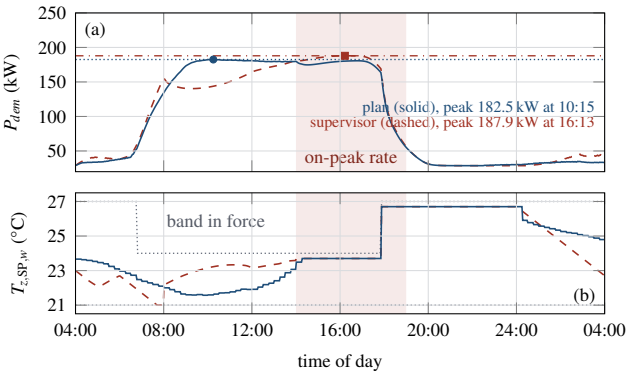



**Figure 9.5** The design day under the plan (scenario 15, solid lines) and under the band supervisor (scenario 8, dashed lines). (a) The fifteen-minute demand: the plan's 182.5 kW peak (dotted line, circle) is set at 10:15 by its own charge and the meter stays under 181 kW through the on-peak window, shaded, where the supervisor's 187.9 kW peak (dash-dotted line, square) falls at 16:13. (b) The west zone setpoint: the plan's quarter-hour steps against the supervisor's 1 °C/h ramp to the 21.0 °C floor at 07:42, with the band in force as gray dotted lines.

cannot do. The win is against the shipped tuning. The supervisor at $K_{pre} = 0.25$ does the same day for \$552.74 at 180.9 kW, \$4.87 better than the plan, and beats it on the heat wave as well, \$668.77 against \$682.12, where the plan ties the shipped tuning's \$684.30 to within \$2.18. Figure 9.4 draws the design-day rows of Table 9.5.

The curtailment days are where the plan buys what no setting of the supervisor buys. With two hours of notice it costs \$421.43 against \$431.16 and draws 16.3 kWh above the cap against 16.1 kWh, at 100 % comfort where the supervisor loses 13.5 min and 25.1 °C min of excursion after the event, and no setting of the supervisor on the grid reaches 100 % on that day. Two thirds of the plan's overage, 10.6 kWh of the 16.3, falls in the event's last quarter hour, because between 13:45 and 14:00 the plan opens the terminals and brings the zones from 25.2 °C to the 24.0 °C ceiling that returns at

two o'clock, which the supervisor does not attempt. With no notice it costs $421.69 against $437.05 and 17.1 kWh against 19.4 kWh. With its preview off, the plan costs $422.20 on the notified day at 97.8 % comfort and 13.9 kWh, because a plan that holds the present price has nothing to arbitrage and sees the cap only once the event is running, so it does what the supervisor does.

The flat daytime tariff is the case the plan loses, by $1.76: $398.41 against $396.65, a 10 kW lower peak ($8.23 of demand charge) bought with $9.99 more energy. The loss is not the demand weight, which is the tariff's own $0.85 per kilowatt, because with the demand charge left out of the objective the day costs $399.35 and with a 190 kW ceiling $396.00. The load map prices both realized days' energy within 0.3 %, which is the size of the difference, so on a day with nothing to arbitrage but the chiller curve the two controllers are inside the model's resolution of each other, the dependence on the rate structure that the optimal-control studies of building thermal storage examine [21].

On the challenge the plan is given the suggested regulatory tuning and economizer with the zone control alone changed, and closes at $589.68 with 100 % comfort, 1.4 kWh above the cap and a 204.4 kW peak, against the suggested configuration's $620.53, 0.2 kWh and 234.8 kW. The demand term the supervisor omits is worth $25.88 on this day, and the charge before the event, which sets the suggested configuration's peak at two o'clock, is paced under a peak 30 kW lower.

The three-hour horizon of scenario 9 is the counterfactual scenario 15 asks for. The plan at three hours costs $578.65 at 199.8 kW, better than the supervisor at the same horizon ($619.98 and 242.2 kW) and $21.04 worse than at 24 hours, because it cannot value the cheap morning until the on-peak window is three hours away. It does not reproduce the supervisor's failure, and the reason is the demand term: the same three-hour plan with the demand charge left out of its objective costs $650.89 at 251.7 kW, the pathology of Figure 9.3 and worse, and at 24 hours without the demand charge $600.28 at 226.5 kW, worse than the supervisor at eight hours. A short horizon charges in a rush whether the charge is a rule or a plan, and what spares the plan is that it prices the peak it is about to set. The lesson of Figure 9.3 survives in a weaker form: the horizon that sees the peak but not the time to charge for it is still the expensive one.

**DID YOU KNOW? A Plan Cannot Rescue Soft Loops**

From the simulator: on the challenge day the 24-hour plan, set above the zone, supply air and duct static loops as handed over, closes at $761.27 with 28.4 % comfort and 112.9 kWh above the cap. Above the suggested loop tuning the same plan closes at $589.68 with 100.0 % comfort (Table 9.5, case 16).

## 9.7 Technology in Practice

### What Measures It

**TECHNOLOGY Sensor: The Zone Temperature Sensor**

In a variable air volume system each zone has a temperature sensor that reports to the controller of its terminal box [22]. In a plain thermostat the sensor and the controller share one housing [23]. Its faults are a dead signal, a constant bias from lost calibration or a poor location, a drifting signal, and an intermittent or noisy one [22]. Space temperature sensors are among the points that most need calibration, and one out of calibration can waste energy for years unnoticed [23]. A precooling field test added sensors inside hollow spheres, which weigh the radiant temperature of the cooled surfaces with the air temperature and so track comfort better [18]. No open, vendor-neutral source gave a typical range or accuracy. The chapter's zone sensor is a 1.0 min lag with seeded noise (Table 9.2).

A building automation system gathers outdoor temperature and humidity, indoor temperature, humidity and carbon dioxide, and the state, temperatures and flows of its equipment [4]. Temperature, humidity, pressure, flow and power are its common sensors [23]. Occupancy is read more coarsely: motion sensors give presence only, while advanced HVAC control needs a head count, and carbon dioxide, often used as its proxy, drifts and needs frequent calibration [4]. In this chapter the zone

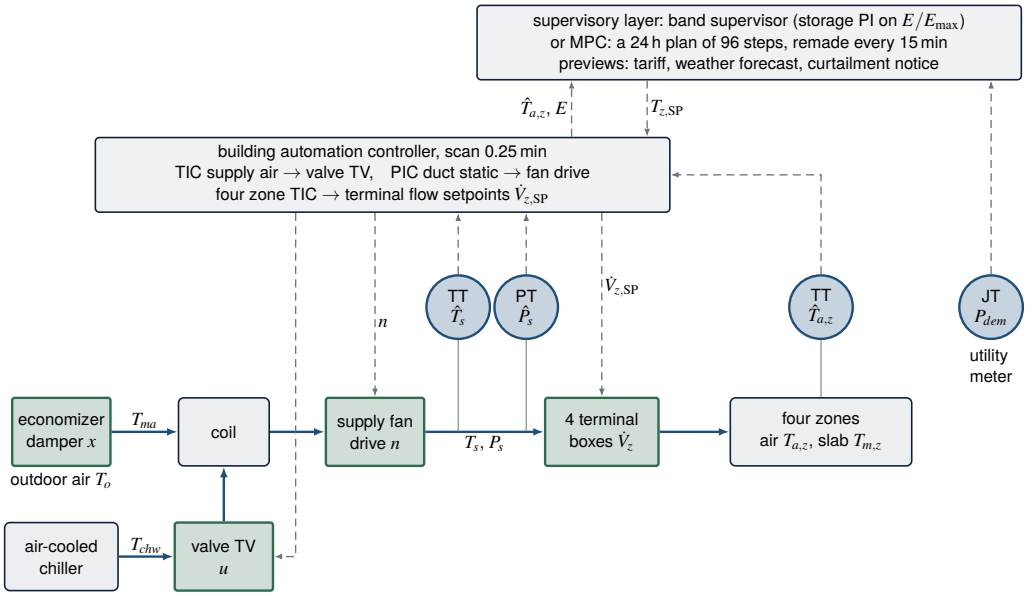



**Figure 9.6** The chapter's loops on field hardware, drawn from Sections 9.3 and 9.5. Green boxes are the final elements (economizer damper, chilled water valve TV, supply fan drive and the four terminal boxes), blue circles the sensors (supply air, duct static, the four zone sensors drawn as one, and the utility meter), dashed arrows signals. Notice that the supervisory layer moves no device: it writes the zone setpoint, and the zone loops turn it into airflow.

sensors, the supply air thermistor and the duct static transmitter of Figure 9.1 are first-order lags of 1.0, 0.30 and 0.08 min, and the utility meter is a 15 min demand window (Table 9.2). Occupancy enters as a known schedule.

### What Moves It

The devices a building controller drives are air dampers, mixing boxes, control valves and, in some cases, fans, pumps and motors [23]. The terminal dampers also steer the fan: a retrofit guide uses their positions to reset the duct static setpoint, raising it when a set number of dampers are wide open and lowering it when most are not [5]. The chapter's chilled water valve and fan drive are 0.50 min lags, the terminals 0.35 min lags, and the economizer damper opens only as far as the coil-leaving target needs (Equation (9.7)).

> **TECHNOLOGY** **Final Element: The VAV Terminal Box**
>
> A pressure-independent terminal box carries an airflow sensor and a damper. Its controller moves the damper to meet the zone's load, and a preset minimum airflow keeps ventilation air flowing when the zone needs no cooling [22]. A common practice sets that minimum at half the maximum, more than ventilation needs, and lowering it to each zone's real need saves fan energy [5]. Faults met in service include a damper stuck open, shut or part way from a failed motor, a broken linkage or corrosion, a wrong minimum or maximum flow setpoint, and a biased or drifting airflow sensor [22]. In the chapter each terminal is a 0.35 min lag capped by what the duct static will pass, with a minimum of 0.30 of zone design flow (Equation (9.4), Table 9.2).

### Where the Controller Runs

Direct digital control measures a variable, compares it with a setpoint and signals a terminal device, using proportional-integral algorithms [23]. Building controls form a hierarchy: local controllers at

the devices answer setpoints, and supervisory levels above them carry precooling and the temperature and pressure resets, usually as rules [4]. Supervisory and lower loops run with cycle times from milliseconds to minutes [24]. For load shifting the report gives a predictive controller a horizon of typically a day and a re-evaluation interval between 5 and 15 min, since the thermal dynamics are slow [4]. The chapter's four loops scan every 0.25 min and its plan looks 24 h ahead and is remade every 15 min, as Figure 9.6 draws them.

> **IN PRACTICE** **Widen the Deadband, Then Keep It Wide**
>
> Many buildings run with little or no deadband, the gap between the heating and cooling setpoints. A federal retrofit guide for offices widens it from 71 and 73 °F to 69 and 75 °F, since most occupants are comfortable between 73 and 79 °F in summer. With direct digital control at the zones it is one global edit at the operator workstation [5]. A federal maintenance guide asks for the widest gap between the two setpoints, an unoccupied cooling setback 5 to 10 °F above the occupied setpoint, and overrides recorded and removed as soon as possible [23]. The chapter's 21 to 24 °C band and 27 °C night setback are such settings, owned by the operator, not the controller.

### What Is Modernizing It

Three developments bear on this chapter's method. Predictive supervisory control is the first: early experiments at building sites reported energy savings of up to 17 to 65 percent against the installed control sequences [4]. Grid signals are the second. An open data schema exchanges electricity prices and demand response commands between buildings and the grid, and some programs use communicating thermostats to precool before a setback during peak hours, which avoids both discomfort and a recovery spike [4]. The third is cheaper sensing and diagnosis: wireless, battery-powered sensors make retrofits more economical, and newer automation systems can detect and diagnose equipment faults and advise the operator [4]. Each one feeds the plan a signal this chapter takes as known: the tariff, the curtailment notice, and a measured state of the fabric.

## 9.8 Limitations

The application is not a building energy model. With no moisture anywhere, two nodes per zone, no core zone and a Carnot chiller, nothing here settles a question about a real building's annual consumption, its plant sizing, or the load it could shed next August. Its figures are realistic for a dry-climate office of this size, at 856 kWh across a 3 K band, a coefficient of performance from 5.43 to 3.10 across one day, and about 40 % of the building's electricity in the HVAC plant, and none of them is a measurement of any particular building.

Three limits bear directly on the results reported here. The two-node zone of Section 9.3.7 reaches the last part of the store faster than a real slab would, so every storage figure is optimistic and the curtailment results are the ones most exposed to that. The demand meter is a first-order lag rather than a clock-aligned integrating window, so a real peak that straddles two window boundaries reads lower than Table 9.4 shows, and the demand charge is the term the horizon sweep of Figure 9.3 is entirely about. The supervisor prices energy and a curtailment penalty and does not price the demand charge at all, which is why the charge before the challenge's event still sets that day's 235 kW peak at two o'clock.

The plan's model leaves out what the plant has and the supervisor's model never needed: the sensor filters, the terminal and duct lags of (9.4), the chilled water loop of (9.9) that walks up when the plant is short, the pump, the part-load dependence of (9.11) and every disturbance slider, whose effect on the zones the estimated load carries while the capacity row keeps the declared capacity. The comparison of Section 9.6.6 is read at that model's resolution, 0.3 % of a day's energy, which is why the $1.76 flat-tariff loss is reported as a tie and not tuned away. The three-hour result differs from the supervisor's by one term rather than by the horizon, because a plan that does not price the demand charge fails at three hours as the supervisor does, at $650.89 against $619.98.

Repeatability is itself unrealistic. A frozen step, a frozen scan period and a frozen seed give the same trajectory everywhere, whereas a real building automation system jitters its scan, quantizes every analog point more coarsely and reorders what it reads. The four features that survived did so because each one forces a decision on whoever designs the controller, and they are the separation of 61 between the two thermal nodes, the shared fan that runs out unequally, the chiller whose efficiency follows the weather, and a controlled variable that is an interval.

> **OPEN CHALLENGE** **Why Predictive Control Is Still Rare in Buildings**
>
> A national report on building controls finds that predictive control has not been widely adopted in buildings despite field demonstrations. Legacy automation hardware has little memory and computation, cloud services bring latency and lost connections, accurate and robust models are costly to develop, train and calibrate, and operators are slow to accept the result. An optimizer's choices can also look like a black box, harder to trust and to tune by hand [4]. This chapter's plan takes its model from the declared constants with no fitted parameter. Progress would be models that update themselves from measured data and plans restated as rules an operator can check, both of which the report proposes [4].

## 9.9 Discussion

The comparison between scenarios 10, 11 and 12 does not come out the way the framing of most demand response programs suggests. Notice is what those programs are built around, and on this plant two hours of it takes 3.4 kWh, a sixth, off the overage and $6 off the day's bill. Permission takes off 136 kWh and $183 against the same cap on the same day, about forty times the energy and thirty times the money, and it is not a control setting. Buying notice and leaving the comfort specification alone optimizes the smaller of the two terms [17], and the same conclusion has been reached from the optimal-control side, where the strategies recovering most of the available reduction are the ones given the widest temperature trajectory [13, 15].

The model of Section 9.3 supports methods beyond the implemented loops. The seventeen-state plant with a stated objective, a hard ventilation floor and a band constraint is already in the shape a predictive controller wants, and the controller of Section 9.5.4 is that shape on two nodes per zone, carrying the demand charge this supervisor omits [6, 7] and an estimate of the structural temperature that nobody instruments. A field implementation on an occupied office building reports its results together with a cost-benefit account [25]. Whether any of it is worth doing is settled before the controller is specified, by whether the band exists on paper. Chapter 6 charges the same kind of store, buildings, pipes and a tank, ahead of a morning heat peak, and Chapter 8 selects the setpoints of a cooling plant whose optimum sits on a constraint that moves with the weather.

> **OPEN CHALLENGE** **Weights for Comfort, Cost and the Grid**
>
> The same report calls the balance among energy efficiency, grid benefits and occupant preferences a key challenge. Weights in one cost function are usually found by iterating in simulation, which means costly retuning once the controller is deployed, and occupant preferences are subjective and may change with time [4]. This chapter prices the band at fixed dollars per kelvin squared and carries one tuning of the plan to every case, and a supervisor retuned once beats it on the design day and the heat wave. Progress would be weights set from measured comfort and from preferences that occupants state, which the report lists as open research [4].

## 9.10 Conclusion

This chapter studied a commercial building and its air handler as a laboratory for control against a specification that is an interval. The simulator integrates seventeen states, four zone air temperatures over four structural temperatures, four terminal flows, the supply air, the chilled water, two actuators and the utility demand meter, with a counter-flow coil, a Carnot-fraction air-cooled chiller whose condensing temperature follows the weather, and a dry-bulb economizer. The two thermal nodes in

one room answer in 5.45 and 335 min, a ratio of 61, and the second holds 856 kWh across a 21.0 to 24.0 °C band. Three proportional-integral loops reach 63 % of a step in 0.55, 1.00 and 11.6 min, and a fourth prices a kilowatt-hour of cooling against the worst hour inside its horizon and spends the band on storage. Precooling takes the design day from $615 on a fixed setpoint to $561 at equal comfort, and half of the $24 it saves against a floating band is demand charge. The horizon that produces it is a real optimum with a range between one and five hours where a forecast costs more than none. On a three-hour curtailment two hours of notice moves the energy drawn above the cap from 19.4 to 16.1 kWh while removing the comfort band moves it to 152.3 kWh. A fourth zone control, a receding-horizon plan on the same two nodes per zone with the demand charge in its objective, closes the design day at $557.61 against $561.19, holds both curtailment days at 100 % comfort for $9.73 and $15.36 less, takes the challenge's peak from 234.8 to 204.4 kW, loses the flat-tariff day by $1.76, and is beaten on the design day and the heat wave by the supervisor retuned once. A fuller model would add a latent coil load and a distributed slab. The building shows what a control system can do with the freedom a specification leaves it.

**KEY POINTS**

- An office is judged against a comfort band, so the setpoint inside it is free and the concrete stores cooling.
- Floating in the band saves energy, and using it as a store saves money: $585 against $561 on the design day.
- A forecast horizon shorter than the time to charge the store costs more than no forecast at all.
- Two hours of notice cut a curtailment's overage from 19.4 to 16.1 kWh, while removing the band raises it to 152.3 kWh.
- A 24-hour plan that prices the demand charge beats the shipped rule, $557.61 against $561.19, and a rule retuned once beats the plan.

**Where else this applies.** A data hall's supervisory layer (Chapter 8) also chooses a setpoint that a loop below it holds, and through one day its lowest-cost chilled water setpoint stays between 15.6 and 16.9 °C (Table 8.6).

## References


[1] J. F. Nicol and M. A. Humphreys, "Adaptive thermal comfort and sustainable thermal standards for buildings," *Energy and Buildings*, vol. 34, no. 6, pp. 563–572, 2002.

[2] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[3] A. Arteconi, N. J. Hewitt, and F. Polonara, "State of the art of thermal storage for demand-side management," *Applied Energy*, vol. 93, pp. 371–389, 2012.

[4] A. Roth and J. Reyna, "Grid-interactive efficient buildings technical report series: Whole-building controls, sensors, modeling, and analytics," U.S. Department of Energy, Washington, DC, Tech. Rep. NREL/TP-5500-75478, 2019.

[5] G. Liu, B. Liu, W. Wang, J. Zhang, R. A. Athalye, D. Moser, E. Crowe, N. Bengtson, M. Effinger, L. Webster, and M. Hatten, "Advanced energy retrofit guide: Office buildings," Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. PNNL-20761, 2011.

[6] F. Oldewurtel, A. Parisio, C. N. Jones, D. Gyalistras, M. Gwerder, V. Stauch, B. Lehmann, and M. Morari, "Use of model predictive control and weather forecasts for energy efficient building climate control," *Energy and Buildings*, vol. 45, pp. 15–27, 2012.

[7] Y. Ma, F. Borrelli, B. Hencey, B. Coffey, S. Bengea, and P. Haves, "Model predictive control for the operation of building cooling systems," *IEEE Transactions on Control Systems Technology*, vol. 20, no. 3, pp. 796–803, 2012.

[8] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E. Perarnau Ollé, J. Oravec, M. Wetter, D. L. Vrabie, and L. Helsen, "All you need to know about model predictive control for buildings," *Annual Reviews in Control*, vol. 50, pp. 190–232, 2020.

[9] S. Wang and J. Burnett, "Variable-air-volume air-conditioning systems: Optimal reset of static pressure setpoint," *Building Services Engineering Research and Technology*, vol. 19, no. 4, pp. 219–231, 1998.

[10] L. Pérez-Lombard, J. Ortiz, and C. Pout, "A review on buildings energy consumption information," *Energy and Buildings*, vol. 40, no. 3, pp. 394–398, 2008.

[11] A. Rabl and L. K. Norford, "Peak load reduction by preconditioning buildings at night," *International Journal of Energy Research*, vol. 15, no. 9, pp. 781–798, 1991.

[12] K. Keeney and J. Braun, "A simplified method for determining optimal cooling control strategies for thermal storage in building mass," *HVAC&R Research*, vol. 2, no. 1, pp. 59–78, 1996.

[13] J. E. Braun, K. W. Montgomery, and N. Chaturvedi, "Evaluating the performance of building thermal mass control strategies," *HVAC&R Research*, vol. 7, no. 4, pp. 403–428, 2001.

[14] J. E. Braun, "Load control using building thermal mass," *Journal of Solar Energy Engineering*, vol. 125, no. 3, pp. 292–301, 2003.

[15] K.-h. Lee and J. E. Braun, "Model-based demand-limiting control of building thermal mass," *Building and Environment*, vol. 43, no. 10, pp. 1633–1646, 2008.

[16] P. Bacher and H. Madsen, "Identifying suitable models for the heat dynamics of buildings," *Energy and Buildings*, vol. 43, no. 7, pp. 1511–1522, 2011.

[17] M. H. Albadi and E. F. El-Saadany, "A summary of demand response in electricity markets," *Electric Power Systems Research*, vol. 78, no. 11, pp. 1989–1996, 2008.

[18] P. Xu, P. Haves, M. A. Piette, and J. Braun, "Peak demand reduction from pre-cooling with zone temperature reset in an office building," Lawrence Berkeley National Laboratory, Berkeley, CA, Tech. Rep. LBNL-55800, 2004, in Proceedings of the ACEEE 2004 Summer Study on Energy Efficiency in Buildings. OSTI 838609.

[19] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[20] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[21] G. P. Henze, C. Felsmann, and G. Knabe, "Evaluation of optimal control for active and passive building thermal storage," *International Journal of Thermal Sciences*, vol. 43, no. 2, pp. 173–183, 2004.

[22] M. R. Brambley, N. Fernandez, W. Wang, K. A. Cort, H. Cho, H. Ngo, and J. K. Goddard, "Final project report: Self-correcting controls for VAV system faults; filter/fan/coil and VAV box sections," Pacific Northwest National Laboratory, Richland, WA, Tech. Rep. PNNL-20452, 2011.

[23] G. P. Sullivan, R. Pugh, A. P. Melendez, and W. D. Hunt, "Operations & maintenance best practices: A guide to achieving operational efficiency, release 3.0," Pacific Northwest National Laboratory for the U.S. Department of Energy Federal Energy Management Program, Richland, WA, Tech. Rep. PNNL-19634, 2010.

[24] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication 800-82r3, 2023.

[25] D. Sturzenegger, D. Gyalistras, M. Morari, and R. S. Smith, "Model predictive climate control of a Swiss office building: Implementation, results, and cost–benefit analysis," *IEEE Transactions on Control Systems Technology*, vol. 24, no. 1, pp. 1–12, 2016.

## 9.11 Exercises

**9.1** ○ Concept check · 2 min

Scenario 7 gives the building a 21 to 24 °C band and a forecast horizon of zero. The day costs \$585 against \$615 on a fixed 23 °C setpoint, and the peak falls only from 214 to 202 kW. Why is this saving not the use of the fabric as a store?

(a) The band's ceiling is too close to the setpoint, so the slab cannot warm far enough to hold any cooling

(b) The ramp limit stops the setpoint from falling fast enough to charge the slab before the afternoon

(c) With no horizon the worst cost ahead is the cost now, so the charge demand stays at zero and the afternoon load is met as it arrives

(d) The economizer fills the slab with cool night air, so the store is already full when the band tightens

**9.2** Concept check · 2 min

On scenario 8's day the bill under the band supervisor is \$585 with no forecast, \$620 with a three-hour horizon and \$561 with an eight-hour one (Figure 9.3). Which rule explains the shape of that curve?

(a) The horizon only has to reach the start of the on-peak window, and anything longer is wasted
(b) A longer horizon always lowers the bill, because the supervisor sees more of the price curve
(c) The forecast error grows with the horizon, so the shortest horizon that sees any price rise is the cheapest
(d) The horizon has to cover the time it takes to charge the store, or the charge is bought in a rush that sets a new demand peak

**9.3** Concept check · 2 min

In scenario 10 the building draws 16.1 kWh above the 140 kW cap, 13.4 kWh of it in the event's third hour, and eight hours of notice instead of two still leaves 5.7 kWh. Why can more notice not bring the overage to zero?

(a) With every terminal on its ventilation minimum the coil still cools the required outdoor air, and that load rises above the cap as the afternoon warms
(b) The supervisor stops charging half an hour before the event, so the store is empty by the third hour
(c) The demand meter lags by fifteen minutes, so energy drawn after the event is still counted against the cap
(d) The chiller loses capacity in the afternoon heat, so the chilled water walks up and every valve opens

**9.4** Theory and calculation · 6 min

Use Equation (9.13) and the constants of Table 9.2.

(a) Compute $E_{\max}$ for the 21 to 24 °C band and the storage added by every tenth of a degree of band width. *Checkpoint:* 855.8 kWh, 28.5 kWh per 0.1 K.

(b) At the design point every zone's air is at 23.00 °C and the slabs are at 24.19, 24.19, 24.53 and 25.84 °C. If the slab reference in Equation (9.13) were the band top itself, by how many kilowatt-hours would $E$ fall at this state? *Checkpoint:* About 469 kWh.

(c) A tenant asks for the band to be narrowed to 22 to 24 °C. How much storage does the building lose? *Checkpoint:* About 285 kWh.

**9.5** Theory and calculation · 6 min

The west zone at the design duty passes 6.25 m³/s with $\rho c_p = 1.206\,\mathrm{kJ/(m^3\,K)}$.

(a) From Equations (9.1) and (9.2), compute the west air node's time constant with the slab held, the slab's time constant with the air held, and their ratio. *Checkpoint:* 5.39 min, 333 min, about 62.

(b) The predictive controller's model, Equation (9.21), takes the cooling $Q_z$ as its input instead of the airflow. What air time constant does that model have, and why is it longer? *Checkpoint:* 8.43 min.

(c) Treat the slab as a single lag with the 335 min the chapter measures. What fraction of a step in the air does the slab follow in two hours of charging, and in eight? *Checkpoint:* 30 % and 76 %.

**9.6** Theory and calculation · 7 min

The supervisor charges in proportion to $\sigma = \mathrm{clamp}(K_{pre}(c_{worst}/c_{now} - 1), 0, 1)$, Equation (9.19), with $K_{pre} = 0.20$.

(a) The chiller's coefficient of performance on the design day is 5.43 at 05:00 and 3.10 at 16:00, when the tariff is \$0.075 and \$0.320 per kilowatt-hour. Compute the cost of a kilowatt-hour

of cooling at both hours, their ratio, and how much of the ratio is the tariff and how much the chiller. *Checkpoint:* 1.38 and 10.3 cents, 7.5 = 4.27 × 1.75.

(b) In scenario 10 the ratio $c_{worst}/c_{now}$ is 1.40 just before the notice arrives at 09:00 and 8.85 just after. Compute $\sigma$ at both. *Checkpoint:* 0.08 and 1 (clamped from 1.57).

(c) The fractional saving $K_{pre}(1 - c_{now}/c_{worst})$ was the other candidate. Compute it for the curtailment's 8.85 and for a tariff peak with a ratio of 3, and say why the chapter rejects it. *Checkpoint:* 0.177 and 0.133.

**9.7** App challenge · 8 min

Load scenario 9, "Ninety minutes of forecast, which is worse than none", default seed, as loaded (Band + precooling, Forecast horizon 90 min). Run each day to its end at speed 240 (240 min per second, so a day takes 6 s).

(a) Read the Today, Peak demand and Comfort tiles at $t = 1440$ min. *Checkpoint:* \$602, 222 kW, 100.0 %.

(b) Set Forecast horizon to 0, then 180, then 480 min, each before Run or with Reset, and record Today and Peak demand for each. *Checkpoint:* \$585 and 202, \$620 and 242, \$561 and 188 kW.

(c) The demand charge is \$0.85 per kilowatt of the day's peak. Split each of the four bills into demand charge and energy, and say which part the horizon moves. *Checkpoint:* In the order 90, 0, 180 and 480 min: energy about \$413, 413, 414 and 401, and demand 189, 172, 206 and 160.

Read the bill to ±\$1 and the peak to ±1 kW. The exported CSV columns `metric_energy_usd` and `metric_demand_usd` give the split directly.

**9.8** App challenge · 9 min

Run the curtailment day three ways, default seed, at speed 240, each for the whole 1440 min, and read the Over the cap, Today and Comfort tiles at the end.

(a) Scenario 10, "A curtailment, with two hours of notice", as loaded, then scenario 11, "The same curtailment, with no notice at all", as loaded. *Checkpoint:* 16.1 kWh, \$431 and 97.8 %, then 19.4 kWh, \$437 and 97.7 %.

(b) In scenario 11 open Show Model, set Notice the utility gives (under Parameters) to 480 min, press Reset and run the day again. *Checkpoint:* 5.7 kWh, \$421, 98.5 %.

(c) Scenario 12, "And the same curtailment with no band at all", as loaded. Compare what two hours of notice and what the band are each worth, in kilowatt-hours above the cap and in dollars. *Checkpoint:* 152.3 kWh, \$614, 100.0 %.

Read the overage to ±0.1 kWh, the bill to ±\$1 and comfort to ±0.1 %.

**9.9** App challenge · 9 min

Load scenario 15, "The same day, planned: a 24 h MPC against the priced rule", default seed, as loaded (Zone control MPC, Plan horizon 24 h, Preview On). Run each day to $t = 1440$ min at speed 240.

(a) Read Today, Peak demand and Comfort, and the last row of the exported CSV columns `metric_energy_usd` and `metric_demand_usd`. *Checkpoint:* \$558, 183 kW and 100.0 %, with \$402.47 and \$155.14 in the file.

(b) Press Reset, set Zone control to Band + precooling and run the day. Read the same five values and say where the plan's \$3.58 comes from. *Checkpoint:* \$561, 188 kW and 100.0 %, with \$401.46 and \$159.73 in the file.

(c) Set Zone control back to MPC, set Plan horizon (under Predictive controller in Advanced) to 3 h, press Reset and run. Compare with the supervisor's \$620 and 242 kW at a three-hour horizon. *Checkpoint:* \$579, 200 kW.

Read the bill to $\pm$\$1, the peak to $\pm$1 kW and the CSV values to $\pm$\$0.05.

**9.10** Application area · 4 min

A facilities manager receives complaints about warm afternoons and proposes two changes to the lease: narrow the comfort band from 21–24 to 22–24 °C, and lower the ceiling allowed during a utility event from 25.5 to 24.5 °C.

(a) What does each change cost on the graded challenge day, as the chapter measures it from the suggested configuration? *Checkpoint:* 5.9 kWh above the cap, then 28.0 kWh and \$665.

(b) Whose decision is the band width, and what should the engineer bring to that decision? *Checkpoint:* The tenants' and the owner's, in the lease.

**9.11** Application area · 3 min

The challenge asks for the dry-bulb economizer to be switched on.

(a) Why is it worth exactly nothing on that day? *Checkpoint:* The outdoor air never falls below 20.0 °C.

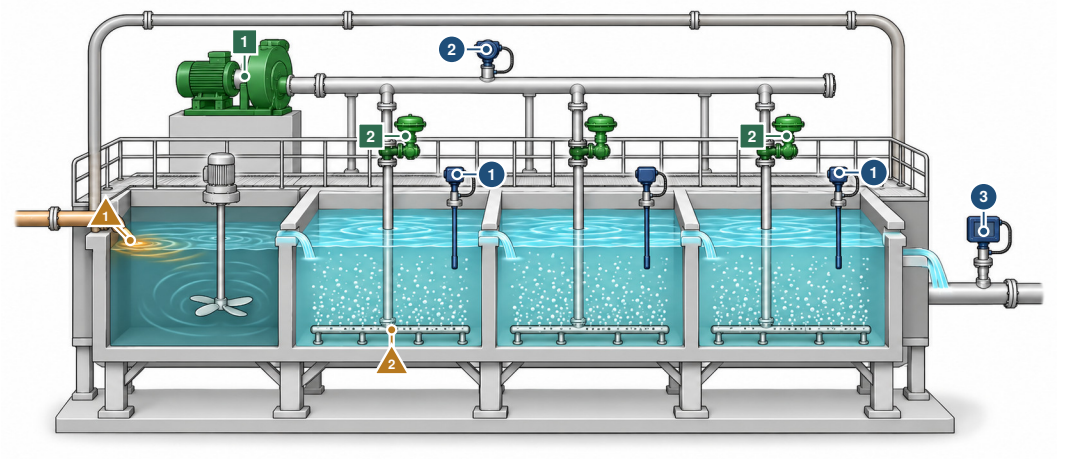
(b) The model has no moisture. Why would the same 18 °C dry-bulb changeover be a poor choice in a humid climate? *Checkpoint:* Latent load, so an enthalpy changeover.

**9.12** Review the agent · 5 min

An AI assistant summarizes scenario 15: "The 24-hour MPC closes the design day at \$557.61 against \$561.19 for the band supervisor, so a plan beats a priced rule on this building. Replace the supervisor with the MPC and the bill will fall on every day like this one." Find the flaw and say how you would check it in the app.

CHAPTER 10

# Wastewater Aeration Simulator for Supervisory Setpoint Selection Under a Discharge Permit




**The controller moves** 1 blower and its speed drive 2 three air valves, one per tank

**The controller reads** 1 dissolved oxygen, one probe per tank 2 air header pressure 3 effluent ammonium analyzer

**What disturbs it** 1 sewage load that follows the day 2 diffuser fouling

*One blower and one header serve three tanks in series, while the ammonia that the permit counts leaves the last tank hours after the air that removed it. Illustration, not to scale.*

IN PLAIN TERMS **Buying Only the Air the Permit Needs**

**The main idea.** Bacteria in large tanks eat the ammonia in sewage, and blowers feed them air. The plant is judged on the ammonia that leaves, averaged over eight hours. The air is a large part of its power bill.

**The challenge.** The ammonia that leaves answers a change of air only hours later. So the air has to be chosen ahead of time. A planner that knows the usual shape of the day saves air, but it lets out more ammonia.

**Why it matters.** On a settled day the planner uses 6.5 percent less power than a simple feedback controller, at about the same average ammonia.

THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 10.2, the facts in the Introduction and in Section 10.5.7, the try-it segments in Section 10.5, and the key points |
| **Study** | 70 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 25 min | the first page, the results (Section 10.5), Technology in Practice (Section 10.6), Limitations (Section 10.7) and the open challenge |

**The control question** An activated sludge plant is judged on an eight-hour mean of effluent ammonia that answers the air in about three hours, while its control system holds a dissolved oxygen setpoint that appears in no permit. Where should that setpoint sit hour by hour, and what is moving it worth at equal compliance, on the first day and on a settled day?

**Learning objectives**

- Compute the oxygen saturation, the nitrifier population and the header time constant of the works from its model equations.
- Explain why a cold snap moves the effluent ammonia and not the nitrifier population, and why the sludge age is the slow handle.
- Compare a fixed setpoint, the ammonia supervisor and most-open-valve control at equal compliance, and say why the first day overstates the supervisor.
- Evaluate the day-ahead MPC against the supervisor on energy, worst rolling mean and nitrogen discharged, with and without the load preview.
- Justify a supervisory investment on this plant in dollars a year.

**Prerequisites** Sections 1.6.2 (velocity form), 1.4.3 (receding horizon), 1.4.6 (soft constraints), 1.4.7 (back-off), 1.4.8 (economic MPC) and 1.6.4 (how an exercise names a run), and Volume 3, Chapter 3 for a dissolved oxygen cascade.

**The app** https://apmonitor.com/apps/wastewater-aeration/

**See also** Volume 4, Chapter 9 for valve-position control on a shared supply and Volume 4, Chapter 2 for the surge line and its override. Volume 19, Chapter 11 and Volume 19, Chapter 8 for supervisory loops that act on slow, sampled measurements of a living population.

## 10.1 Introduction

Every regulatory loop on the simulated works of this chapter behaves itself. The air header pressure settles in 17.3 s, the dissolved oxygen in each aerated tank settles in 6.20 min, and both are ordinary proportional integral loops that a first course covers [1]. The hard question is not how to hold a dissolved oxygen setpoint. It is what that setpoint ought to be, hour by hour, when the quantity that can be measured in 45 s carries no legal force and the quantity the discharge permit is written about responds in about three hours and is reported as an eight-hour rolling mean.

That gap is what makes the plant worth simulating rather than drawing. A block diagram of a cascade shows an outer loop writing an inner setpoint and says nothing about whether the outer loop is worth installing, because the answer depends on the size of a bill, the shape of a load curve and the wording of a limit. The misconception this application corrects is quantitative rather than structural. Aeration is 50 to 60 percent of a treatment works' electricity [2], so a control change that touches aeration sounds as though it should be worth tens of percent. At 5 m of diffuser submergence the oxygen saturation the bubbles work against is 12.39 mg/L, so a 2 mg/L dissolved oxygen setpoint removes only 16.1 % of the transfer driving force, and the entire practical setpoint range from 1.2 to 3.0 mg/L spans 32 % of a day's aeration energy and no more. Everything a setpoint policy can win has to come out of that third. Read at equal compliance, load-following the setpoint against an ammonia analyzer is worth 194 kWh out of 4446 kWh on the first day of operation and 66 kWh out of 4355 kWh on a settled day, while a most-open-valve controller on the same air header is worth 267 kWh out of 4335 kWh and gives up nothing at all.

**DID YOU KNOW? Water, Wastewater and the Aeration Share**

Drinking water and wastewater systems account for about 3 to 4 percent of United States energy use, and the energy of water and wastewater utilities is 35 percent of a typical municipal energy budget [3]. Inside a treatment works, aeration can take from 25 to as much as 60 percent of total plant energy [4].

The class of problem is general. Whenever a plant carries a cheap fast measurement that is easy to control and an expensive slow measurement that decides whether the plant is compliant or profitable, the engineering question moves from regulation to the choice of the controlled variable and its setpoint, which is the subject of self-optimizing control [5]. Blend ratio against a laboratory property, furnace excess oxygen against a stack analyzer, reactor temperature against a product specification and coagulant dose against a settled turbidity (Volume 3, Chapter 6) all have this shape, and in each case a supervisory loop with an integral time measured in hours sits above a regulatory loop measured in minutes [1]. The second structure this plant teaches is equally portable. Three throttled users share one air header, so the header pressure is a free variable that no quality measurement constrains, and walking it down until the busiest valve is nearly open removes a pressure drop that was being manufactured and then destroyed. Valve position control of that kind applies to chilled water, instrument air and fuel gas headers with no change of form [6].

However, the published case for ammonia-based aeration control is usually made on a trial lasting days to weeks, and a supervisory loop with an eight-hour integral time takes about a day to forget the setpoint it was commissioned from. The application measures both numbers. This chapter prices a supervisory dissolved oxygen setpoint on a simulated activated sludge plant against a legal limit and an electricity bill, so that the value of load-following aeration is measured at equal compliance and on a settled day rather than quoted from a first-day trial. The equal-compliance accounting finds 4 % on day one and 1.5 % on day three from a configuration nobody touched in between, and it finds the header loop nobody markets ahead of the marketed analyzer loop by a factor of four. The kinetics, the transfer correlation and the tuning arithmetic are all standard.

Section 10.2 traces the history that produced the current answer. Section 10.3 sets out a sixteen-state reduced activated sludge model with its air system, its blower surge line, its instrument chain and its rolling-mean permit, closely enough to rebuild. Section 10.4 describes the four proportional integral loops and the day-ahead predictive supervisor that can replace the ammonia loop. Section 10.5 measures fifteen scenarios, from three valves in manual to a graded February day. Section 10.6 describes the probes, blowers and controllers of real works, Section 10.7 states what the model leaves out, Section 10.8 separates the three decisions the plant carries, and the exercises of Section 10.10 run the same scenarios.

## 10.2 Where Aeration Control Came From

The activated sludge process was reported in 1914 as a laboratory observation that sewage aerated in the presence of accumulated solids nitrified within a few hours instead of a few days, and the same paper recorded that the accumulated sludge had to be retained between batches for the effect to survive [7]. Retention of solids is the whole of the mechanism, and it is the reason the plant of this chapter has a state with a time constant of days.

Air was originally supplied at whatever rate the blowers ran. The first control action available to an operator was the number of machines in service, and the operating rule was a visual one, because there was no instrument in the tank. Nitrification design became quantitative when growth kinetics were attached to the population. The Monod expression for a specific growth rate limited by a single substrate [8] supplied the form, and the recognition that the sludge age fixes the average growth rate a population is permitted to have converted nitrification from an observation into a design calculation with a safety factor on solids retention time [9]. The consolidation of that work into the activated sludge model family gave a common structure for the kinetics, the stoichiometry and the temperature corrections used in every simulator since [10].

Dissolved oxygen control arrived with the membrane probe. Once a probe could report a tank concentration continuously, the airflow could be trimmed against a setpoint rather than staged, and the whole subsequent history of instrumentation, control and automation on treatment works is organized around which measurement became available when [11]. Two consequences followed quickly. The first is that a dissolved oxygen setpoint is a proxy rather than an objective, because nothing in a discharge permit mentions it. The second is that holding a proxy perfectly can hide a fault, since a controller that opens a valve to compensate for fouled diffusers reports no error at all. Measurement of the alpha factor and its collapse over a diffuser grid's service life made the size of that hidden cost explicit and put oxygen transfer efficiency, rather than dissolved oxygen, at the center of aeration energy work [12, 13]. Modeling the air side as a system of blowers, headers and valves rather than as an ideal source of oxygen followed for the same reason, because the pressure the blower makes and the valves throttle away is invisible on a dissolved oxygen trend [14].

Ammonium analyzers reaching the reliability needed for closed-loop service is the change that produced the current answer. Once effluent or end-of-train ammonia could be read every few minutes, the dissolved oxygen setpoint could be moved against the variable the permit is written about rather than held at a value chosen for the worst hour, and aeration control reviews from the last fifteen years treat that supervisory layer as the standard modern structure [15]. Optimization studies over a diurnal load place the achievable saving in the single digits of percent rather than the tens once compliance is held constant [16], and practice guidance distinguishes the feedback form used here from a feedforward form that anticipates the load [17]. Energy benchmarking of operating works supplies the other half of the argument, because the specific electricity consumption that distinguishes a good plant from a poor one is dominated by equipment condition and load factor rather than by setpoint policy [2].

> HISTORY **2003: Smaller Blowers and Most-Open-Valve Logic**
>
> A federal survey of energy measures at treatment works tells of a medium-sized city's activated sludge plant whose organic load had fallen well below its design. With only one of its inlet-throttled centrifugal blowers running, the dissolved oxygen still sat between 4.5 and 8.0 mg/L, because the machine could not turn down far enough. In July 2003 the plant replaced two of its blowers with units of half the power, fitted new membrane oxygen probes, and replaced single-loop oxygen control with direct airflow control carrying most-open-valve logic. The change saved about 1,000,000 kWh a year, about 9 percent of the plant's energy, and paid back in about three years [4]. Scenario 10 of this chapter meets the same turndown limit at night, and scenario 9 prices the same header logic.

## 10.3 Process Model

The simulated works treats 20 000 $m^3$/d of separately sewered municipal sewage carrying 600 kg/d of ammoniacal nitrogen. Screened influent enters an anoxic zone of 1500 $m^3$ stirred by 15 kW of mixers, passes through three aerated tanks of 1700 $m^3$ each and leaves by a final channel carrying the ammonium analyzer. Mixed liquor is returned to the head of the anoxic zone at 3.00 times the influent flow from the end of the aerobic train and at 0.75 times it as return activated sludge, so the anoxic zone and each aerated tank see 65.97 $m^3$/min while the sewage itself crosses the basin in 7.92 h. Air is blown into the three tanks through fine-bubble diffusers 5 m under water, one butterfly valve per tank, from a common header of 250 $m^3$ fed by a single variable-speed centrifugal blower. The arrangement, the instrument tags and each symbol the equations use are drawn in Figure 10.1.

The clock is in minutes throughout, which is the only base in which a 17.3 s header transient and a four-day biomass recovery are both legible. Every kinetic rate is quoted per day on its slider and converted once, in the parameter declaration, by a factor of 1440. Concentrations are in g/$m^3$, which equals mg/L, flows in $m^3$/min, air in standard cubic meters per minute at 1.20 kg/$m^3$, pressures in kPa gauge except inside the blower work expression, and power in kW. Table 10.1 sorts the controlled,

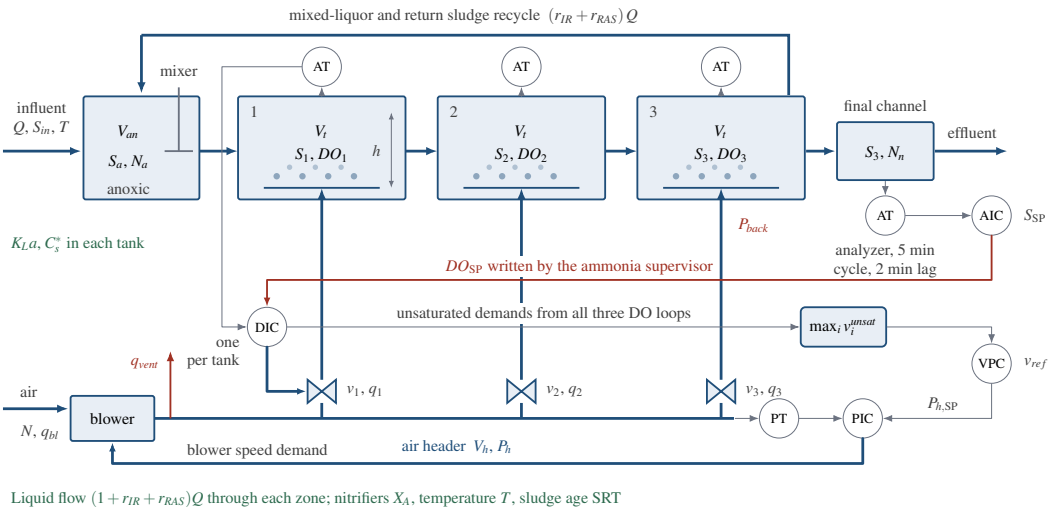



**Figure 10.1** The simulated works and every symbol used in the equations. Influent $Q$ at strength $S_{in}$ joins a recycle of $(r_{IR}+r_{RAS})Q$ in the anoxic zone, then crosses three aerated tanks in series. Each tank has its own diffuser grid $h$ below the surface, its own butterfly valve $v_i$ passing $q_i$ from the header at gauge pressure $P_h$, and its own dissolved oxygen probe. The blower speed $N$ holds the header against the diffuser back pressure $P_{back}$, and $q_{vent}$ is the air blown off to keep the machine off its surge line. The ammonium analyzer on the final channel measures $S_3$ and drives the dissolved oxygen setpoint written to all three tank controllers.

**Table 10.1** What the Control System Interacts With, at the Design Condition of 13 °C and Average Load. A Range Quoted Against a Controlled Variable Is the Travel of Its Setpoint.

| Symbol | Role | Nominal | Range |
|---|---|---|---|
| $DO_1$ | CV, measured | 2.00 mg/L | 0.3–4.0 mg/L |
| $S_3$ | CV, measured | 0.378 g N/m$^3$ | 0.3–4.0 g N/m$^3$ |
| $P_h$ | CV, measured | 62.0 kPa | 55–95 kPa |
| $v_{\max}$ | CV, inferred | 68.9 % | 70–98 % |
| $v_1, v_2, v_3$ | MV | 68.9/50.0/31.7 % | 0–100 % |
| $N$ | MV | 59.5 % | 48–100 % |
| $DO_{SP}$ | MV of AIC | 2.00 mg/L | 0.5–3.0 mg/L |
| $P_{h,SP}$ | MV of VPC | 62.0 kPa | 55–95 kPa |
| $T$ sewage | DV, measured | 13.0 °C | 6–24 °C |
| $Q$ multiplier | DV, measured | 1.00× | 0.60–2.50× |
| $S_{in}$ multiplier | DV, measured | 1.00× | 0.50–2.00× |
| diurnal swing | DV, measured | 1.00× | 0–1.50× |
| $\alpha F$ | DV, unmeasured | 0.55 | 0.30–0.85 |
| carbon load | DV, unmeasured | 1.00× | 0.60–1.60× |
| $S_a$ | state | 5.86 g N/m$^3$ | $\geq 0$ |
| $S_1, S_2, S_3$ | state | 3.39/1.40/0.378 g N/m$^3$ | $\geq 0$ |
| $N_a, N_n$ | state | 9.55/15.02 g N/m$^3$ | $\geq 0$ |
| $DO_1, DO_2, DO_3$ | state | 2.00/2.00/1.80 mg/L | 0–20 mg/L |
| $X_A$ | state | 101.6 g VSS/m$^3$ | $\geq 0$ |
| $P_h$ | state | 62.0 kPa | 0–200 kPa |
| $N$ | state | 59.5 % | 0–100 % |
| $v_1, v_2, v_3$ | state | 68.9/50.0/31.7 % | 0–100 % |
| $T_w$ basin | state | 13.0 °C | 6–24 °C |

manipulated, disturbance and state variables at the design condition every steady scenario opens from.

### 10.3.1 Load, Nitrogen and the Nitrifier Population

Influent flow and strength follow a measured diurnal shape rather than a sine, because the argument for load-following depends on the day being both predictable and lopsided. Twenty-four hourly factors are interpolated with a Catmull-Rom spline so that the load has no corner in it, the strength peaks about an hour after the flow does, and the two multiplied give an ammoniacal nitrogen load running from 222 kg/d at 03:53 to 1032 kg/d at 09:54 about a 600 kg/d mean. Ammonia is carried through four completely mixed zones in series, as shown in Equations (10.1) and (10.2).

$$V_{an}\frac{\mathrm{d}S_a}{\mathrm{d}t} = QS_{in}(1-f_{ass}) + Q_r S_3 - Q_{an}S_a \tag{10.1}$$

$$V_t\frac{\mathrm{d}S_i}{\mathrm{d}t} = Q_{an}(S_{i-1}-S_i) - \frac{\mu_A(T,S_i,DO_i)}{Y_A}X_A V_t \tag{10.2}$$

where $Q$ is the influent flow in $\mathrm{m^3/min}$ at ammoniacal strength $S_{in}$ in $\mathrm{g\,N/m^3}$, $V_{an}$ and $V_t$ are the anoxic and aerated tank volumes in $\mathrm{m^3}$, $S_a$ is the anoxic ammonia, $S_i$ is the ammonia in aerated tank $i$ with $S_0 \equiv S_a$, $Q_r = (r_{IR}+r_{RAS})Q$ is the total recycle, $Q_{an} = Q + Q_r$ is the flow through every zone, $f_{ass}$ is the influent nitrogen fraction assimilated into heterotrophic cells, $Y_A$ is the autotroph yield in g VSS per g N and $X_A$ is the nitrifier concentration in $\mathrm{g\,VSS/m^3}$. Both sides of (10.1) carry grams of nitrogen per minute. The specific growth rate is Monod in ammonia and in oxygen with an Arrhenius correction on the maximum, as shown in Equation (10.3).

$$\mu_A = \hat{\mu}_{A,20}\,\theta_A^{T-20}\,\frac{S}{K_{NH}+S}\,\frac{DO}{K_{O,A}+DO} \tag{10.3}$$

where $\hat{\mu}_{A,20}$ is the maximum specific growth rate at 20 °C in 1/min, $\theta_A$ is the temperature coefficient, $K_{NH}$ is the ammonia half-saturation constant and $K_{O,A}$ is the oxygen half-saturation constant [8, 10]. The oxygen factor in (10.3) is the entire justification for holding a dissolved oxygen setpoint, and because it saturates the justification runs out. With $K_{O,A} = 0.5\,\mathrm{mg/L}$ the factor is 50 % at 0.5 mg/L, 67 % at 1 mg/L, 80 % at 2 mg/L, 86 % at 3 mg/L and 90 % at 4.5 mg/L.

The population is a state in its own right, growing at the mean of the three tank rates, decaying, and removed by wasting at the reciprocal of the solids retention time, as shown in Equation (10.4).

$$\frac{\mathrm{d}X_A}{\mathrm{d}t} = \left(\bar{\mu}_A - b_A(T) - \frac{1}{\mathrm{SRT}}\right)X_A, \qquad b_A = b_{A,20}\,\theta_B^{T-20} \tag{10.4}$$

where $\bar{\mu}_A$ is the mean of the three tank growth rates, $b_{A,20}$ is the decay rate at 20 °C with temperature coefficient $\theta_B$, and SRT is the solids retention time in minutes. Setting the bracket in (10.4) to zero gives the least intuitive property of the plant. At steady state the wasting rate, not the temperature, fixes the average growth rate the nitrifiers are permitted to have, so $\bar{\mu}_A = b_A(T) + 1/\mathrm{SRT}$ moves only from 0.2411/d at 13 °C to 0.2305/d at 9 °C on a seven-day sludge age, against a maximum rate that falls from 0.4917 to 0.3723/d over the same interval. The steady population is therefore 101.6 g VSS/m³ at 13 °C and 102.2 g VSS/m³ at 9 °C, and the effluent ammonia carries the whole of the loss by rising from 0.378 to 1.378 g N/m³. Cutting the sludge age from 7 to 5 days is what actually moves the population, to 79.5 g VSS/m³ in the steady state and through 96.5, 92.6, 89.7, 87.4 and 85.7 g VSS/m³ on the first five days after the decision.

Nitrate is carried as two pools, one in the anoxic zone and one for the whole aerobic train, since nitrate is conserved along the train and nothing consumes it there. Denitrification in the anoxic zone is carbon-limited and scales with the raw sewage arriving rather than with the nitrate on offer, as shown in Equations (10.5) and (10.6).

$$V_{an}\frac{\mathrm{d}N_a}{\mathrm{d}t} = Q_r N_n - Q_{an}N_a - V_{an}r_{DN} \tag{10.5}$$

$$r_{DN} = k_{DN,20}\,\theta_{DN}^{T-20}\,\frac{N_a}{K_{NO3}+N_a}\,\frac{QS_{in}}{\bar{Q}\bar{S}_{in}} \tag{10.6}$$

where $N_a$ and $N_n$ are the anoxic and aerobic nitrate concentrations in g N/m$^3$, $r_{DN}$ is the denitrification rate in g N/m$^3$ per minute, $k_{DN,20}$ is that rate at 20 °C with temperature coefficient $\theta_{DN}$, $K_{NO3}$ is the nitrate half-saturation constant and $\bar{Q}\bar{S}_{in}$ is the daily mean nitrogen load. The aerobic pool $N_n$ is fed by $Q_{an}(N_a - N_n)/(3V_t)$ plus the mean nitrification rate of the three tanks. At the design condition the anoxic zone holds 9.55 g N/m$^3$ of nitrate and the aerobic train 15.02 g N/m$^3$.

### 10.3.2 Oxygen Transfer, the Air Header and the Blower

Dissolved oxygen in each tank is a balance of what the upstream zone brings, what the diffusers transfer and what the biology consumes, as shown in Equation (10.7).

$$\frac{\mathrm{d}DO_i}{\mathrm{d}t} = \frac{Q_{an}}{V_t}\,(DO_{i-1} - DO_i) + K_L a_i\,(C_s^{*} - DO_i) - R_i \tag{10.7}$$

where $DO_0 \equiv 0$ because liquor arrives from the anoxic zone with no oxygen in it, $R_i$ is the total oxygen uptake rate in that tank in g/(m$^3$ min), and $C_s^{*}$ is the saturation concentration the bubbles see. Omitting the recycle term in (10.7) understates the air tank 1 needs by 6.5 %, because at a combined recycle of 3.75 the oxygen-free stream entering tank 1 is nearly four times the influent. The transfer coefficient is a power law in diffuser flux and the saturation carries half the submergence as static head, as shown in Equations (10.8) and (10.9).

$$K_L a_i = \alpha F\, k_{La,0} \left(\frac{q_i}{V_t}\right)^{n} \tag{10.8}$$

$$C_s^{*} = \beta\, C_{s,w}(T) \left(1 + \frac{\rho g h}{2 P_{atm}}\right) \tag{10.9}$$

where $q_i$ is the air to tank $i$ in Nm$^3$/min, $n = 0.75$ is below unity because the bubbles coalesce and rise faster as the flux rises, $\alpha F$ lumps the mixed-liquor alpha factor with diffuser fouling, $\beta$ is the salinity correction and $C_{s,w}$ is a cubic fit to the clean-water solubility tables accurate to 0.15 g/m$^3$ between 5 and 30 °C [12]. At 5 m and 13 °C, (10.9) returns 12.39 mg/L against 10.50 mg/L for the same water at the surface. Inverting (10.8) and (10.7) at steady state gives $q_i \propto [R_i/(C_s^{*} - DO_i)]^{1/n}$, so the air rises faster than linearly in the demand and faster than linearly in the dissolved oxygen demanded. At the design load, holding 3 mg/L instead of 2 mg/L costs 17.3 % more air and holding 1 mg/L saves 14.7 %, and at constant oxygen uptake the air per additional milligram is 1.97 times as steep at 3.5 mg/L as at 0.5 mg/L.

Each valve is an orifice in series with an inherent exponential characteristic, and the header is the only capacity in the air system, as shown in Equations (10.10) and (10.11).

$$q_i = C_v\, f(v_i) \sqrt{P_h - P_{back}}, \qquad f(v) = \frac{\mathrm{e}^{R_v v/100} - 1}{\mathrm{e}^{R_v} - 1} \tag{10.10}$$

$$\frac{V_h}{P_{atm}} \frac{\mathrm{d}P_h}{\mathrm{d}t} = q_{bl}(N, P_h) - \textstyle\sum_i q_i - q_{vent} \tag{10.11}$$

where $C_v$ is the wide-open valve capacity, $R_v$ is the rangeability exponent, $\rho$ is the liquor density, $g$ is gravitational acceleration, $P_{back} = \rho g h + \Delta P_{dif} = 53.05$ kPa is the pressure needed before any air reaches the water, and $V_h$ is the header volume. With 250 m$^3$ of pipework, one Nm$^3$/min of imbalance moves the header 0.405 kPa/min. Both terms on the right of (10.11) push back on pressure, at −1.72 Nm$^3$/min per kPa from the blower curve and +6.82 Nm$^3$/min per kPa from the three valves at the design point, so the closed balance has a time constant of 17.3 s and a gain of 0.660 kPa per percent of blower speed. That is a factor of 21 faster than the 6.20 min of a tank, and the ratio is free rather than tuned. The fermenter of Volume 3, Chapter 3 is the opposite case, with its lags within a factor of two of one another, so there the speed of the inner loop has to be earned by tuning.

The blower is a single variable-speed centrifugal machine with a fitted quadratic head curve, a fan law and a surge line, as shown in Equations (10.12) and (10.13).

$$q_{bl} = \sqrt{\frac{a_0 (N/100)^2 - P_h}{a_2}} \ \text{when positive, otherwise } 0 \tag{10.12}$$

$$q_{vent} = k_{vent} \max\left(0,\ \frac{N}{100} q_{surge} - q_{bl}\right) \tag{10.13}$$

where $a_0$ is the shut-off head at full speed, $a_2$ is the droop of the fitted head curve, $q_{surge}$ is the surge flow at full speed and $k_{vent}$ is the anti-surge valve gain. Equation (10.13) is a tight proportional anti-surge controller collapsed into the flow balance, and every standard cubic meter it discards has been compressed and paid for. Volume 4, Chapter 2 treats surge itself and the protective override that keeps a machine off it. Shaft and motor power follow the adiabatic compression of the delivered plus vented air, as shown in Equation (10.14).

$$W = \frac{\dot{m}\, c_p T_1}{\eta_{is}\, \eta_m} \left[ \left( \frac{P_{atm} + P_h}{P_{atm} - \Delta P_f} \right)^{0.283} - 1 \right] \tag{10.14}$$

where $W$ is the shaft and motor power in kW, $\dot{m} = \rho_{air}(q_{bl} + q_{vent})/60$ in kg/s, $c_p = 1005\,\mathrm{J/(kg\,K)}$ and $T_1$ is the inlet temperature. At the design condition (10.14) returns 158.9 kW for 122.1 $\mathrm{Nm^3/min}$ against 62 kPa, which is 1.64 kg of oxygen dissolved per kWh and 12.8 % transfer efficiency, both realistic field values for a five-meter fine-bubble grid in mixed liquor [13]. Adding the 15 kW of anoxic mixers and dividing by an aeration share of 0.55 gives a site draw of 316 kW, or 0.379 $\mathrm{kWh/m^3}$ treated.

### 10.3.3 The Permit, Which Is a Mean and Not a Fence

The discharge limit is an eight-hour rolling mean of effluent ammonia, held as a ring of 96 five-minute bins rather than as an exponentially weighted average, as shown in Equation (10.15).

$$\bar{S}_e(t) = \frac{1}{T_p} \int_{t-T_p}^{t} S_3(\tau)\, \mathrm{d}\tau \ \le\ S_{limit}, \qquad S_3 \le S_{max} \tag{10.15}$$

with $T_p = 480\,\mathrm{min}$, $S_{limit} = 2.5\,\mathrm{gN/m^3}$ and an absolute upper tier $S_{max} = 5.0\,\mathrm{gN/m^3}$ that is not averaged. A ring of real bins reproduces both halves of a rolling mean, because a morning excursion stays on the record until it falls out of the window and a quiet afternoon genuinely pays for it. That property is what makes a load-following setpoint legitimate rather than merely cheap, and the upper tier exists because a plant could otherwise average its way through anything.

### 10.3.4 Constants, the Opening State and the Numerics

The constants are collected in Table 10.2. Most can be reached from the parameter panel, five cannot, and that split matters because the hidden five are the ones a student has to infer from a test rather than read off a screen.

No scenario opens on a transient. The opening state is solved for rather than approached, and the population comes from bisection on the residual $\bar{\mu}_A(X_A) - b_A - 1/\mathrm{SRT}$ at the daily average load, because a plant that has been running since March does not start the day at whatever the current hour happens to demand. The ammonia chain is then swept to convergence at the load actually arriving at $t = 0$, since the anoxic zone is fed by a recycle that comes from the end of the train and the loop has to be iterated rather than marched. Valve positions, header pressure and blower speed are then solved from the oxygen demand so that (10.7) and (10.11) are stationary. The application's test suite verifies this by evaluating every derivative at the opening state and requiring the largest relative value to be below $10^{-6}$.

**Table 10.2** Constants of the Model. The Five Marked Hidden Never Reach the Screen. Kinetic Rates Are Held per Minute and Presented per Day.

| Symbol | Meaning | Value | Unit | Visibility |
|---|---|---|---|---|
| $\bar{Q}$ | dry-weather average flow | 13.889 | $m^3$/min | adjustable |
| $\bar{S}_{in}$ | average influent ammonia | 30.0 | g N/$m^3$ | adjustable |
| $f_{ass}$ | influent N assimilated | 0.12 | — | adjustable |
| $V_{an}$ | anoxic zone volume | 1500 | $m^3$ | adjustable |
| $V_t$ | volume of each aerated tank | 1700 | $m^3$ | adjustable |
| $h$ | diffuser submergence | 5.0 | m | adjustable |
| $r_{IR}$ | internal recycle ratio | 3.00 | $\times Q$ | adjustable |
| $r_{RAS}$ | return sludge ratio | 0.75 | $\times Q$ | adjustable |
| $\hat{\mu}_{A,20}$ | maximum autotroph growth | 0.80 | 1/d | adjustable |
| $b_{A,20}$ | autotroph decay | 0.12 | 1/d | adjustable |
| $\theta_A$ | growth temperature coefficient | 1.072 | — | adjustable |
| $K_{NH}$ | ammonia half-saturation | 0.70 | g N/$m^3$ | adjustable |
| $K_{O,A}$ | oxygen half-saturation | 0.50 | mg/L | adjustable |
| $Y_A$ | autotroph yield | 0.24 | g VSS/g N | adjustable |
| SRT | solids retention time | 7.0 | d | adjustable |
| $k_{DN,20}$ | anoxic denitrification rate | 0.18 | g N/$m^3$/min | adjustable |
| $K_{NO3}$ | nitrate half-saturation | 0.30 | g N/$m^3$ | adjustable |
| $\theta_{DN}$ | denitrification coefficient | 1.08 | — | adjustable |
| $\beta$ | salinity correction | 0.95 | — | adjustable |
| $k_{La,0}$ | transfer correlation coefficient | 2.50 | 1/min | adjustable |
| $n$ | airflow exponent | 0.75 | — | adjustable |
| $R_{c,0}$ | carbonaceous demand at mean load | 2656 | g/min | adjustable |
| $f_{endog}$ | endogenous share of that demand | 0.25 | — | adjustable |
| $V_h$ | air header volume | 250 | $m^3$ | adjustable |
| $\Delta P_{dif}$ | diffuser pressure loss | 4.0 | kPa | adjustable |
| $C_v$ | valve capacity, wide open | 55.0 | $Nm^3/min/\sqrt{kPa}$ | adjustable |
| $R_v$ | valve rangeability exponent | 2.5 | — | adjustable |
| $\tau_v$ | air valve travel time | 0.25 | min | adjustable |
| $a_0$ | blower shut-off head | 275.0 | kPa | adjustable |
| $a_2$ | blower curve droop | 0.00238 | kPa $min^2/Nm^6$ | adjustable |
| $q_{surge}$ | surge flow at full speed | 82 | $Nm^3$/min | adjustable |
| $N_{\min}$ | minimum blower speed | 48 | % | adjustable |
| $\tau_{bl}$ | blower speed ramp | 0.20 | min | adjustable |
| $\eta_{is}$ | isentropic efficiency | 0.72 | — | adjustable |
| $\eta_m$ | motor and drive efficiency | 0.95 | — | adjustable |
| $\Delta P_f$ | inlet filter and silencer loss | 2.0 | kPa | adjustable |
| $W_{mix}$ | anoxic zone mixers | 15 | kW | adjustable |
| $\tau_{DO}$ | DO probe time constant | 0.75 | min | adjustable |
| $\tau_{Ph}$ | header transmitter lag | 0.05 | min | adjustable |
| $t_{cyc}$ | analyzer cycle time | 5.0 | min | adjustable |
| $\theta_{NH}$ | analyzer transport lag | 2.0 | min | adjustable |
| $\tau_T$ | basin thermal lag | 480 | min | adjustable |
| $T_p$ | permit averaging window | 480 | min | adjustable |
| $S_{limit}$ | rolling-mean permit | 2.5 | g N/$m^3$ | adjustable |
| $S_{max}$ | absolute upper-tier limit | 5.0 | g N/$m^3$ | adjustable |
| $\eta_{aer}$ | aeration share of site power | 0.55 | — | adjustable |
| $\theta_B$ | decay temperature coefficient | 1.029 | — | hidden |
| $Y_{O/N}$ | oxygen per unit N oxidized | 4.33 | g $O_2$/g N | hidden |
| $k_{vent}$ | anti-surge valve gain | 2.0 | — | hidden |
| $T_1$ | blower inlet temperature | 293 | K | hidden |
| $\rho_{air}$ | standard air density | 1.20 | kg/$m^3$ | hidden |

The engine of Section 1.6.1 advances the physics under a fourth-order Runge-Kutta integrator whose step is frozen at 0.02 min (1.2 s). Controllers fire every 0.1 min (6 s), which is the scan a works distributed control system actually runs, and two samples of everything are logged per simulated minute, so one day of plant is 72 000 steps and 2880 records. The default seed is 20 260 914.

### 10.3.5 Assumptions

*Four completely mixed zones and no clarifier.* There is no dispersion model, no dead volume and no settling tank. Return activated sludge is a flow ratio and the mixed-liquor solids are a parameter, so the application cannot show sludge bulking, solids washout in a storm, or any of the settleability problems that actually close plants down.

*One lumped autotroph population.* Real nitrification is two organisms in series, ammonia to nitrite and nitrite to nitrate, and nitrite accumulation under transient low-oxygen conditions is a documented failure mode this structure cannot represent. Nitrous oxide, which is where a genuinely low dissolved oxygen setpoint stops being free, is not modeled at all [18]. The application enforces a setpoint floor as a control limit instead, which is what a real plant does but not why.

*Heterotrophic oxygen demand is a forcing, not a state.* The carbonaceous demand is scaled by the arriving load and split between the three tanks in fixed proportions of 0.50, 0.30 and 0.20, so it responds to the influent instantly where a real basin's stored substrate would smear it over an hour or two. This application's morning air peak is therefore slightly sharper than a real one.

*Denitrification is carbon-limited and proportional to arrival.* There is no readily biodegradable COD state and no separate anoxic biomass, so (10.6) gets the direction and the magnitude of the nitrate right and would get the response to a methanol dose wrong.

*Oxygen transfer is one power law with one condition slider.* Alpha, beta and fouling are lumped, there is no dependence on mixed-liquor solids and no flux distribution across the grid, and (10.9) uses the mid-depth approximation, which overstates transfer near the surface and understates it at the floor.

*One blower standing in for a hall.* There are no inlet guide vanes, no staging of multiple machines, no discharge temperature limit and no motor thermal model, so $a_0$ is the intercept of a fit rather than a measured head and $q_{surge}$ is the minimum flow of the whole hall rather than of one impeller. The anti-surge controller is (10.13) inside the flow balance rather than a loop a student can detune.

*The analyzer fails only by being slow.* A real wet-chemistry instrument also drifts between calibrations, fails to a last good value that looks entirely plausible, and reports its own health. A supervisory loop that trusts a stuck reading is a more common industrial failure than one that is badly tuned, and this application cannot stage it.

*Sludge age is a parameter rather than a wasting flow.* The application can show what a solids retention time decision does to the population and cannot show the settlement-tank capacity or the sludge-treatment load that paying for it costs.

## 10.4 Controller Design

### 10.4.1 The Shared Algorithm

All six controllers, three dissolved oxygen loops and three single loops above them, run the velocity-form algorithm of Section 1.6.2, written here in the symbols the later equations use, as shown in Equations (10.16) and (10.17).

$$e_k = a\,(\mathrm{SP}_k - \mathrm{PV}_k) \tag{10.16}$$

$$P_k = K_c e_k, \qquad D_k = -K_c \tau_D\, a \left.\frac{\mathrm{d}\mathrm{PV}^f}{\mathrm{d}t}\right|_k$$

$$\Sigma_k = \Sigma_{k-1} + (P_k - P_{k-1}) + (D_k - D_{k-1}) \tag{10.17}$$

with $a = +1$ on the two reverse acting loops and $a = -1$ on the two direct acting ones. The running sum $\Sigma_k$ collapses to $P_k + D_k$ under whatever gains are in force, and the filtered measurement $\mathrm{PV}^f$ advances by an exact zero-order-hold recursion. Every loop in this application uses $\tau_D = 0$, so the derivative term contributes nothing and the filter is inert. The output is formed and the integral is

**Table 10.3** The Implemented Loops. All Four Are Proportional Integral With $\tau_D = 0$ and a 0.1 min Scan.

| Loop | CV | MV | $K_c$ | $\tau_I$ (min) | Action |
|---|---|---|---|---|---|
| PIC | $P_h$ | $N$ | 1.20 | 0.5 | reverse |
| DIC ×3 | $DO_i$ | $v_i$ | 8.0 | 6 | reverse |
| VPC | $v_{\max}$ | $P_{h,\mathrm{SP}}$ | 0.15 | 45 | direct |
| AIC | $S_3$ | $DO_{\mathrm{SP}}$ | 1.20 | 480 | direct |

back-calculated against whatever was applied, as shown in Equation (10.18).

$$\begin{aligned} u_k^{raw} &= u_{bias} + \Sigma_k + I_{k-1} \\ u_k &= \min(u_{\max}, \max(u_{\min},\ u_k^{raw})) \\ I_k &= I_{k-1} + \left( \frac{K_c}{\tau_I} e_k + \frac{u_k - u_k^{raw}}{T_t} \right) \Delta t_c \end{aligned} \tag{10.18}$$

where the integral update is disabled when its gain is zero. $T_t$ is the tracking time constant and defaults to $\tau_I$, and $\Delta t_c = 0.1$ min is the scan period of every loop [1]. The running sum $\Sigma_k$ is the $S_k$ of Equation (1.15) and $u_k^{raw}$ is the $v_k$ of Equation (1.16). Anti-windup is back-calculation on the integral alone and $\Sigma_k$ is never clamped, for the reason Section 1.6.2 gives. No loop carries a rate limit. In manual the integral of (10.18) is replaced by $I_k = u^{man} - u_{bias} - \Sigma_k$, which lands the output on the operator's number, so a transfer either way is bumpless.

### 10.4.2 Four Loops, Three Orders of Magnitude Apart

Table 10.3 lists what each loop measures, what it moves, and the tuning the application ships. Ordering the four by speed is the design decision, and the four signs are where an error is easiest to make.

More blower speed raises the header and more valve opening raises the tank oxygen, so both of those process gains are positive and both controllers are reverse acting with $a = +1$ in (10.16). More dissolved oxygen means less effluent ammonia, and raising the header pressure puts more differential across every valve so that each has to close, so both supervisory gains are negative and both controllers are direct acting with $a = -1$. Getting the ammonia loop's sign backward produces a controller that answers rising ammonia by turning the air down, which is stable for about an hour and then is not, and the application's test suite pins all four directions by nudging each loop away from the point it is already tracking.

The header pressure controller writes blower speed over 48 to 100 %. Its lower bound is not a motor limit, because below about 48 % the machine cannot raise the header above $P_{back}$ at all, so the turndown is set by the depth of the tank. The three dissolved oxygen controllers write valve position over 0 to 100 % against setpoints in the ratio 1.00, 1.00 and 0.90, which is tapered aeration carried into the setpoints as well as into the diffuser density. A gain of 8 %/mg/L against a measured process gain of 0.180 mg/L per percent of valve and a 6.20 min lag is about half of what an internal model control rule would return, and the detuning is deliberate because the same three controllers see three different oxygen demands that all move together through the day.

### 10.4.3 Most-Open-Valve Control

The valve position controller holds the busiest air valve near a target opening by moving the header pressure setpoint, over 55 to 95 kPa with a default target of 90 %. Two details decide whether it works. It has to be slow, because it sits above three dissolved oxygen loops and becomes a disturbance to them if it moves faster than they settle, hence 45 min of integral time against their six. And it has to watch the demand rather than the shaft, as shown in Equation (10.19).

$$v_{\max} = \max_i\ u_i^{raw}, \qquad P_{h,\mathrm{SP}} = \mathrm{PI}\big(v_{ref},\ v_{\max}\big) \tag{10.19}$$

where $u_i^{raw}$ is the unsaturated output of dissolved oxygen loop $i$ from (10.18) rather than its clamped value. A valve pinned at 100 % cannot report that it wanted 130 %, so a controller reading valve position stops raising the header exactly when the plant most needs it raised [6]. The same structure returns load to an absorption chiller in Volume 4, Chapter 9.

**Pause.** Try Exercise 10.2 before reading on.

### 10.4.4 The Ammonia Supervisor and Its Instrument

The ammonium analyzer is modeled as a transport lag of 2 min from the final channel to the kiosk, followed by a sample and hold on a 5 min instrument cycle. Additive colored noise with a standard deviation of 0.05 g N/m$^3$ is applied before the delay and held along with the sample. What the supervisor reads is therefore a staircase that is already several minutes old. The supervisor writes the dissolved oxygen setpoint over a hard travel of 0.2 to 4.5 mg/L and an operator travel of 0.5 to 3.0 mg/L, the floor because a basin below about 0.5 mg/L goes anoxic inside the floc and starts producing nitrous oxide [18], and the ceiling because (10.3) is 86 % saturated at 3 mg/L and 90 % at 4.5 mg/L while the air still costs full price.

An integral time of 480 min on a loop whose measurement updates every five minutes looks absurd and is correct. A change in dissolved oxygen takes about three hours to reach the analyzer through the basin and the kinetics, and the quantity the loop is defending is itself an eight-hour mean by (10.15), so the supervisor is tuned to the plant it supervises rather than to the instrument it reads. Whenever a supervisor is out of service the corresponding controller is run in tracking mode against the operator's own number, so all four loops are seeded from the plant at $t = 0$ and none of them steps its actuator on the first scan.

### 10.4.5 A Day-Ahead Predictive Supervisor

A model predictive controller (MPC) is the fourth value of the aeration control option, after manual, fixed dissolved oxygen and ammonia-based control. Fixed dissolved oxygen remains the default, and only scenario 14 opens on the MPC. It moves the handle the ammonia supervisor moves, the dissolved oxygen setpoint of all three tanks, and everything beneath that setpoint acts as shipped, from the three dissolved oxygen loops and their taper to the header loop, most-open-valve control when selected, the blower's speed limits and its anti-surge vent. Every $T_s = 15$ min one optimization plans the setpoint over $H = 96$ steps, a full day of one load cycle and three permit windows, in 19 move blocks of 1, 1, 1, 1, 2, 2, 2, 2, 4, 4, 4, 4, 8, 8, 8, 8, 12, 12 and 12 steps. Only the first block takes effect, staying in force until the next decision, which is the receding horizon of Section 1.4.3 [19, 20]. A case study of predictive control on an activated sludge process is reported in [21]. On every scan the setpoint is clamped to the supervisor's operator floor and ceiling, 0.5 and 3.0 mg/L by default.

The controller's model is written from Section 10.3 with the declared values of Table 10.2 and the design disturbances (13 °C, $\alpha F = 0.55$, nominal carbon load), never with the run's own settings, so a learner's change to the plant appears to the controller as mismatch. It keeps the four ammonia balances and one lag for the dissolved oxygen loops, as shown in Equation (10.20).

$$
\begin{aligned}
V_{an}\frac{\mathrm{d}\hat{S}_a}{\mathrm{d}t} &= \hat{Q}\hat{S}_{in}(1-f_{ass}) + \hat{Q}_r\hat{S}_3 - \hat{Q}_{an}\hat{S}_a + V_{an}\,d \\
V_t\frac{\mathrm{d}\hat{S}_i}{\mathrm{d}t} &= \hat{Q}_{an}\left(\hat{S}_{i-1}-\hat{S}_i\right) - \frac{\mu_A\left(13,\hat{S}_i,\widehat{DO}_i\right)}{Y_A}X_A^d V_t \\
\tau_c\frac{\mathrm{d}\widehat{DO}_1}{\mathrm{d}t} &= DO_{\mathrm{SP}} - \widehat{DO}_1, \qquad \widehat{DO}_i = \gamma_i\widehat{DO}_1
\end{aligned}
\tag{10.20}
$$

where hats mark the controller's quantities, $i = 1,2,3$, $\hat{Q}$ and $\hat{S}_{in}$ come from the preview, $\mu_A$ is (10.3) at 13 °C, $X_A^d = 101.58\,\mathrm{g\,VSS/m^3}$ is the design-point population of (10.4) held constant, $d$ is an unexplained ammonia load into the anoxic zone in g N/m$^3$ per minute, $\gamma_2 = 1.00$ and $\gamma_3 = 0.90$ are

the operator's taper ratios, and $\tau_c = 4.3\,\text{min}$ is the closed-loop answer of the dissolved oxygen loops, the time constant plus dead time fitted to ±0.3 mg/L setpoint steps at 2 mg/L on a flat, noise-free day (fits at 1 to 3 mg/L leave residuals of at most 2.2 %). The predicted power $\hat{W}_k$ applies (10.8) to (10.14) at the declared values to the air each tank needs for its forecast carbonaceous and nitrification uptake, at the measured header pressure held for the day, with the surge vent and the 15 kW of mixers. A setpoint step from 2.0 to 2.5 mg/L changes the effluent ammonia six hours later by $-0.0646\,\text{g N/m}^3$ on the plant and $-0.0641\,\text{g N/m}^3$ on the model, and a step to 1.5 mg/L by $+0.1337$ and $+0.1319\,\text{g N/m}^3$. Driven open loop through scenario 5's day by the probe readings, the model follows the effluent ammonia with an RMS error of $0.137\,\text{g N/m}^3$ and the power with 1.6 kW, 0.9 % of the mean.

The oxygen factor of (10.3) makes (10.20) strongly nonlinear in the variable the plan moves. At steady state the gain from dissolved oxygen to effluent ammonia is $-0.001\,\text{g N/m}^3$ per mg/L at 04:00 and 3 mg/L but $-18\,\text{g N/m}^3$ per mg/L at 20:00 and 0.5 mg/L, and along the plans actually chosen the sensitivity over one decision period still runs from $-0.0069$ to $-0.465\,\text{g N/m}^3$ per mg/L, a factor of 67. The framework's shared linear MPC class carries one fixed model and cannot follow that, so this application forms its own linear time-varying prediction and hands the condensed program to the framework's shared quadratic program solver, the linearization along the plan that Section 1.4.2 lists among the prediction models of this volume. At each decision the model is advanced by fourth-order Runge-Kutta in 2.5 min substeps along the current plan, linearized by finite differences at all 96 steps, and the sensitivity of every step's state to every block's setpoint $u_j$ is accumulated. The plan minimizes Equation (10.21).

$$\begin{aligned} \mathscr{J} = &\sum_{k=1}^{H} \hat{W}_k T_s + \sum_{h=1}^{24} \left[ \sigma_1 \left( \varepsilon_h + \eta_h \right) + \sigma_2 \left( \varepsilon_h^2 + \eta_h^2 \right) \right] \\ &+ \lambda \sum_{j=1}^{19} \left( u_j - u_{j-1} \right)^2 \\ \text{s.t. } & \hat{\bar{S}}_{e,k} \le \bar{S}_c + \varepsilon_h, \quad \hat{S}_{3,k} \le S_{max} - 0.5 + \eta_h \\ & \varepsilon_h, \eta_h \ge 0, \quad DO_{lo} \le u_j \le DO_{hi} \end{aligned} \tag{10.21}$$

Here $h$ is the hour holding step $k$, $T_s$ is in hours, $u_0$ is the setpoint applied now, $DO_{lo}$ and $DO_{hi}$ are the operator floor and ceiling, and $\varepsilon_h$ and $\eta_h$ are slacks. The predicted rolling mean $\hat{\bar{S}}_{e,k}$ is the window of (10.15) formed from the predicted $\hat{S}_3$ and, for the part already past, the controller's ring of the last 32 quarter-hour means of the analyzer reading. The plan's ceiling $\bar{S}_c = 2.2\,\text{g N/m}^3$ is the challenge's target, $0.3\,\text{g N/m}^3$ inside $S_{limit}$, a back-off in the sense of Section 1.4.7. Both ceilings are soft with one slack per hour, with the linear and quadratic prices of Equation (1.7), at $\sigma_1 = 2000$ kWh per g N/m$^3$ and $\sigma_2 = 2 \times 10^4$ kWh per (g N/m$^3$)$^2$, and the move price is $\lambda = 100$ kWh per (mg/L)$^2$. Energy is linear and nothing below the ceilings is priced. The program has 67 variables and 278 inequality rows, the smallest eigenvalue of its Hessian is 1.30, so it is strictly convex, and a dual active-set method [22] solves it.

Up to four such programs form one sequential quadratic program, warm started from the previous plan shifted by one decision as in real-time iteration [23]. Each pass relinearizes along the last plan, moves no block by more than 1.0 mg/L, and keeps the longest of the steps 1, 1/2, 1/4 and 1/8 that lowers the nonlinear objective, stopping when none does or no block would move 0.01 mg/L. Over 6530 plans, every scenario under the MPC and nine mismatched days, no program ended other than optimal and no plan was rejected, and 60 % used all four passes. A compass search over all 19 blocks of the same objective, started from the returned plan, lowers it by at most 0.002 % at five of six decisions examined and by 0.12 % at a run's first plan, whose first two passes hit the step limit.

A steady-state Kalman filter carries the four ammonia states and $d$. Every 0.1 min scan, under every aeration structure, it integrates (10.20) with the probe readings in place of $\widehat{DO}_i$, and at each decision it corrects all five states from the analyzer's held reading against the model's $\hat{S}_3$ of 4.5 min

earlier, the declared transport lag plus half the cycle. The gain belongs to the 15 min model at the design point, with process variances per decision of $10^{-4}$ (g N/m$^3$)$^2$ on each concentration and $10^{-4}\varphi$ on $d$, $\varphi = 1$, and a measurement variance of 0.0225 (g N/m$^3$)$^2$, the analyzer noise plus 0.02 for its staircase. The random walk $d$ makes the plan offset-free, as an input-disturbance estimate does in Section 1.4.5 [24].

The preview is a switch. Typical day feeds the declared average flow and strength through the application's diurnal curve at the clock hour, a calendar rather than the day's measured load, which on the application's ordinary dry day is exactly the load that arrives, while any other day reaches the plan only through the analyzer. Off feeds the declared average all day, which is all the supervisor knows of the hours ahead, and is the like-for-like comparison. No perfect preview is offered. The prices $\lambda$ and $\varphi$ were set once, before any comparison, on scenario 7's day. Of 35 pairs, 23 kept the setpoint travel per day within the supervisor's 12.58 mg/L at 1.30 g N/m$^3$, the worst rolling mean at or below 2.25 g N/m$^3$ and no minute over the permit, and the largest $\lambda$ within 1 % of the least energy was taken, $\lambda = 100$ with $\varphi = 1$ at 8023 kWh over days 2 and 3 against 8016 kWh. It lay on the grid's edge, and $\lambda = 300$ cost 1.3 % more.

The plan is rejected whenever the solver reports anything but optimal, a move is not finite or it leaves the floor and ceiling, and the ammonia supervisor, which tracks the applied setpoint on every scan, takes over until the next accepted plan. Under the MPC the dashed setpoint line on the ammonia trend still shows that fallback supervisor's 1.90 g N/m$^3$ aim, which is nothing the plan aims at. Three rejections forced at 600 min in scenario 14 with the noise off move the setpoint by 0 mg/L, and the plan is back at 645 min. The controller reads the probe readings, the analyzer's held reading, the measured header pressure, the clock hour, the applied setpoint, the operator's floor, ceiling and taper and its three panel settings, and never the plant's states, parameters or disturbances. A test feeds a second plant, different in 14 parameters, 7 disturbances and every state, the first plant's instrument readings and obtains the same 39 plans, so the plant, its instruments and (10.20) stay three separate models. In two runs of scenario 14 in Chrome at a speed of 200 simulated minutes per second a plan takes a mean of 13.1 and 13.2 ms, 17.4 and 17.6 % of one core, at 59.6 to 59.8 frames per second, and at a speed of 20 at most two frames in 15 s take 33.4 ms.

## 10.5 Results

Fifteen scenarios build up in a laboratory order, and Table 10.4 lists them. Each pins the opening mode, the structure in service, the four tunings, the disturbance settings and whatever events are timed into it, and declares its objective ahead of the run. All run results come from the app's engine at the default seed. Unless stated otherwise, settings are applied before Run, and Reset retains those settings and recreates the same initial plant and controller state (Section 1.6.1). Counterfactual model benchmarks are identified explicitly.

### 10.5.1 The Design Condition and the Two Fast Loops

Scenario 2 flattens the diurnal swing without changing its average and closes the two regulatory loops. Ammonia falls 5.86, 3.39, 1.40 and 0.378 g N/m$^3$ through the four zones on airflows of 67.7, 36.6 and 17.8 Nm$^3$/min through valves at 68.9, 50.0 and 31.7 %. The tanks are not three identical stages. At 3.39 g N/m$^3$ the ammonia factor in (10.3) is already 83 % of its maximum, so tank 1 nitrifies close to flat out and removes a fixed mass per minute, while tank 3 at 0.378 g N/m$^3$ sits at 35 % of the factor and works at a third of the rate on a tenth of the concentration.

Scenario 1 makes the header coupling visible by hand. After one minute of Run, opening valve 1 from 68.9 to 78.9 % and touching nothing else pulls the header from 62.0 to 59.7 kPa, and because (10.10) is a square root in the differential, tank 2 loses 4.93 Nm$^3$/min and tank 3 loses 2.4 Nm$^3$/min without either valve moving. Half an hour later tank 1 is at 2.73 mg/L, tank 2 at 1.63 mg/L and tank 3 at 1.17 mg/L on its way to 1.10 mg/L. Scenario 4 shows what that coupling costs when the inner loop is thrown away. With the header controller detuned to $K_c = 0.12$ % per kPa

**Table 10.4** The Fifteen Scenarios and What Each One Isolates

| # | Scenario | What it exercises |
|---|---|---|
| 1 | Three valves by hand | parallel users on one header |
| 2 | Flat load, loops closed | tapered aeration, Monod in ammonia |
| 3 | Dissolved oxygen steps | process gain that drifts under the loop |
| 4 | Header loop detuned then repaired | why the inner loop must be fast |
| 5 | A whole day at 2.0 mg/L | load, lag and the evidence gap |
| 6 | The same day at 1.2 mg/L | the fixed-setpoint frontier |
| 7 | Ammonia-based control | equal-compliance accounting |
| 8 | The supervisor at $\tau_I = 10$ min | tuning to the instrument |
| 9 | Most-open-valve control | removing an unused pressure drop |
| 10 | A dry Sunday night | blower surge line, reduced marginal saving |
| 11 | A three-day cold snap | the state with no fast handle |
| 12 | Liquors at 09:00 and at 01:00 | timing, returned mass and preceding state |
| 13 | Diffusers foul over a year | a failure that holds every setpoint |
| 14 | The eight-hour mean, planned a day ahead (MPC) | planning the permit's own mean against the supervisor |
| 15 | Graded February day | the whole structure under a budget |

and $\tau_I = 8$ min, a 1.2 mg/L setpoint step sags the header to 55.4 kPa and leaves all three tanks 0.68, 0.61 and 0.46 mg/L from setpoint fifteen minutes later, on errors none of them caused. Repairing the tuning mid-run and taking the same step back down keeps the header inside 61.9 to 64.9 kPa and the fifteen-minute errors at 0.03, 0.11 and 0.26 mg/L.

Scenario 3 measures the gain the dissolved oxygen loop is tuned against and shows it moving in two directions at once. The equal-percentage trim in (10.10) makes the valve gain rise as it opens while the 0.75 exponent in (10.8) makes the transfer gain fall as the air rises, and the two nearly cancel. Measured at fixed header pressure the gain is 0.198, 0.180 and 0.163 mg/L per percent of valve at setpoints of 1, 2 and 3 mg/L, a drift of 18 % across the whole range. Setting the rangeability exponent to zero for a linear trim gives 0.191, 0.144 and 0.110 mg/L per percent, a drift of 42 %, with the valve sitting at 41.2 % open instead of 68.9 %. That cancellation is why equal-percentage trim is specified on air valves, and it is one of the few places where the right answer is to change the plant rather than to schedule the controller.

### 10.5.2 A Whole Day, and What the Setpoint Range Is Worth

Scenario 5 runs midnight to midnight at a fixed 2.0 mg/L. The 24-hour mean effluent ammonia is 1.100 g N/m$^3$ and the peak is 3.27 g N/m$^3$ at 11:45, an hour and three-quarters after the load peaked at 09:54, because the sewage takes 7.92 h to cross the basin. The worst eight-hour rolling mean is 2.277 g N/m$^3$ against a 2.5 g N/m$^3$ limit, and the day costs 4335 kWh, which is 181 kW around the clock and 0.394 kWh per cubic meter treated.

Two facts in that run set up everything after it. The morning peak was entirely predictable, because the load starts climbing at 06:00 and peaks at 09:54 while the analyzer does not see the consequence until 11:45, and the same curve repeats every weekday. And at 04:00 the effluent ammonia is 0.011 g N/m$^3$, two orders of magnitude inside the limit, while the plant spends 47.4 Nm$^3$/min and 76.7 kW holding 2.0 mg/L of oxygen that nothing is using. Dropping the setpoint to 1.0 mg/L at that hour takes the ammonia to 0.023 g N/m$^3$ on 40.9 Nm$^3$/min and 68.3 kW.

Scenario 6 starts at 1.2 mg/L and costs 3813 kWh, with a worst rolling mean of 3.786 g N/m$^3$ and instantaneous peak 4.79 g N/m$^3$. Its initialization uses that lower oxygen profile. For a comparison that holds the opening plant state fixed, scenario 5 is loaded and run for one minute at its supplied setpoint before the target is changed. Each trial retains the same opening plant state and runs to 1440 minutes. Sweeping that common state gives the frontier this plant lives on, plotted in Figure 10.2 and tabulated in Table 10.5. The 1.2 mg/L benchmark sits at its far end, saving 516 kWh against

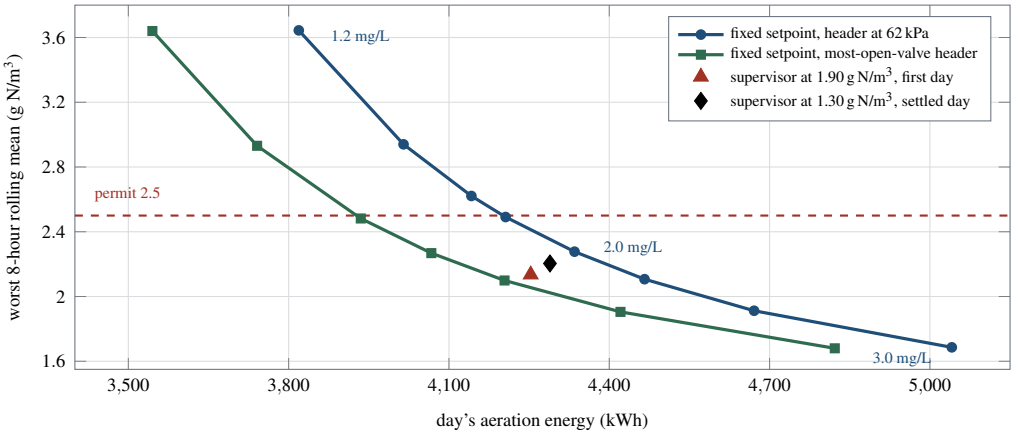



**Figure 10.2** The fixed-setpoint frontier over one diurnal day, with and without most-open-valve control on the air header. Each marker on a curve is one fixed dissolved oxygen setpoint between 1.2 and 3.0 mg/L, and every run starts from the same equilibrium, with the DO target changed after one minute. The header loop moves the whole curve left by about 270 kWh at unchanged compliance, while the ammonia supervisor moves one point along the curve by 194 kWh on its first day and 66 kWh once settled. Circles mark the header at 62 kPa, squares the most-open-valve header, the triangle and the diamond the supervisor's first and settled days, and the dashed horizontal line the permit.

**Table 10.5** Fixed-Setpoint Frontier Over One Diurnal Day, Read After the First Eight Hours So That the Rolling-Mean Ring Is Full. Every Run Starts From the Same 2.0 mg/L Equilibrium, With Targets Changed After One Minute.

| $DO_{SP}$ | header at 62 kPa | | most-open valve | |
|---|---|---|---|---|
| (mg/L) | energy (kWh) | $\bar{S}_e$ (g N/m$^3$) | energy (kWh) | $\bar{S}_e$ (g N/m$^3$) |
| 1.2 | 3819 | 3.644 | 3545 | 3.640 |
| 1.5 | 4015 | 2.940 | 3741 | 2.931 |
| 1.8 | 4206 | 2.491 | 3935 | 2.482 |
| 2.0 | 4335 | 2.277 | 4067 | 2.268 |
| 2.2 | 4466 | 2.107 | 4204 | 2.099 |
| 2.5 | 4671 | 1.912 | 4421 | 1.905 |
| 3.0 | 5041 | 1.686 | 4822 | 1.680 |

the 2.0 mg/L day while the worst rolling mean reaches 3.644 g N/m$^3$ and the instantaneous peak 4.66 g N/m$^3$. The curve is steep at the bottom and flat at the top, because the half milligram from 1.2 to 1.7 mg/L buys 1.02 g N/m$^3$ of rolling mean for 323 kWh while the half milligram from 2.5 to 3.0 mg/L buys 0.22 g N/m$^3$ for 370 kWh. End to end the whole practical range is 1222 kWh, 32 % of the cheapest day on it, which is the arithmetic that bounds every claim made for a setpoint policy on this plant.

### 10.5.3 What Load-Following Is Worth, Read Honestly

The fine matching values 2.02 mg/L and 2.17 mg/L below are counterfactual model benchmarks, between the interface's 0.05 mg/L setpoint increments. The other fixed-setpoint runs use the exposed grid.

Scenario 7 puts the analyzer in charge of the dissolved oxygen setpoint, aiming at 1.90 g N/m$^3$ with $K_c = 1.20\,\mathrm{mg/L}$ per g N/m$^3$, $\tau_I = 480\,\mathrm{min}$ and a travel limit of 0.5 to 3.0 mg/L. The setpoint starts where the operator left it and walks down all night, reaching its 0.5 mg/L floor at 05:07, its

**Table 10.6** The First Day Is Not the Settled Day. Identical Configurations Run for Three Consecutive Diurnal Days, Energy and Worst Rolling Mean per Day. Each row is a fresh run. Fixed targets other than 2.00 mg/L change after one minute. Supervisor aims are applied before Run. The last four rows run scenario 14's settings, the MPC of Section 10.4.5 three ways and the ammonia supervisor tuned with the day known (aim 1.40 g N/m$^3$, ceiling 2.5 mg/L, $K_c = 4.8$, $\tau_I = 900$ min).

| | day 1 | | day 2 | | day 3 | |
|---|---|---|---|---|---|---|
| Configuration | kWh | $\bar{S}_e$ | kWh | $\bar{S}_e$ | kWh | $\bar{S}_e$ |
| Fixed 2.00 mg/L | 4335 | 2.277 | 4341 | 2.245 | 4342 | 2.221 |
| Fixed 2.02 mg/L | 4348 | 2.258 | 4354 | 2.226 | 4355 | 2.201 |
| Fixed 2.17 mg/L | 4446 | 2.130 | 4452 | 2.094 | 4453 | 2.066 |
| Supervisor at 1.90 g N/m$^3$ | 4253 | 2.135 | 4109 | 2.563 | 4105 | 2.598 |
| Supervisor at 1.30 g N/m$^3$ | 4471 | 1.919 | 4307 | 2.183 | 4289 | 2.204 |
| MPC, typical day | 4018 | 2.211 | 4010 | 2.219 | 4011 | 2.207 |
| MPC, preview off | 4306 | 2.597 | 4309 | 2.630 | 4306 | 2.604 |
| MPC, ceiling 2.15 g N/m$^3$ | 4044 | 2.162 | 4032 | 2.169 | 4032 | 2.157 |
| Supervisor in hindsight | 4352 | 1.995 | 4230 | 2.175 | 4212 | 2.213 |

3.0 mg/L ceiling at 10:05 and 1.3 mg/L again by late afternoon. Between midnight and 07:00 the plant spends 564 kWh against the fixed setpoint's 646 kWh, and from 07:00 to midnight the two spend 3689 kWh each to the kilowatt-hour. It buys the night outright and a cleaner morning for nothing.

Over that first day the supervisor finishes at a worst rolling mean of 2.135 g N/m$^3$ on 4253 kWh against the fixed setpoint's 2.277 g N/m$^3$ on 4335 kWh, which is better on both counts and is something no point on the frontier of Table 10.5 manages against any other. Comparing at nearby compliance rather than at equal setpoint, the fixed setpoint that reaches 2.135 g N/m$^3$ is 2.17 mg/L and costs 4446 kWh, so the saving is 194 kWh, or 4.4 %.

Running the identical configuration for a second and a third day is what changes the answer, and Table 10.6 reports it. The supervisor settles at 4109 and 4105 kWh on rolling means of 2.563 and 2.598 g N/m$^3$, cheaper than the fixed setpoint and outside the 2.5 g N/m$^3$ limit, while the fixed setpoint barely moves at 4341 and 4342 kWh. Nothing broke. An eight-hour integral takes about a day to forget the 2.0 mg/L setpoint it was seeded from, and while it is forgetting it runs richer than its own aim asks for. Aiming at 1.30 g N/m$^3$ instead gives a settled day of 4289 kWh at 2.204 g N/m$^3$ against the fixed setpoint's 4342 kWh at 2.221 g N/m$^3$, and at nearby compliance a fixed 2.02 mg/L reaches 2.201 g N/m$^3$ on 4355 kWh, so the settled saving is 66 kWh, or 1.5 %. Four percent on the day of commissioning and one and a half percent for the rest of the loop's life is the honest pair, and the reason the first number is the one usually quoted is that most trials run for a week.

🕮 **Pause.** Try Exercise 10.3 before reading on.

Scenario 8 is scenario 7 with the supervisor's integral time at 10 min, which is a sensible number for a flow loop. Over the day the fast supervisor finishes at 2.440 g N/m$^3$ on 4206 kWh against the slow one's 2.135 g N/m$^3$ on 4253 kWh and the fixed setpoint's 2.277 g N/m$^3$ on 4335 kWh, so it has made the effluent worse than having no supervisor at all while moving the setpoint 16.8 mg/L against 15.8 mg/L to do it. The fast loop saves 47 kWh against the slow one while using more of the compliance margin. A flat-load comparison with the analyzer noise enabled makes the activity visible. Starting from scenario 2, enabling the ammonia supervisor and 0.05 g N/m$^3$ analyzer noise, and measuring the final ten hours of a three-day run gives 10.91 mg/L of setpoint travel over a 0.70 mg/L band for the fast loop, against 6.72 mg/L over 0.27 mg/L for the slow one. The input load is constant throughout, and the staircase, transport delay and sampled noise still drive controller activity.

### 10.5.4 The Loop Nobody Sells

Scenario 9 switches most-open-valve control in at a flat load. Comparing means over the opening hour and minutes 300–360, the header comes down from 62.00 to 56.03 kPa, the busiest valve opens

from 68.94 to 87.74 %, the blower slows from 59.54 to 57.65 %, and total aeration power falls from 173.93 to 160.96 kW, a reduction of 7.5 %. The largest dissolved-oxygen excursion is 0.113 mg/L in tank 1. Oxygen transferred per kilowatt-hour rises from 1.64 to 1.78 kg. Nothing was given up anywhere, because the 6.0 kPa that disappeared was being manufactured by the blower and destroyed across three butterfly valves. Over a full diurnal day the same loop takes the fixed-setpoint plant from 4335 to 4067 kWh, which is 267 kWh or 6.2 %, with the rolling mean unmoved at 2.268 g N/m$^3$ against 2.277 g N/m$^3$. The two loops add, since the supervisor with the header loop underneath it finishes its first day at 4025 kWh.

> **TRY IT IN THE APP 9 · Most-Open-Valve Control**
>
> Load the scenario, press Run and let it pass the switch at minute 60. Watch the header setpoint walk down while the three dissolved oxygen traces hardly move. Between the opening hour and minutes 300 to 360 the header falls from 62.00 to 56.03 kPa and the aeration power from 173.93 to 160.96 kW.

### 10.5.5 Two Losses Hidden by Successful Regulation

Scenario 13 takes the diffuser condition from 0.55 to 0.48 to 0.38, which is about a year of service on membrane diffusers, with the dissolved oxygen loops in service throughout. The dissolved oxygen returns near its target after each step, although tank 1 falls to 1.37 mg/L during the last transition. Effluent ammonia briefly reaches 0.408 g N/m$^3$ and remains below the permit. No consent alarm fires. What moves is everything nobody watches. Airflow goes from 122 to 146 to 200 Nm$^3$/min, the valves from 69/50/32 to 86/65/43 %, the blower from 59.5 to 75.7 %, the power from 174 to 205 to 276 kW, and oxygen per kilowatt-hour from 1.64 to 1.37 to 1.00 kg. That last hundred kilowatts is $106,000 a year at twelve cents. There is no dissolved oxygen alarm that will ever fire, because holding the dissolved oxygen is exactly what the loops are doing, and the only way to see the loss is to monitor the air it takes to hold it [2, 13].

Scenario 10 is a dry Sunday night at three quarters of the usual flow and nine tenths of the usual strength. By 04:10 the tanks want 31.3 Nm$^3$/min while the machine is passing 37.1 Nm$^3$/min against a surge line at 40.0 Nm$^3$/min, so the blow-off valve of (10.13) vents 5.8 Nm$^3$/min and 1002 standard cubic meters over the seven hours of the run. The dissolved oxygen stays within 0.13 mg/L of setpoint the whole time. What the surge line takes is not the power, it is the saving. Taking the setpoint from 2.0 to 0.5 mg/L at that load cuts the air the tanks receive from 33.7 to 27.2 Nm$^3$/min, which is 19 %, while the machine turns down only from 37.9 to 35.6 Nm$^3$/min and the plant draws 61.4 kW against 64.3 kW, which is 4.5 %. The same setpoint change at the daily average load moves the plant from 173.9 to 133.2 kW, or 23 %. Every cubic meter the tanks stop taking is made up by the blow-off valve instead, so the air discarded over the comparison rises from 584 to 2134 standard cubic meters while the meter falls only from 545 to 477 kWh. Both comparison runs start at the supplied 1.5 mg/L, change target after one minute and continue for seven more hours. Selecting 0.5 mg/L before Run instead initializes a long-term washed-out equilibrium, rather than the operating plant. Neither failure has a tuning answer, and both have capital or scheduling answers that need exactly the number the scorecard is showing.

### 10.5.6 Cold, and a Disturbance That Should Not Have Arrived

Scenario 11 drops the sewage from 13 to 9.5 °C on the first morning of three days and recovers it on the third. The nitrifier population does not move, wandering between 97.6 and 101.9 g VSS/m$^3$ with the diurnal load and ending where it began, exactly as (10.4) requires. The effluent ammonia carries the whole loss, tripling from 0.378 to 1.19 g N/m$^3$ at 9.5 °C, with the rolling mean reaching 4.47 g N/m$^3$, the supervisor pinned on its 3.0 mg/L ceiling for twenty-three hours of the three days and the limit exceeded for eighteen. Air buys back about half, because at 9 °C raising the setpoint from 2.0 to 3.5 mg/L takes the ammonia from 1.378 to 0.901 g N/m$^3$ against 0.378 g N/m$^3$ in the warm. The other half is not available at any price, and the handle that would deliver it is the sludge

age, which moves the population by 5.1 g VSS/m$^3$ on the first day after the decision and is still moving on the fifth.

**Pause.** Try Exercise 10.1 before reading on.

Scenario 12 represents dewatering liquors by multiplying influent nitrogen strength by 1.45 for four hours, without adding a separate hydraulic stream. The first pulse starts at 09:00 on day 1. Effluent reaches 7.95 g N/m$^3$, the rolling mean reaches 5.26 g N/m$^3$, and the day costs 4890 kWh. The second pulse starts at 01:00 on day 2, whose peak effluent is 3.22 g N/m$^3$, worst rolling mean 2.43 g N/m$^3$ and energy 4153 kWh. Those consecutive days differ in their preceding state and in added mass: the pulses add 71.21 kg and 18.56 kg of nitrogen respectively above the same diurnal baseline. They are not an equal-mass timing comparison.

A counterfactual run applying only the 01:00 pulse on the first day, from the same opening state as the morning run, reaches 2.92 g N/m$^3$ peak effluent and 2.086 g N/m$^3$ worst rolling mean on 4336 kWh. The 554 kWh difference from the morning pulse therefore combines timing and the smaller injected mass. The native lesson shows the benefit of the nighttime operating pattern, but does not isolate its value per kilogram returned.

### 10.5.7 Planning the Eight-Hour Mean a Day Ahead

Scenario 14 runs scenario 7's ordinary day three times under the MPC of Section 10.4.5, then asks for the supervisor at 1.30 g N/m$^3$ on the same day. Each comparison changes only the named setting over three days on the default seed unless ten seeds are stated, and reads worst means from the logged half-minute record, 0.001 to 0.002 g N/m$^3$ below the every-step reading of Table 10.6.

> **TRY IT IN THE APP** **14 · The Eight-Hour Mean, Planned a Day Ahead (MPC)**
>
> Load the scenario and run three days on MPC with the typical-day preview, at 200 simulated minutes per second. On the third day watch the dissolved oxygen setpoint: the plan never asks for more than 2.49 mg/L, and by 20:00 it has let it fall to 0.86 mg/L while the effluent climbs to 3.22 g N/m$^3$. The three days cost 4018, 4010 and 4011 kWh (Table 10.6).

The MPC spends 4018, 4010 and 4011 kWh on its three days at worst means of 2.210, 2.218 and 2.206 g N/m$^3$ with no minute over the permit. The supervisor at 1.30 g N/m$^3$, the aim that settles at nearly the same compliance, spends 4471, 4307 and 4289 kWh at 1.918, 2.182 and 2.203 g N/m$^3$, and interpolated across its aims to the MPC's third-day worst mean it needs 4288 kWh, so the plan uses 277 kWh, or 6.5 %, less. That saving needs the calendar. With the preview off the MPC knows what the supervisor knows and loses on both counts, at 4306 kWh and 2.602 g N/m$^3$ on the third day with 362 min over the permit in three days, and at all ten seeds of Table 10.7. The best supervisor setting found with the day known, from 1638 combinations of aim, floor, ceiling, gain and integral time, spends 8442 kWh over days 2 and 3 at 2.212 g N/m$^3$ with four times the shipped gain and the maximum integral time of 900 min, and 268 more combinations around it reach only 8434 kWh at 2.218 g N/m$^3$, so the plan's 8021 kWh sits 4.9 % below that hindsight bound. Figure 10.3 draws the third day of Table 10.6.

The third day, shown in Figure 10.4, says where the energy goes. The supervisor spends 2.9 hours on its 3.0 mg/L ceiling around the morning peak, while the plan never sets more than 2.49 mg/L. At 20:00 the supervisor still asks for 2.23 mg/L and holds the effluent under 1.61 g N/m$^3$, while the plan has let the setpoint fall to 0.86 mg/L and lets the effluent reach 3.22 g N/m$^3$ at 20:37, since that ammonia will be averaged with the empty hours after midnight, and its rolling mean stays between 2.01 and 2.21 g N/m$^3$ from 15:00 to midnight. It transfers 6272 against 6333 kg of oxygen a day at lower dissolved oxygen, so on less air. Its night costs 469 against 443 kWh.

Table 10.7 takes the saving apart one switch at a time at the MPC's third-day worst mean of 2.206 g N/m$^3$. The day's shape is worth about 247 kWh, from 4429 kWh to 4182 kWh for a plan whose ceiling sits on the instantaneous ammonia, and planning the permit's own mean a further

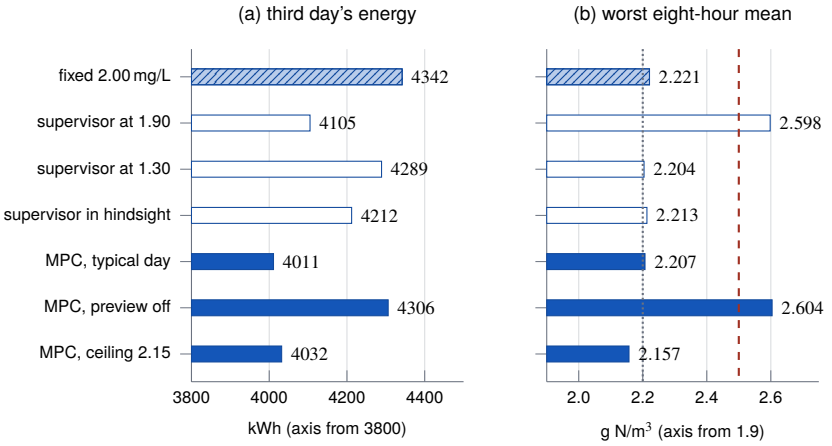



**Figure 10.3** The third day of Table 10.6: (a) the energy and (b) the worst eight-hour rolling mean, for the fixed setpoint (hatched bar), the ammonia supervisor (open bars) and the MPC (dark bars). The red dashed line is the 2.5 g N/m$^3$ permit and the gray dotted line the plan's 2.2 g N/m$^3$ ceiling. Notice that the two bars past the permit are the supervisor aimed at 1.90 g N/m$^3$ and the plan with its preview off, while the plan with the typical day is the shortest bar in (a) near the ceiling in (b).

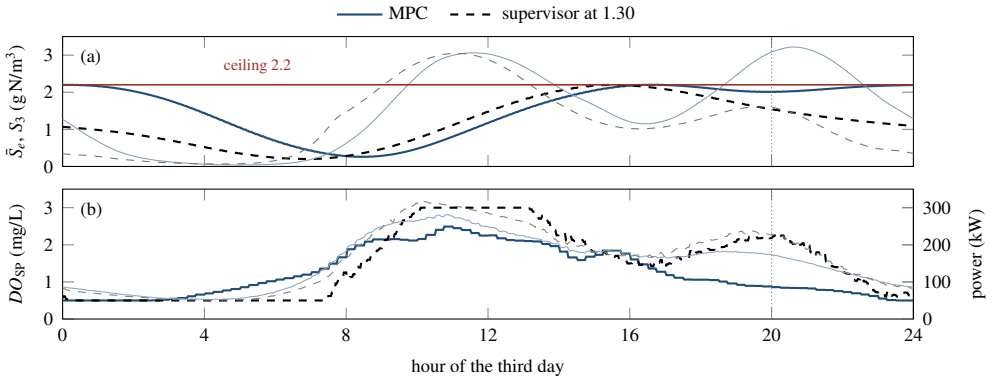



**Figure 10.4** Scenario 14's third day under the MPC (blue, solid) and the ammonia supervisor at 1.30 g N/m$^3$ (black, dashed). (a) Eight-hour rolling mean (thick) against the plan's 2.2 g N/m$^3$ ceiling, and effluent ammonia (thin). (b) Dissolved oxygen setpoint (thick, left axis) and total aeration power (thin, right axis). The dotted line marks 20:00.

171 kWh. The hour's price of air is worth nothing measurable, because a plan pricing every milligram-hour of dissolved oxygen at one flat rate lands on the same 4011 kWh. Feedback cannot reproduce the result either, since a review copy of the supervisor reading the analyzer's eight-hour mean, not shipped, exceeded the permit at all 63 settings tried, with aims of 1.9 to 2.3 g N/m$^3$, gains of 0.15 to 4.8 and integral times of 240 to 900 min.

**Table 10.7** Where the Plan's Saving Comes From, and Ten Noise Seeds. Upper Block: Third-Day Energy Interpolated to a Worst Eight-Hour Mean of 2.206 g N/m$^3$. Lower Block: Days 2 and 3, Mean and Range Over Seeds 1 to 10 Past the Default. Worst Means Read From the Logged Half-Minute Record.

| Configuration (third day) | Energy at a worst mean of 2.206 g N/m$^3$ (kWh) |
|---|---|
| Supervisor at 1.30 g N/m$^3$ (worst 2.203) | 4289, interpolated 4288 |
| MPC, preview off, ceiling on the instantaneous ammonia | about 4429 |
| MPC, preview off, ceiling on the eight-hour mean | not reached (ceiling 1.0 gives 4662 at 2.240) |
| MPC, typical day, ceiling on the instantaneous ammonia | about 4182 |
| MPC, typical day, ceiling on the eight-hour mean (shipped) | 4011 |
| Same, every milligram-hour of oxygen at one flat price | 4011 |
| Supervisor reading the eight-hour mean (review copy) | over the permit at all 63 settings, lowest worst mean 2.716 |

| Configuration (ten seeds) | energy (kWh) | worst mean (g N/m$^3$) | minutes above 2.20 | permit minutes | N, day 3 (kg) |
|---|---|---|---|---|---|
| MPC, typical day | 8028 [8017, 8038] | 2.218 [2.214, 2.222] | 165 [87, 212] | 0 | 35.9 |
| MPC, preview off | 8622 [8613, 8631] | 2.607 [2.570, 2.625] | 759 [740, 780] | 351 | 34.4 |
| Supervisor at 1.30 g N/m$^3$ | 8596 [8590, 8600] | 2.205 [2.200, 2.212] | 11 [0, 31] | 0 | 29.0 |
| Supervisor in hindsight | 8443 [8437, 8448] | 2.218 [2.210, 2.224] | 57 [32, 72] | 0 | 30.8 |
| MPC, ceiling 2.15 g N/m$^3$ | 8073 [8061, 8084] | 2.168 [2.163, 2.172] | 0 | 0 | 35.1 |

**+ DID YOU KNOW? A Plan Without Its Forecast Loses**

From the simulator: on the third day of scenario 14 the plan with the typical-day preview spends 4011 kWh at a worst rolling mean of 2.207 g N/m$^3$. With the preview off it spends 4306 kWh at 2.604 g N/m$^3$, over the permit, while the ammonia supervisor aimed at 1.30 g N/m$^3$ spends 4289 kWh at 2.204 g N/m$^3$ (Table 10.6).

The plan rides its ceiling. Its plans made at 0, 1440, 2040 and 2640 min predict a peak eight-hour mean of 2.192, 2.200, 2.200 and 2.200 g N/m$^3$ over the day ahead, and the following 24 hours realize 2.210, 2.218, 2.218 and 2.206 g N/m$^3$ through its model error. Over ten seeds its worst mean exceeds 2.20 g N/m$^3$ on all 30 seed-days, for 87 to 212 min in three days and by at most 0.022 g N/m$^3$, and ends 0.013 g N/m$^3$ above the supervisor's on average and higher at every seed, so the headline is energy at a compliance 0.013 g N/m$^3$ worse. For information only, a ceiling of 2.15 g N/m$^3$, which any learner can set, stays under 2.20 g N/m$^3$ on every seed-day for 45 kWh more over two days, still 523 kWh below the supervisor with a worst mean no higher than the supervisor's at all ten seeds.

What the plan spends is ammonia. On the third day it discharges 35.9 kg of nitrogen against 29.0 kg, 6.9 kg or 24 % more, because a plan told only a ceiling spends all of it, and it takes every margin the shipped settings leave. In scenarios 2, 3, 4, 9 and 13 the worst mean rises from 0.38 to 0.44 to between 2.09 and 2.13 g N/m$^3$ and the nitrogen discharged from 7.4 to 7.9 to between 28.0 and 28.8 kg a day, for 7.9 to 8.3 % less energy over three days, and in scenario 10 the worst mean rises from at most 0.95 to 2.21 g N/m$^3$ for 11.5 % less. On heavier days than its model it also breaks the permit, 80 to 948 min in three days against the supervisor's 499 to 1207 min. Scenario 11's cold snap costs it 12 007 against 11 738 kWh, 2.3 % more, for a worst mean of 3.948 against 4.468 g N/m$^3$ and no upper-tier minute against 170, and scenario 12's two days of liquors cost 9195 against 9043 kWh, 1.7 % more, with both structures over the permit.

### 10.5.8 Graded February Day

Scenario 15 is one ordinary February day at 12.8 °C with no event of any kind on it, and every gain an order of magnitude out. The target asks for a rolling mean never above 2.20 g N/m$^3$ after the first eight hours, nothing above the 5.0 g N/m$^3$ upper tier, and under 4180 kWh on the meter. As

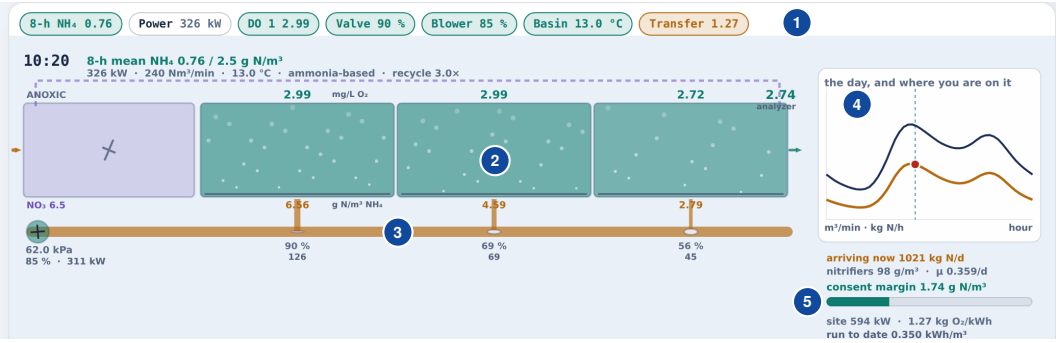



**Figure 10.5** Scenario 7 in progress at 10:20 on the first day of ammonia-based aeration control, cropped to the plant view. (1) The chips: rolling mean, power, tank 1 oxygen, busiest valve, blower speed, basin temperature and oxygen transferred per kilowatt-hour. (2) The three aerated tanks with their oxygen and ammonia readings. (3) The air header and its three valves, with the header pressure at the blower end. (4) The load inset, whose marker sits just past the 09:54 crest of the day, beside the 1021 kg/d arriving now. (5) The consent margin under the permit. By that minute the ammonia supervisor had driven the dissolved oxygen setpoint from 0.5 mg/L overnight to its 3.0 mg/L ceiling fifteen minutes earlier (Section 10.5), the three dissolved oxygen loops were tracking it through their air valves, and the effluent ammonia was still climbing toward a peak the basin would not deliver until 11:15. The rolling mean at that minute was 0.754 g N/m$^3$, which the scene prints rounded, still under the 2.5 g N/m$^3$ permit line.

handed over the day finishes at a rolling mean of 10.5 g N/m$^3$, 10.465 g N/m$^3$ at every step, with 527.56 min above the upper tier on 3832 kWh, so it fails on ammonia while it is inside the energy budget. The inner-to-outer repair sequence is appropriate because the header loop is underneath everything and the dissolved oxygen loops cannot even be judged until it is right. Repairing the two regulatory loops alone reaches 4298 kWh at 2.39 g N/m$^3$, which is close on energy and fails the limit. Adding most-open-valve control at a 92 % target saves 286 kWh, leaving 4011 kWh with a worst rolling mean of 2.384 g N/m$^3$. Energy now meets the budget but ammonia still fails. The suggested supervisor aims at 1.65 g N/m$^3$ with a 3.2 mg/L ceiling. It finishes at 4123 kWh and a worst rolling mean of 2.062 g N/m$^3$, with no upper-tier time, meeting every clause with about 57 kWh to spare. Applying these settings before Run or retaining them through Reset gives the same result. An aim of 1.30 g N/m$^3$, which settles scenario 7, over-aerates this day.

The arithmetic at the end of that day is the point of the challenge. A fixed setpoint hand-picked at 2.25 mg/L, with the header optimized at a 94 % target, also clears the target at 4159 kWh and 2.160 g N/m$^3$. Thirty-seven kilowatt-hours, under one percent, separates the two answers on this particular day. The difference is that 2.25 mg/L is a number somebody chose for this day, and on the same works two degrees warmer it is spending money on oxygen nothing will use, while nobody goes back and lowers it in June. The supervisor spends that four percent only on the days that need it.

The MPC does not meet this target either. On the day as handed over it plans above the detuned loops and ends at 9.764 g N/m$^3$. In place of the suggested supervisor, eleven seeds end at 3772 to 3790 kWh and worst means of 2.206 to 2.219 g N/m$^3$, failing the 2.20 g N/m$^3$ clause at all eleven by its model error, and 2.790 g N/m$^3$ with the preview off. With its ceiling at 2.15 g N/m$^3$ it meets every clause at all eleven seeds on 3800 to 3821 kWh, against the suggested supervisor's 4118 to 4128 kWh.

## 10.6 Technology in Practice

### What Measures It

> **TECHNOLOGY Sensor: The Dissolved Oxygen Probe**
>
> Aeration basins use three kinds of oxygen probe. A membrane electrode, long the most common, lets oxygen pass a membrane to a cathode that reduces it, and the resulting signal is correlated with the concentration. It needs calibration weekly or monthly and a new membrane about every quarter. A galvanic electrode holds its accuracy longer. An optical probe reads how dissolved oxygen quenches the glow of a dye, needs less calibration, and has only a sensor cap to replace, about once a year. In mixed liquor every probe gathers biological growth or grease that makes its reading inaccurate, so probes are wiped weekly to monthly or cleaned by air blasts, water sprays or wipers [4]. No open, vendor-neutral source gave a typical accuracy. The chapter's probe is a 0.75 min lag (Table 10.2).

Automated aeration control takes its input from oxygen probes in the basins, read by a process controller [4]. The quantity the permit names is measured another way. The federal reference method for ammonia nitrogen in wastes is wet chemistry: alkaline phenol and hypochlorite react with the ammonia to form indophenol blue, whose color is read by a colorimetric detector on automated continuous-flow equipment. Its applicable range is 0.01 to 2.0 mg/L as nitrogen, higher concentrations are diluted first, and about 60 samples an hour can be analyzed [25]. In this chapter the oxygen probes are 0.75 min lags, the header transmitter a 0.05 min lag, and the ammonium analyzer a 2 min transport lag followed by a 5 min sample and hold (Table 10.2).

### What Moves It

The total air is set at the blowers: positive displacement machines through variable-frequency drives, multistage centrifugal machines often by inlet throttling, and newer single-stage machines by variable speed, inlet guide vanes and variable discharge diffusers. The air is then divided among the basins by valves, by hand in small plants and by automatic valves in larger ones. Most-open-valve control adjusts the header pressure so that the most open valve stays essentially full open, and it is now commonly integrated into new aeration control systems [4]. The chapter's three butterfly valves are equal-percentage orifices with a 0.25 min travel time, and its blower follows a 0.20 min speed ramp (Table 10.2).

> **TECHNOLOGY Final Element: The Variable-Speed Centrifugal Blower**
>
> Aeration blowers are low-pressure, high-volume compressors that push air to the basins at typically 5 to 14 psi. A positive displacement machine gives a constant volume over a wide range of pressure, while a centrifugal machine gives a wide range of flow over a narrow range of pressure. A federal survey puts a variable-speed multistage centrifugal blower at 60 to 70 percent nominal efficiency with a turndown to 50 percent of rated flow, and reports variable-speed operation 15 to 20 percent more efficient than throttling. Surge does more damage to a single-stage machine, which typically comes with its own surge protection [4]. The chapter's blower is one variable-speed machine with a surge line and a blow-off vent (Equations (10.12) and (10.13)).

### Where the Controller Runs

Most aeration control systems are programmable logic controllers networked under a supervisory control and data acquisition system, and larger facilities sometimes use a distributed control system [4]. Control loops may be nested, with one loop's setpoint taken from another, and supervisory and lower-level loops run continuously with cycle times from milliseconds to minutes [26]. The chapter's six proportional integral controllers scan every 0.1 min, and its MPC remakes a 96-step plan every 15 min above the oxygen loops, in the place the ammonia supervisor would hold, as Figure 10.6 draws them.

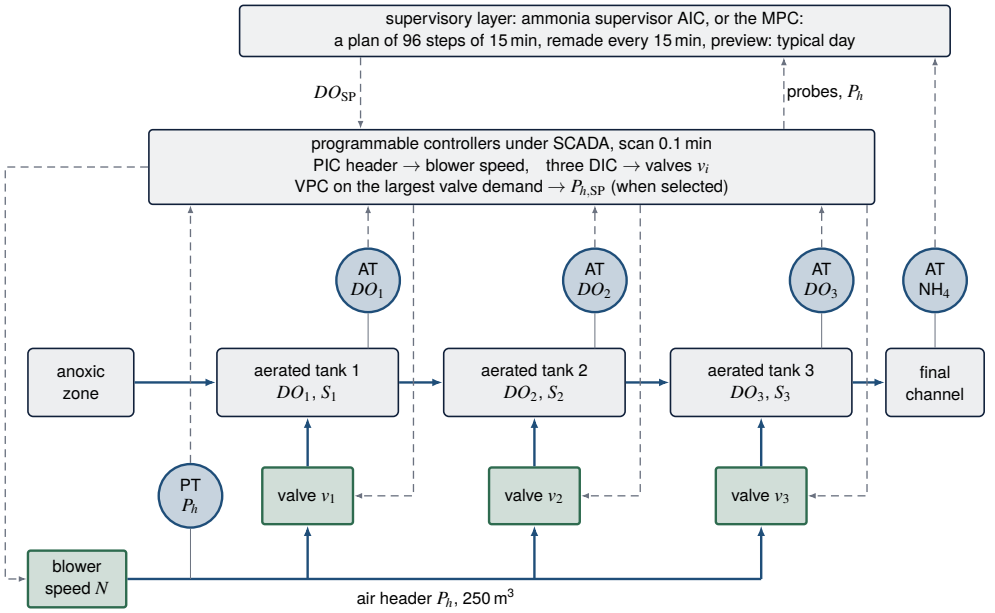



**Figure 10.6** The chapter's loops on field hardware, drawn from Sections 10.3 and 10.4. Green boxes are the final elements (blower speed and the three butterfly valves), blue circles the sensors (header pressure, one oxygen probe per tank and the ammonium analyzer on the final channel), dashed arrows signals. Notice that the supervisory layer moves no device: it writes one oxygen setpoint, and the three oxygen loops and the header loop turn it into air.

> **IN PRACTICE** **Hunting, and the Cascade Under the Oxygen Loop**
>
> The common automated arrangement is a cascade. Each basin's oxygen loop sets an airflow controller on its valve, and a pressure transducer on the main header drives the blowers, so a valve that opens pulls the header down and the pressure loop calls for more air. Oxygen takes minutes or longer to settle after the air changes, so a quick loop overshoots and moves the valve again, which operators call hunting. Conservative tuning stops it but leaves the loop slow, and the curved relation between oxygen and airflow makes tuning harder [4]. This chapter's oxygen loops move their valves directly, a variant the survey also names (Table 10.3).

### What Is Modernizing It

Three developments bear on this chapter's method. Optical probes are the first: they need less calibration and maintenance, and a more reliable probe could pave the way for more automated oxygen control. Model-based setpoints are the second. One commercial scheme runs an on-line process simulation from measured temperature, ammonia, nitrate and influent flow, with an ammonia analyzer in the anoxic zone and a nitrate analyzer at the end of the aerobic zones, and returns oxygen setpoints and the internal recycle flow in real time [4]. This chapter's MPC takes the same feedforward step with a calendar preview. The third is a direct reading of transfer: a low-cost monitor of oxygen transfer efficiency helps predict when diffusers need cleaning, and a study estimated an average energy efficiency improvement of 15 percent [4]. That is the quantity scenario 13 shows a probe cannot see.

## 10.7 Limitations

A single day's scorecard is not a steady state, and this is the limitation that most affects how the results above should be read. The supervisor is seeded bumplessly from whatever dissolved oxygen setpoint

was in force, and with an eight-hour integral time it takes about a day to forget it, which is why scenario 7 is worth 4.4 % at equal compliance on its first day and 1.5 % on its third. That behavior is a real property of slow supervisory loops rather than a numerical artifact, and the application exists partly to make it visible, but any figure quoted from one simulated day has to say which day it was.

The biology is thinner than the air system. One lumped autotroph population with Monod kinetics cannot show nitrite accumulation, and nitrous oxide is absent altogether, which removes the mechanism that makes a very low setpoint genuinely expensive rather than merely risky [18]. Sludge age is a parameter rather than the consequence of a wasting flow, so the application shows what a solids retention time decision does and not what it costs in settlement capacity. There is no clarifier at all, so the failure modes that actually take plants out of service are outside the model. The heterotrophic demand is a forcing rather than a state, which sharpens the morning air peak.

The air system is more complete and still simplified. One machine with a fitted head curve stands in for a hall of three, so the turndown behavior of scenario 10 is the turndown of a staged hall represented by a single surge line rather than the turndown of one impeller, and no answer involving staging can be tested. Oxygen transfer is one power law with alpha, beta and fouling lumped into a single slider, and the mid-depth approximation in (10.9) is a standard simplification rather than a profile [12]. The valve model has no stiction, no hysteresis and no finite stroking rate beyond its 0.25 min travel time, any of which would add a limit cycle that tuning cannot remove.

The instrument chain fails only by being slow. Drift, a stuck reading that looks plausible and an instrument health signal are all absent, and a supervisory loop that trusts a frozen analyzer is a more common industrial failure than one that is badly tuned. The numerics are also more repeatable than a control system, because a fixed integration step and a fixed scan reproduce one trajectory on every machine while a distributed control system does no such thing.

Three scenarios are stylized on purpose. A diffuser condition that falls from 0.55 to 0.38 in five hours compresses a year of fouling into one run, which no physical grid does, but it isolates a cost that is invisible on the controlled variable. The returned liquor stream arrives as a step, because a step is the disturbance whose response can be read off a trend, while a real dewatering run ramps. The cold snap moves the sewage three and a half degrees in a few hours rather than over a week, so that the population and the ammonia can be watched together inside a three-day run.

The MPC's model of (10.20) is thinner again. It fixes the nitrifier population, the basin at 13 °C, the diffusers at $\alpha F = 0.55$ and the carbonaceous load, omits the nitrate pools and the limits of the loops and valves beneath it, and holds the header pressure for a day. Open loop on scenario 7's day its ammonia error is 0.145 g N/m$^3$ RMS, but 1.93 g N/m$^3$ at 9.5 °C and 4.16 g N/m$^3$ at a 1.45 strength, and at $\alpha F = 0.38$ its power error is 46 % of the mean. Its preview is a calendar exact only on the application's ordinary day, its ceiling is chosen by hand, and since nothing in its objective values the nitrogen discharged it spends every margin.

> ⚠ **OPEN CHALLENGE** **Savings Claims Without Independent Data**
>
> A federal survey found that control of nitrification on ammonia, nitrate and nitrite measurements was growing overseas but had few full-scale applications in the United States. For one commercial scheme it reported savings claimed by the supplier that no independent full-scale data had verified, and it warned that savings are site specific and depend on the control in place before the upgrade [4]. This chapter's first and settled days show how far one answer moves with the day and the reference chosen. Progress would be the independent full-scale data the survey found missing, read at equal compliance as Section 10.5 reads its own runs.

## 10.8 Discussion

The plant carries three decisions with different answers, and separating them is the purpose of the application.

The first is where a setpoint should sit, and it is an economic question with a legal constraint rather than a control question. Table 10.5 is a frontier and not an optimum, and the shape of it comes from two saturating mechanisms working in opposite directions. What a milligram of dissolved oxygen buys falls because (10.3) is a Monod term, and what it costs rises because the air needed goes as $(C_s^* - DO)^{-1/n}$. The product is a cost per unit of ammonia removed that is 56.6 times larger between 2.5 and 3.5 mg/L than it is between 0.5 and 1.5 mg/L. Choosing a point on such a curve is exactly the self-optimizing control question of which variable to hold and at what value [5], and the answer here is that no fixed point on it is right for both February and June.

The second is whether a supervisory loop earns its instrument, and the honest answer on this works is that it does not earn it on electricity. A settled saving of 66 kWh a day at twelve cents is about $2,900 a year against an ammonium analyzer costing several times that in capital and a few thousand a year in reagent, and the accounting only gets worse if the comparison is made at equal setpoint rather than at equal compliance [16]. What the loop actually buys is the February margin. On the graded day at 12.8 °C a fixed 2.0 mg/L finishes at 2.39 g N/m$^3$ and needs raising to 2.25 mg/L to hold 2.160 g N/m$^3$, which is 3995 against 4159 kWh, four percent more every day of the year because nobody lowers it again in June. A supervisor spends that four percent only on the days that need it, and the general shape of that argument, a small continuous prize against a large occasional one, recurs in most supervisory control business cases [15, 17].

The third is the one with no vendor. Most-open-valve control removes about 6 kPa of manufactured and destroyed pressure for 267 kWh a day and moves no controlled variable by more than 0.113 mg/L, which is four times what the analyzer loop returns on a settled day. Its absence from most plants is not a technical matter, because the header pressure is a commissioning number that nobody feels entitled to change, and the same structure and the same reticence appear wherever throttled users share a supply header [6]. Diffuser fouling belongs in the same category, since a hundred kilowatts of drift is invisible on every alarm list and visible immediately in oxygen transferred per kilowatt-hour [13].

Beyond the four loops implemented, the same equations carry an optimization problem. A sixteen-state balance with actuator limits, a surge constraint and a rolling-mean permit suits model predictive control, because the permit of (10.15) is a constraint on a moving average rather than a setpoint and an optimizer can hold it directly instead of through a margin chosen by hand [27]. The predictability of the load curve makes the same formulation a feedforward one, which is the distinction practice guidance draws within ammonia-based control [17]. An estimator running against the same balances recovers the diffuser condition and the nitrifier population from airflow, power and the analyzer, which turns scenario 13 from a lesson into a soft sensor. Section 10.4.5 builds the first two of these ideas, as an economic MPC in the sense of Section 1.4.8, but not the third, and still needs a ceiling chosen by hand 0.3 g N/m$^3$ inside the permit, of which its model error spends up to 0.018 g N/m$^3$.

## 10.9 Conclusion

This chapter studied an interactive activated sludge works whose regulatory loops are easy and whose setpoint is not. The simulator integrates sixteen states covering ammonia and nitrate through four zones, dissolved oxygen in three aerated tanks, one nitrifier population, an air header, a blower and three valves, with a 17.3 s pressure loop under a 6.20 min oxygen loop under an ammonia measurement that answers in three hours and is judged as an eight-hour mean. One physical number sizes every result, because 5 m of submergence puts the saturation the bubbles see at 12.39 mg/L, so a 2 mg/L setpoint is 16.1 % of the driving force and the whole practical setpoint range spans 32 % of a day's aeration energy. Inside that range, ammonia-based control read at equal compliance is worth 194 kWh of 4446 kWh on its first day and 66 kWh of 4355 kWh once its eight-hour integral has settled, while most-open-valve control on the header is worth 267 kWh of 4335 kWh with the rolling mean unmoved at 2.27 g N/m$^3$. A supervisor tuned like a regulatory loop at 10 min of integral time ends the day dirtier than no supervisor at all, at 2.440 g N/m$^3$ against 2.277 g N/m$^3$. A cold snap leaves the nitrifier population at 102 g VSS/m$^3$ and puts the entire loss on the effluent, because a seven-day

sludge age fixes the growth rate the population is permitted to have. A model predictive option planning the eight-hour mean a day ahead uses 278 kWh less than the supervisor at 1.30 g N/m$^3$ on the settled day, a saving owed to the typical load curve and to the permit's own measure, which it loses without that curve and pays for with 24 % more nitrogen discharged.

A fuller model would add a two-step nitrification with nitrite, a nitrous oxide term that prices a low setpoint properly, a clarifier with a solids inventory, and a staged blower hall so that the turndown answer can be designed rather than diagnosed. As it stands, the simulator is a place to decide where a supervisory setpoint belongs and to quote honestly what moving it is worth.

**KEY POINTS**

- The works is judged on an eight-hour mean of effluent ammonia, while its loops hold an oxygen setpoint no permit names.
- Read at equal compliance, the ammonia supervisor is worth 194 kWh on its first day and 66 kWh on a settled day, and most-open-valve control 267 kWh with nothing given up.
- The day-ahead MPC that plans the permit's own mean uses 277 kWh, or 6.5 %, less than the supervisor on the settled day, and loses that saving without the typical-day preview.
- The plan pays with 24 % more nitrogen discharged.

**Where else this applies.** A town's water network (Chapter 11) also plans a day of pumping against a tariff, and there the band rule breaks the pressure floor on every seed while the scheduled plan with modulated valves passes all eleven (Table 11.4).

## References

[1] D. E. Seborg, T. F. Edgar, D. A. Mellichamp, and F. J. Doyle, III, *Process Dynamics and Control*, 4th ed. Hoboken, NJ: John Wiley & Sons, 2016.

[2] S. Longo, B. M. d'Antoni, M. Bongards, A. Chaparro, A. Cronrath, F. Fatone, J. M. Lema, M. Mauricio-Iglesias, A. Soares, and A. Hospido, "Monitoring and diagnosis of energy consumption in wastewater treatment plants. A state of the art and proposals for improvement," *Applied Energy*, vol. 179, pp. 1251–1268, 2016.

[3] U.S. Environmental Protection Agency, "Energy efficiency in water and wastewater facilities: A guide to developing and implementing greenhouse gas reduction programs," U.S. Environmental Protection Agency, Washington, DC, Local Government Climate and Energy Strategy Series EPA-430-R-09-038, 2013.

[4] ——, "Evaluation of energy conservation measures for wastewater treatment facilities," U.S. Environmental Protection Agency, Office of Wastewater Management, Washington, DC, Tech. Rep. EPA 832-R-10-005, Sep. 2010.

[5] S. Skogestad, "Plantwide control: the search for the self-optimizing control structure," *Journal of Process Control*, vol. 10, no. 5, pp. 487–507, 2000.

[6] C. L. Smith, "Valve position control," in *Advanced Process Control: Beyond Single-Loop Control*. Hoboken, NJ: John Wiley & Sons, 2010, pp. 179–197.

[7] E. Ardern and W. T. Lockett, "Experiments on the oxidation of sewage without the aid of filters," *Journal of the Society of Chemical Industry*, vol. 33, no. 10, pp. 523–539, 1914.

[8] J. Monod, "The growth of bacterial cultures," *Annual Review of Microbiology*, vol. 3, no. 1, pp. 371–394, 1949.

[9] G. Tchobanoglous, H. D. Stensel, R. Tsuchihashi, and F. L. Burton, *Metcalf and Eddy Wastewater Engineering: Treatment and Resource Recovery*, 5th ed. New York, NY: McGraw-Hill, 2014.

[10] M. Henze, C. P. L. Grady, W. Gujer, G. V. R. Marais, and T. Matsuo, "A general model for single-sludge wastewater treatment systems," *Water Research*, vol. 21, no. 5, pp. 505–515, 1987.

[11] G. Olsson, "ICA and me – A subjective review," *Water Research*, vol. 46, no. 6, pp. 1585–1624, 2012.

[12] M. K. Stenstrom and R. G. Gilbert, "Effects of alpha, beta and theta factor upon the design, specification and operation of aeration systems," *Water Research*, vol. 15, no. 6, pp. 643–654, 1981.

[13] D. Rosso, L. E. Larson, and M. K. Stenstrom, "Aeration of large-scale municipal wastewater treatment plants: state of the art," *Water Science and Technology*, vol. 57, no. 7, pp. 973–978, 2008.

[14] L. Rieger, J. Alex, W. Gujer, and H. Siegrist, "Modelling of aeration systems at wastewater treatment plants," *Water Science and Technology*, vol. 53, no. 4-5, pp. 439–447, 2006.

[15] L. Åmand, G. Olsson, and B. Carlsson, "Aeration control – a review," *Water Science and Technology*, vol. 67, no. 11, pp. 2374–2398, 2013.

[16] L. Åmand and B. Carlsson, "Optimal aeration control in a nitrifying activated sludge process," *Water Research*, vol. 46, no. 7, pp. 2101–2110, 2012.

[17] L. Rieger, R. M. Jones, P. L. Dold, and C. B. Bott, "Ammonia-based feedforward and feedback aeration control in activated sludge processes," *Water Environment Research*, vol. 86, no. 1, pp. 63–73, 2014.

[18] M. J. Kampschreur, H. Temmink, R. Kleerebezem, M. S. M. Jetten, and M. C. M. van Loosdrecht, "Nitrous oxide emission during wastewater treatment," *Water Research*, vol. 43, no. 17, pp. 4093–4103, 2009.

[19] D. Q. Mayne, J. B. Rawlings, C. V. Rao, and P. O. M. Scokaert, "Constrained model predictive control: Stability and optimality," *Automatica*, vol. 36, no. 6, pp. 789–814, 2000.

[20] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[21] M. O'Brien, J. Mack, B. Lennox, D. Lovett, and A. Wall, "Model predictive control of an activated sludge process: A case study," *Control Engineering Practice*, vol. 19, no. 1, pp. 54–61, 2011.

[22] D. Goldfarb and A. Idnani, "A numerically stable dual method for solving strictly convex quadratic programs," *Mathematical Programming*, vol. 27, no. 1, pp. 1–33, 1983.

[23] M. Diehl, H. G. Bock, and J. P. Schlöder, "A real-time iteration scheme for nonlinear optimization in optimal feedback control," *SIAM Journal on Control and Optimization*, vol. 43, no. 5, pp. 1714–1736, 2005.

[24] G. Pannocchia and J. B. Rawlings, "Disturbance models for offset-free model-predictive control," *AIChE Journal*, vol. 49, no. 2, pp. 426–437, 2003.

[25] J. W. O'Dell, "Method 350.1: Determination of ammonia nitrogen by semi-automated colorimetry, revision 2.0," U.S. Environmental Protection Agency, Environmental Monitoring Systems Laboratory, Cincinnati, OH, Tech. Rep., Aug. 1993.

[26] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication 800-82r3, 2023.

[27] J. D. Hedengren, R. Asgharzadeh Shishavan, K. M. Powell, and T. F. Edgar, "Nonlinear modeling, estimation and predictive control in APMonitor," *Computers & Chemical Engineering*, vol. 70, pp. 133–148, 2014.

## 10.10 Exercises

**10.1** Concept check · 2 min

In Scenario 11 the sewage drops from 13 to 9.5 °C for two days with the ammonia supervisor in service. The nitrifier population wanders between 97.6 and 101.9 g VSS/m$^3$ and ends where it began, while the effluent ammonia triples. Why does the population not fall?

(a) The supervisor raises the dissolved oxygen to its ceiling, which gives back the growth rate the cold took away

(b) Decay slows with the same temperature coefficient as growth, so the two effects cancel in the balance

(c) Wasting fixes the average growth rate at steady state, so the ammonia rises until the nitrifiers grow that fast again

(d) Three days is too short for a state with a time constant of days, so a longer run would show the loss

**10.2** Concept check · 1 min

The most-open-valve controller reads the largest *unsaturated* output of the three dissolved oxygen loops, not the measured valve positions. Why?

(a) The unsaturated output is quieter than a stem position, so the slow loop can carry a higher gain

(b) The valves take a quarter of a minute to travel, a lag that would destabilize a 45-minute integral

(c) Reading the demand lets the loop act before the dissolved oxygen moves, as a feedforward would

(d) A valve pinned fully open cannot report that it wanted more, so a position reading stops raising the header

**10.3** Concept check · 2 min

On its first day the ammonia supervisor of Scenario 7, aimed at 1.90 g N/m$^3$, saves 194 kWh against a fixed setpoint at nearby compliance. Run on with nothing changed, it settles at 4109 and 4105 kWh with worst rolling means of 2.563 and 2.598 g N/m$^3$, outside the permit. What made the first day look better than the loop is?

(a) Its eight-hour integral needs about a day to forget the 2.0 mg/L it was seeded from, so day one ran richer than its aim

(b) The nitrifier population declined at the low overnight oxygen, so the later days needed less ammonia removed

(c) Analyzer noise accumulated in the integral over three days and pushed the setpoint away from its aim

(d) The second and third days carried a lighter load, so the same aim let more ammonia through

**10.4** Theory and calculation · 6 min

Equation (10.9) gives the saturation the bubbles see. Take $\beta = 0.95$, a clean-water solubility $C_{s,w} = 10.50\,\mathrm{mg/L}$ at 13 °C, $\rho = 1000\,\mathrm{kg/m^3}$, $g = 9.81\,\mathrm{m/s^2}$, $P_{atm} = 101.325\,\mathrm{kPa}$ and $h = 5\,\mathrm{m}$.

(a) Compute $C_s^*$. *Checkpoint:* 12.39 mg/L.

(b) What share of the transfer driving force does a 2 mg/L setpoint remove? *Checkpoint:* 16.1 %.

(c) At a fixed uptake $R$, the steady air is $q \propto [R/(C_s^* - DO)]^{1/n}$ with $n = 0.75$. Find how the air for one more mg/L depends on $DO$, and compute its ratio at 3.5 mg/L against 0.5 mg/L. *Checkpoint:* $\propto (C_s^* - DO)^{-(1/n+1)}$, a ratio of 1.97.

**10.5** Theory and calculation · 7 min

At the flat average load the works treats 20 000 m$^3$/d at 30 g N/m$^3$, assimilates a fraction $f_{ass} = 0.12$ of it and discharges 0.378 g N/m$^3$. Use $V_t = 1700\,\mathrm{m^3}$, $Y_A = 0.24$, $b_{A,20} = 0.12/\mathrm{d}$, $\theta_B = 1.029$, 13 °C and a seven-day sludge age.

(a) Compute the steady mean growth rate $\bar{\mu}_A$ from Equation (10.4). *Checkpoint:* 0.2411/d.

(b) Summing Equation (10.2) over the three tanks, the nitrifiers remove $3\bar{\mu}_A X_A V_t / Y_A$ of ammonia a day. At steady state this equals the load less assimilation less the effluent. Compute $X_A$. *Checkpoint:* 520.4 kg N/d nitrified, $X_A = 101.6\,\mathrm{g\,VSS/m^3}$.

(c) Cut the sludge age to five days. Compute $\bar{\mu}_A$, and $X_A$ if the nitrified mass did not change. The model's steady state is 79.5 g VSS/m$^3$. What effluent ammonia does that imply? *Checkpoint:* 0.2982/d, 82.1 g VSS/m$^3$, about 1.2 g N/m$^3$.

**10.6** Theory and calculation · 7 min

The header balance of Equation (10.11) sets how fast the inner loop can be. At the design point the three valves pass 122.1 Nm$^3$/min from a header at 62.0 kPa, the back pressure is $P_{back} = 53.05\,\mathrm{kPa}$, the blower curve gives −1.72 Nm$^3$/min per kPa, and $V_h = 250\,\mathrm{m^3}$.

(a) From Equation (10.10), show that $\partial \sum q_i / \partial P_h = \sum q_i / (2(P_h - P_{back}))$ at fixed valve positions, and evaluate it. *Checkpoint:* 6.82 Nm$^3$/min per kPa.

(b) Linearize Equation (10.11) at fixed speed and compute the header time constant. Compare it with the 6.20 min of an aerated tank. *Checkpoint:* 0.289 min or 17.3 s, about 21 times faster.

(c) In Scenario 1 opening valve 1 pulls the header from 62.0 to 59.7 kPa. Valves 2 and 3 do not move and passed 36.6 and 17.8 $Nm^3$/min. Predict the flow each loses. *Checkpoint:* About 5.0 and 2.5 $Nm^3$/min (the run reads 4.9 and 2.4).

**10.7** App challenge · 7 min

Load Scenario 5, *5 · A whole day at a fixed setpoint*, at the default seed and run it past 1440 min at a speed of 60 (60 simulated minutes per second, about 25 s). Export the CSV (its first lines start with #) and use the row at `time_min` = 1440.

(a) Read `metric_energy_kWh`, the largest `nh4_avg_gNm3` after 480 min, and the largest `nh4_gNm3` with its clock time (`clock_h`). *Checkpoint:* 4335 kWh (±1), 2.275 g N/$m^3$ (±0.003), 3.27 g N/$m^3$ at about 11:45.

(b) Load Scenario 6, *6 · The same day, cheaper, and over the consent*, as shipped (1.2 mg/L), and repeat. Also read `metric_consent_min` at 1440. What does the lower setpoint save, and what does it cost? *Checkpoint:* 3813 kWh, 3.784 g N/$m^3$, 4.79 g N/$m^3$, 603 min over the permit. A saving of 12 %.

(c) The frontier of Table 10.5 puts 1.2 mg/L at 3819 kWh and 3.644 g N/$m^3$. Why does Scenario 6 differ? *Checkpoint:* A different opening state.

**10.8** App challenge · 9 min

Load Scenario 7, *7 · Ammonia-based aeration control*, at the default seed and run it past 4320 min at a speed of 200 (about 22 s). Export the CSV. A day's energy is the difference of `metric_energy_kWh` between the rows at its start and end, and its worst mean is the largest `nh4_avg_gNm3` from 480 min into the day to its end.

(a) Give each day's energy and worst mean, and `metric_consent_min` at 4320. Tolerances: 1 kWh and 0.003 g N/$m^3$. *Checkpoint:* 4253, 4109, 4105 kWh. 2.134, 2.562, 2.597 g N/$m^3$. 222 min.

(b) The fixed 2.0 mg/L setpoint of Scenario 5 spends 4335, 4341 and 4342 kWh on three days at worst means of 2.277, 2.245 and 2.221 g N/$m^3$ (Table 10.6). Which conclusion does day one support, and which does day three? *Checkpoint:* Day one: cheaper and cleaner. Day three: cheaper and illegal.

(c) Reload the scenario, set *Effluent ammonia setpoint* to 1.30 before Run and repeat (a). Compare the third day with the fixed 2.02 mg/L of Table 10.6, 4355 kWh at 2.201 g N/$m^3$. *Checkpoint:* 4471, 4307, 4289 kWh. 1.918, 2.182, 2.203 g N/$m^3$. 66 kWh or 1.5 % saved.

**10.9** App challenge · 10 min

Load Scenario 14, *14 · The eight-hour mean, planned a day ahead (MPC)*, at the default seed and run it past 4320 min at a speed of 200. Read each day as in the previous exercise. Under the MPC the dashed setpoint line of the ammonia trend shows the fallback supervisor's 1.90 g N/$m^3$, not the plan's ceiling.

(a) Give each day's energy and worst mean, and the largest `nh4_gNm3` on day three with its clock time. *Checkpoint:* 4018, 4010, 4011 kWh. 2.210, 2.218, 2.206 g N/$m^3$. 3.22 g N/$m^3$ at about 20:37.

(b) Reload, set *MPC load preview* to *Off* before Run and repeat the energies and worst means. Read `metric_consent_min` at 4320. *Checkpoint:* 4306, 4309, 4306 kWh. 2.595, 2.628, 2.602 g N/$m^3$. 362 min.

(c) The supervisor at 1.30 g N/m$^3$ spends 4289 kWh on its third day and discharges 29.0 kg of nitrogen, the MPC 35.9 kg. What does the plan save, and how many kilowatt-hours does it buy with each extra kilogram of nitrogen? *Checkpoint:* 278 kWh, 6.5 %, about 40 kWh per kg.

**10.10** Application area · 4 min

The chapter prices three changes to the aeration control of this works at twelve cents a kilowatt-hour: the settled ammonia supervisor saves 66 kWh a day, most-open-valve control 267 kWh a day, and a fixed setpoint chosen for the February day (2.25 mg/L, 4159 kWh) costs more than 2.0 mg/L with the same header loop (3995 kWh) on every day it is left in place.

(a) Put each in dollars a year. *Checkpoint:* About $2,900, $11,700 and $7,200.

(b) An ammonium analyzer costs several times the first figure in capital and a few thousand dollars a year in reagent. Which change would you make first, and what does the analyzer actually buy? *Checkpoint:* The header loop. The analyzer buys not having to set the worst-day setpoint all year.

**10.11** Application area · 4 min

The works has one variable-speed blower whose surge flow at full speed is 82 Nm$^3$/min, falling in proportion to speed, with a minimum speed of 48 % and an anti-surge gain $k_{vent} = 2$. In Scenario 10, a dry Sunday night, the tanks take 31.3 Nm$^3$/min at 04:10 while the machine passes 37.1 Nm$^3$/min against a surge line of 40.0 Nm$^3$/min.

(a) What is the surge flow at the minimum speed, and what does Equation (10.13) vent at 04:10? Check the balance. *Checkpoint:* 39.4 Nm$^3$/min. 5.8 Nm$^3$/min, equal to $37.1 - 31.3$.

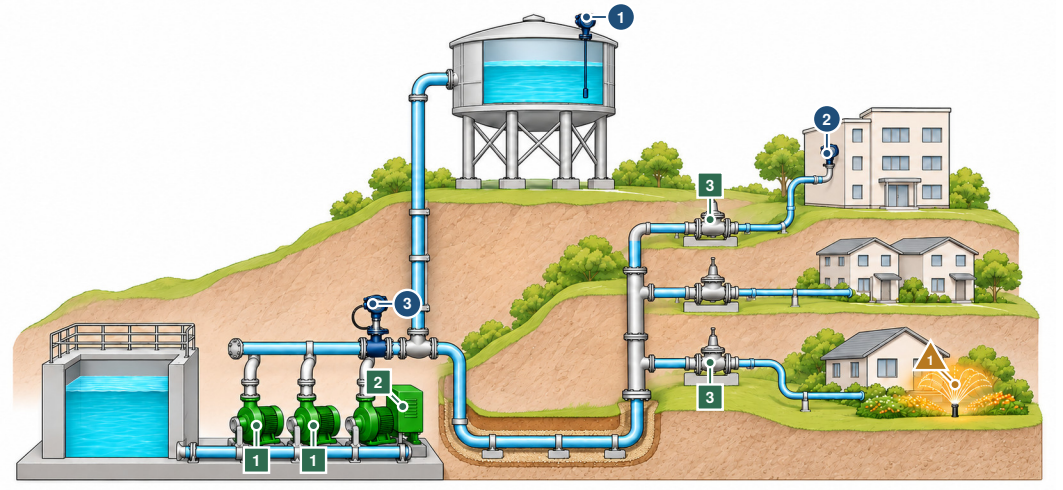
(b) At that load, taking the setpoint from 2.0 to 0.5 mg/L cuts the air to the tanks by 19 % and the power by 4.5 %. Explain, and name two equipment remedies and the number that sizes them. *Checkpoint:* The vent makes up the air the tanks stop taking. A smaller companion blower, wider turndown, or a basin taken offline, sized on the vented volume.

**10.12** Review the agent · 5 min

An AI assistant reviews the graded challenge, Scenario 15, and writes: *"Press Reveal suggested tuning, then set Aeration control to MPC. The plan holds its 2.2 g N/m$^3$ ceiling on the eight-hour mean, which is exactly the challenge's 2.20 g N/m$^3$ clause, and it plans the day ahead, so the day passes with about 350 kWh more to spare than the suggested supervisor."* Find the flaw and say how you would check it in the app.

CHAPTER 11

# Water Distribution Simulator That Schedules Integer Pumps Under a Pressure Floor




**The controller moves** 1 two fixed-speed pumps, on or off 2 pump on a variable-speed drive 3 pressure-reducing valves

**The controller reads** 1 tank level 2 pressure at the highest customer 3 station flow

**What disturbs it** 1 customer demand

*The pumps lift water to the hilltop tank, which feeds the zones through their valves, and the hour the pumps run decides both the energy bill and the pressure at the highest house. Illustration, not to scale.*

### IN PLAIN TERMS When the Pumps Run Decides the Bill

**The main idea.** A town's pumps lift water up to a tank on a hill, and the tank feeds the homes below. Power costs more in the evening, so a planner decides each hour which pumps to run.

**The challenge.** Each pump is either on or off, so the planner must choose among a huge number of patterns. The tank must stay high enough to give the top house its pressure, but higher pressure pushes more water out through leaks.

**Why it matters.** The usual rule, starting a pump when the tank runs low, pays far more for power than the planner and still lets the top house lose pressure.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 11.2, the try-it and fact segments in Sections 11.4 and 11.6, and the key points |
| **Study** | 45 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 15 min | the first page, the results (Section 11.6), Technology in Practice (Section 11.7), Limitations (Section 11.8) and the open challenge |

**The control question** A town's pumps are on or off, its tank is the only store, and the same tank level and valve settings decide the pumping bill, the pressure at the highest house and the water lost through leaks. When should the pumps run, and what does an integer schedule re-solved every hour add to a level band and to pressure management?

**Learning objectives**

- Compute a pump's operating point, power and least working speed from its curve and the affinity laws.
- Explain why a level band starts its pumps at the wrong hour and lets the hill lose its pressure floor first.
- Interpret an LP relaxation, its fractional pumps and the gap that branch and bound proves.
- Measure what a re-solve after a pump trip and modulated valves are each worth on the same day.
- Recognize a floor that no schedule and no valve setting can hold.

**Prerequisites** Sections 1.4.3, 1.4.6, 1.4.8, 1.4.10, 1.6.2 and 1.6.4, and Chapter 9 for storage bought against a tariff.

**The app** https://apmonitor.com/apps/water-network-pumping/

**See also** Volume 8, Chapter 14 for integer servers ordered ahead of a boot delay, and Volume 8, Chapter 12 for on-off compressors staged on one header and the starts they cost.

## 11.1 Introduction

At midnight a hilltop tank 600 $m^2$ in plan holds 4.0 m of water on a floor 68 m above the clearwell that feeds it. On the same day, the same town and the same pumps, a rule that starts a pump when the tank falls to 2.0 m and stops it at 6.0 m spends \$241.9 on energy and buys 571 kWh of it at the \$0.20/kWh evening price, and it lets the highest customer fall under a 25 m pressure floor at 06:20, thirteen minutes before the lead pump starts. An hourly mixed-integer schedule spends \$128.1, buys nothing at the peak price and holds the floor. The tank is the same. What differs is when the pumps run.

The electricity for pumping is a large part of the operating cost of a water supply network [1], and scheduling it against a tariff has been framed as an optimization problem since the 1970s, solved by dynamic programming [2, 3], linear programming [1] and, more recently, branch and bound over integer pump states [4], and reviewed as a whole for its models, forecasts and methods [5, 6]. Pressure completes the picture, since leaks discharge more as the pressure rises [7, 8], and reducing excess pressure through valve settings is an established route to lower losses [9–11]. Predictive and real-time control of urban water networks gathers both in one controller [12, 13]. The broader class is any plant whose actuators are on or off, whose inventory buffers a demand against a time-varying price, and whose losses depend on a state the controller chooses. Unit commitment of generators, chiller and compressor staging and batch scheduling have that shape. The refrigeration pack of Volume 8, Chapter 12 stages two fixed-speed compressors and one on a drive, the same arrangement as the pumps here, and the branch and bound of Land and Doig was written for exactly such discrete programs [14].

The plant targets one belief in particular, that a full tank is a safe tank and that a pump should run whenever the level is low. In this model a full tank at the wrong hour raises the energy bill and the leakage together, while a tank filled at night and drawn through the evening peak holds the floor at a lower cost. However, the literature a student meets first divides the problem in two. Scheduling studies fix the hydraulics and minimize cost [1, 4], pressure management studies fix the pumping and minimize leakage [10], and the public-domain network solver of the U.S. Environmental Protection Agency simulates the hydraulics of a schedule it is given rather than choosing one [15]. This chapter works with a browser simulator of a small town's water system in which the integer schedule, the valves, the pressure floor and the pressure-dependent leakage are coupled in one run. It measures what a level band, an hourly branch and bound, modulated valves and a re-solve after a pump trip achieve on the same day. The steady network equations, the leakage law, the linear program, the branch and

bound and the proportional-integral loops are all established methods. What the simulator adds for study is a quasi-static network whose node and tank balances close to the integrator's precision, a branch and bound on the framework's shared simplex that reports its relaxation, tree and gap to the learner, and a strict split between instruments and truth, so that no controller or estimator reads the simulated state. Ten cases with completion rules run from pumps worked by hand to a two-pump loss that no schedule can survive and a hot-day challenge.

Section 11.2 traces the domain. Section 11.3 sets out the network, its instruments and its parameters, and Section 11.4 shows why the hour a pump runs decides both the hill's pressure and the leakage. Section 11.5 describes the level band, the hourly schedule and its branch and bound, the flow and valve loops and the two estimators. Section 11.6 measures the ten cases and the challenge and verifies the code, Section 11.7 describes the instruments, pumps, valves and control platforms of real networks, Section 11.8 states the limitations, and the exercises in Section 11.10 run the cases.

## 11.2 From Pumping Costs to Pressure Management

A dynamic programming solution for the pumping costs of a water supply network was published by 1975 [2], and progressive optimality, an iterative dynamic program over the hours of the day and the subsystems of a network, scheduled eight reservoirs and seven pumping stations against a 24-hour demand forecast and a varying energy cost in 1989 [3]. Linear programming entered with a 24-hour minimum-cost schedule that accounted for pump efficiencies, the tariff structure and the network's linearized hydraulics, solved quickly enough for real-time use [1]. A 1994 review classified the methods of the time by system type, hydraulic model, demand model and optimization method, linear, dynamic or nonlinear programming [5]. Branch and bound formulations with piecewise linear component models later found schedules whose performance in a full hydraulic simulation lay within the solver's optimality gap and the uncertainty of the demand forecast [4]. Reviews in 2017 and 2019 collected the operational optimization and real-time control literature that grew from these roots [6, 13], and model predictive control of drinking water and drainage networks was reported as an application of control in 2013 [12].

Leakage followed a separate path. Network models that included pressure- dependent demand and leakage terms were proposed in 1985 [7], and excess pressure was minimized through valve settings to reduce leakage in 1989 [9]. Optimal settings of flow reduction valves followed in 1998 [10], and open- and closed-loop pressure control for leakage reduction in 2000 [11]. The power law between leak flow and pressure carries an exponent $N_1$ that is 0.5 for a fixed orifice and rises toward 1.5 when the leak opening itself grows with pressure, the fixed and variable area discharges (FAVAD) concept [16], and an analysis of elastically deforming leaks showed the exponent tending to 0.5 at low and 1.5 at high pressure [8]. The minimum service pressure itself is set by design standards. The recommended standards of a board of state and provincial health and environmental managers ask for at least 20 psi (140 kPa) at ground level at every point under all conditions of flow and a normal working pressure not under 35 psi (240 kPa) [17]. The simulator joins the two paths, because the hour a pump runs decides the tank level, and the tank level and the valves decide both the pressure at the hill and the leakage of the town.

> ⧖ HISTORY **1950: A Board of States Writes the Rules for Tanks and Pumps**
>
> In 1950 a board of state public health managers around the Great Lakes and the upper Mississippi created a Water Supply Committee, with one associate from each member state, to review water works practices, policies and procedures. A province joined in 1978. The committee's report was first published in 1953 and has been revised and republished many times since, as a guide to the design of public water supplies. Its chapter on storage asks that pumps be controlled from tank levels, with the signal sent by telemetry when appreciable head is lost between the source and the tank, and that level indication be provided at a central location. It also asks that the high and low levels of a tank that gives a system its pressure differ by no more than 30 feet [18]. The level band of Section 11.5 is a rule of that kind.

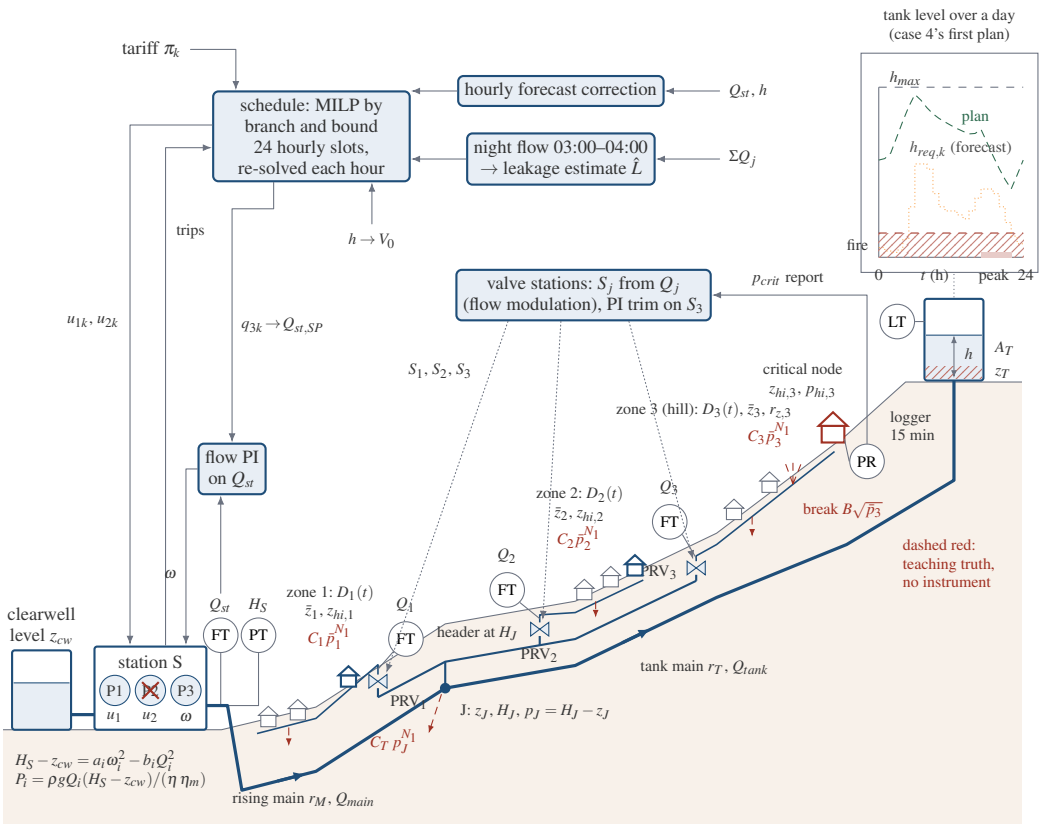



**Figure 11.1** The town's water system in elevation, as the Network view draws it, with the control layer above. The station S lifts water from the clearwell at $z_{cw}$ with pumps P1 and P2 on or off ($u_1, u_2$) and P3 at speed $\omega$, up the rising main ($r_M, Q_{main}$) to the junction J. From J the tank main ($r_T, Q_{tank}$) fills the tank of area $A_T$ on the summit, and the header at $H_J$ feeds each zone through a pressure-reducing valve of loss $k_v$ and outlet setting $S_j$. Each zone draws $D_j(t)$ through a main $r_{z,j}$ and leaks $C_j \bar{p}_j^{N_1}$ at its mean pressure, the trunk leaks $C_T p_J^{N_1}$, and a break in zone 3 discharges $B\sqrt{\bar{p}_3}$. Leakage and the break are teaching truth (dashed red) that no instrument reads. P2 is crossed as it is after the challenge's 18:00 trip. Signals arriving at a block are tagged by the instrument symbols ($Q_{st}, h, Q_j$) instead of drawn. The inset is case 4's first plan, made at midnight (dashed), against its floor rows $h_{req,k}$ (dotted), a forecast of the level the critical node needs with no pump running, not a measurement. Not to scale.

## 11.3 The Town as Modeled

The network is a declared teaching town, not any utility's, as shown in Figure 11.1. Water leaves a clearwell held at the treatment plant's level $z_{cw} = 100\,\text{m}$ through a station S of three pumps, two at fixed speed and one on a variable-speed drive, and climbs a 3 km rising main of resistance $r_M$ to a junction J in town at $z_J = 125\,\text{m}$. From J a 1.2 km tank main of resistance $r_T$ runs up to an elevated tank whose floor is at $z_T = 168\,\text{m}$, and a town header carries the junction's head $H_J$ to three pressure-reducing valves, one at each zone. Zone 3 lies on the hill, and its highest customer, at $z_{hi,3} = 142\,\text{m}$, is the critical node at which the pressure floor is judged. Mean customer demand is 57 L/s (25, 20 and 12 L/s by zone). The engine integrates in hours with a step of 10 s.

Each variable of Table 11.1 carries its role. The tank's volume is the only hydraulic state, because the network is solved as steady flow at every step. Nine further states are ledgers integrated beside it

**Table 11.1** Variables by Role, With Defaults

| Variable | Role | Default | Range |
|---|---|---|---|
| $p_{hi,3}$ | CV, critical node (true) | floor 25 m | 15–35 m floor |
| $h$ | CV, tank level | $h_0 = 4.0$ m | 0–7 m |
| $u_1$, $u_2$ | MV, fixed-speed pumps | off | 0 or 1 |
| $\omega$ | MV, pump 3 speed | stopped | 0, 60–100 % |
| $S_j$ | MV, valve outlet heads | 34, 34, 40 m above valve | 15–60 m |
| $D_j(t)$ | DV, zone demand | 25, 20, 12 L/s mean | noise 0–20 % |
| $\pi(t)$ | DV, tariff | 0.07, 0.10, $0.20/kWh | peak 17:00–22:00 |
| events | DV, trips, break, hot day | — | as scripted |
| $H_J$ | algebraic, junction head | solved | — |
| $Q_i$, $Q_j$ | algebraic, flows | solved | — |
| $V = A_T h$ | state, tank volume | 2400 m$^3$ | 0–4200 m$^3$ |
| $E$, $C_e$ | state ledgers, kWh and $ | 0 | — |
| water ledgers | state, m$^3$ (seven) | 0 | — |

so that the balances can be checked: energy, its cost, and the water pumped, delivered, leaked, lost to a break, spilled over the top, not served from an empty tank and added by a teaching shortcut.

### 11.3.1 Pumps, Pipes and Valves

Each running pump lifts water from the clearwell to the station head $H_S$ along a parabola, and the variable-speed pump's curve scales with its speed ratio by the affinity laws, as shown in Equation (11.1).

$$\begin{aligned} H_S - z_{cw} &= a_i \omega_i^2 - b_i Q_i^2, \qquad \omega_1 = \omega_2 = 1 \\ H_{up} - H_{down} &= r\,Q|Q|^{0.852}, \qquad r = \frac{10.67\,L}{C^{1.852} D^{4.871}} \\ H_{out,j} &= \min\left(S_j,\; H_J - k_v Q_j^2\right) \end{aligned} \tag{11.1}$$

where a check valve keeps any pump from running backward and a pump gives no flow past its shutoff head. The second line is the Hazen–Williams loss in meters and cubic meters per second. The source code of the EPANET 2 solver writes the same resistance as $4.727\,L C^{-1.852} d^{-4.871}$ in feet and cubic feet per second [15], which converts to 10.667, so the app's 10.67 is 0.03 % above it, a ratio of 1.00030 at 20, 60 and 120 L/s on the rising main. The third line is a pressure-reducing valve that holds its outlet head $S_j$ or, when the junction cannot supply it, lies wide open with the loss $k_v Q_j^2$. The drive stops pump 3 below 60 % speed. By the affinity laws the pump cannot lift water into the tank at all below 78.6 % speed with the tank empty and 82.6 % with it full.

### 11.3.2 Zones, Leakage and the Junction

A zone's flow is its demand plus the leakage its mean pressure drives, and its highest customer sees the outlet head less the full loss of the zone main, as shown in Equation (11.2).

$$\begin{aligned} Q_j &= D_j(t) + C_j\,\bar{p}_j^{N_1} + \delta_{j3}\,B\sqrt{\bar{p}_3} \\ \bar{p}_j &= H_{out,j} - \bar{z}_j - \tfrac{1}{2} r_{z,j} Q_j^{1.852} \\ p_{hi,j} &= H_{out,j} - z_{hi,j} - r_{z,j} Q_j^{1.852} \\ Q_{main} &= \textstyle\sum_i Q_i = Q_{tank} + \sum_j Q_j + C_T\,p_J^{N_1} \end{aligned} \tag{11.2}$$

where $p_J = H_J - z_J$, leakage is zero at or below zero pressure, and the break term $B$ is nonzero only after a main breaks in zone 3. The leakage law at $N_1 = 1.18$ is a teaching value inside the FAVAD range of 0.5 for fixed holes to 1.5 for flexible pipes [16]. At $N_1 = 0.5$ and 1.5 the law turns a halved pressure into 29.3 % and 64.6 % less leakage, where the article quotes 30 % and 65 %, and at 1.18 a tenth less pressure gives 11.7 % less leakage. The leakage coefficients are set so that leakage is about a fifth to a quarter of the water pumped.

The series elements are eliminated so that one unknown is left, the junction head $H_J$. For a trial $H_J$ each zone's flow is the root of an increasing function of $Q_j$, the station's flow the root of a decreasing one, and the junction's imbalance then falls strictly with $H_J$, so its root is unique. Every root is found by Newton's method safeguarded by a bracket to a declared tolerance of $10^{-7}$ m$^3$/s. Over every case the worst residual is $1.0 \times 10^{-8}$ m$^3$/s, no step fails to converge and the junction needs at most 35 iterations.

### 11.3.3 Tank, Demand and Power

The tank integrates exactly the flow the solve returns, demand follows a daily pattern with colored noise, and each pump's electrical power follows its efficiency curve, as shown in Equation (11.3).

$$\begin{aligned} A_T \frac{\mathrm{d}h}{\mathrm{d}t} &= Q_{tank}, \qquad 0 \le h \le h_{max} \\ D_j(t) &= \bar{D}_j \, \phi(t) \, m_{hot}(t) \, (1 + \sigma_D n_j(t)) \\ P_i &= \frac{\rho g Q_i (H_S - z_{cw})}{\eta(Q_i) \, \eta_m \, \eta_d^{\delta_{i3}}} \\ \eta &= \eta_{max}(2x - x^2), \qquad x = \frac{Q_i}{\omega_i Q_{bep,i}} \\ \frac{\mathrm{d}E}{\mathrm{d}t} &= \textstyle\sum_i P_i, \qquad \frac{\mathrm{d}C_e}{\mathrm{d}t} = \pi(t) \textstyle\sum_i P_i \end{aligned} \tag{11.3}$$

where $\phi(t)$ interpolates 24 hourly multipliers normalized to a mean of one (from 0.458 at night to 1.627 at the 07:00 morning peak, with a second peak at 18:00), $m_{hot} = 1.12 + 0.25 \exp(-((t_h - 16)/3)^2)$ on a hot day and 1 otherwise, and $n_j$ is first-order colored noise of correlation time 0.5 h drawn independently for each zone and held through each step. The tank spills at $h_{max} = 7$ m and pauses the run if it empties. The efficiency peaks at the best-efficiency flow, $x$ is capped at 1.9, $\eta_m = 0.94$ is the motor and $\eta_d = 0.97$ the drive on pump 3. Energy is priced at \$0.07/kWh from 23:00 to 06:00, \$0.20/kWh in the five-hour evening peak from 17:00 and \$0.10/kWh otherwise, and an optional solar tariff adds \$0.04/kWh from 10:00 to 15:00. Leaked water is valued at \$0.30/m$^3$. All prices are teaching values.

### 11.3.4 Instruments

Every controller reads only the instruments. The tank level transmitter reads $h$ with 1 cm of colored noise, the station flowmeter $Q_{st}$ with 0.5 %, the station and valve pressure transmitters with 0.1 m, each zone meter $Q_j$ at its valve with 1 %, and the energy meter with 1 %. A pressure logger at the critical house samples $p_{hi,3}$ every 15 min with 0.1 m of noise and holds each report until the next. Run contacts and trip relays report each pump, and the day's weather forecast is a declared input. Leakage, the break, the true demand and the true pressures are never measured, and the comma-separated (CSV) export of Section 1.6.4 carries them in columns prefixed `truth_`.

### 11.3.5 Parameters, Initialization and Numerics

Every parameter appears in Table 11.2, all of them teaching values and all adjustable in the Advanced panel. Each case starts at midnight with the tank at its case's level and the instruments lined out on the network as the case's pumps and valves leave it. Its noise starts from the case's seed, as Section 1.6.1 describes. The tank and ledgers are integrated by explicit Euler at 10 s, so the network is solved once per step and the tank integrates exactly the flow of that solve. Control decisions are taken once a minute, and the historian writes one row per simulated minute. With pump 1 alone filling the tank at 3 m, the solve gives 65.52 L/s at a head rise of 74.52 m, against a static lift of 71 m, and an independent bisection on the pump and system curves gives the same crossing.

**Table 11.2** Model Parameters (Teaching Values, All Adjustable)

| Symbol | Meaning | Value | Unit |
|---|---|---|---|
| $z_{cw}$, $z_J$, $z_T$ | clearwell, junction, tank floor | 100, 125, 168 | m |
| $A_T$, $h_{max}$, $h_{fire}$ | tank area, overflow, fire reserve | 600, 7, 1 | $m^2$, m, m |
| $L$, $D$ | rising main; tank main | 3000, 0.40; 1200, 0.40 | m |
| $L_{z,j}$, $D_{z,j}$ | zone mains 1, 2, 3 | 2500, 2000, 1500; 0.30, 0.30, 0.25 | m<br>m |
| $C$ | Hazen–Williams coefficient | 120 | — |
| $\bar{z}_j$, $z_{hi,j}$ | zone mean, highest customer | 105, 118, 132; 112, 126, 142 | m<br>m |
| $z_{V,j}$ | valve elevations | 110, 122, 130 | m |
| $k_v$ | valve loss wide open | 312 | $s^2/m^5$ |
| $a_i$ | shutoff heads, pumps 1, 2, 3 | 110, 105, 110 | m |
| $b_i$ | curve coefficients | 8264, 12500, 8264 | $s^2/m^5$ |
| $Q_{bep,i}$ | best-efficiency flows | 55, 40, 55 | L/s |
| $\eta_{max,i}$ | peak efficiencies | 0.80, 0.78, 0.80 | — |
| $\eta_m$, $\eta_d$, $\omega_{min}$ | motor, drive, least speed | 0.94, 0.97, 0.6 | — |
| $\bar{D}_j$ | mean customer demand | 25, 20, 12 | L/s |
| $\sigma_D$, $\tau_D$ | demand noise, correlation | 5, 0.5 | %, h |
| $N_1$ | leakage exponent | 1.18 | — |
| $C_j$, $C_T$ | leakage, zones, trunk[a] | 59.7, 57.9, 43.5; 29.7 | — |
| $B$ | break coefficient, zone 3[b] | 1.27 | — |
| $\pi$ | night, day, peak, solar | 0.07, 0.10, 0.20, 0.04 | $/kWh |
| $p_w$, $c_s$ | water value, wear per start | 0.30, 2 | $/$m^3$, $ |

[a] In $10^{-6}$ $m^3/s$ per $m^{N_1}$. [b] In L/s per $\sqrt{m}$.

### 11.3.6 Assumptions

The hydraulics are quasi-static, so pipe transients and water hammer at a trip, which last seconds, are outside a model whose lesson takes hours. Each zone is lumped into a valve, a main to its highest customer and a mean pressure for leakage, while a real network has thousands of pipes and loops. Demand is driven, not pressure dependent, so customers draw their pattern whatever the pressure, and the run pauses rather than starve the town if the tank empties. The pressure floor of 25 m, about 35 psi, the fire reserve of 1 m and the leakage law are teaching values. A real floor comes from the utility's service standard and the state's design rules [17], a real fire reserve from the fire flow and duration the local authority requires, and a real exponent from a night step test on the zone. Pump wear is a declared cost per start, and demand charges and water quality are absent.

## 11.4 Why the Hour Matters

### 11.4.1 A Tank That Sets the Hill's Pressure

With every pump off the critical node is fed by the tank alone, so its pressure falls with the level and with the town's draw on the tank main. The declared model holds the critical node at the floor plus a 0.5 m margin with no pump running when the tank stands at 0.28 m at half the mean demand, 1.53 m at the mean and 3.73 m at 1.6 times the mean, near the morning peak. In case 1, with every pump off from a 2.5 m tank, the level falls to 1.20 m by 05:00 and the critical node goes under the floor at 05:35. The morning peak draws up to 108.6 L/s of demand and leakage, against 68 to 69 L/s from one pump, so pump 1 started at 05:00 lifts the hill from 25.97 to 26.73 m in a minute and still loses the floor for 120 min, lowest 24.55 m at 07:59. Both pumps from 04:00 hold it, lowest 26.11 m at 07:03. The tank is an integrator, and the pumps must lead the peak by hours, as the gates of Chapter 5 release water one travel time ahead of each published order.

> **TRY IT IN THE APP** **1 · Manual Pumps: Fill the Tank, Read the Hill**
>
> Load the scenario and press Run without touching a pump. Every pump stays off, so the tank alone feeds the town. Watch the tank level fall from 2.5 m to 1.20 m by 05:00 and the critical node go under the 25 m floor at 05:35.

### 11.4.2 A Tank That Also Sets the Leakage

The same level that protects the hill drives leakage everywhere upstream of a valve that is not holding its setting. Leakage rises by 0.17 L/s for every meter of tank level at mean demand in the declared model, so a tank held high all day pays for its margin in water. A pressure-reducing valve set for the peak gives too much at night. In case 3, with the schedule running, fixed valves leak 1252 $m^3$ of 5219 $m^3$ pumped, 24 %. A night cut of 2 m on every valve from 00:00 to 05:00 saves 13 $m^3$, while a 3 m cut breaks the floor for the whole window, lowest 24.53 m at 04:57, because a uniform cut is limited by the zone with the least margin. Modulated valves leak 1165 $m^3$ with the floor held, zone 1 falling from 377 to 331 $m^3$, zone 2 from 359 to 327 and zone 3 from 270 to 262, while the trunk's 245 $m^3$, upstream of every valve, does not move.

> **DID YOU KNOW?** **A Town That Lost Almost a Third of Its Water**
>
> A public water system in the southern United States serving 6,000 people found that about 30 percent of the water it produced went to apparent and real losses. With pressure monitors, a pressure regulating device and a lower operating pressure, its losses fell to 22 percent of the water produced [19].

## 11.5 The Control Layers as Programmed

The header MAN/AUTO runs the pump station. In Manual the learner switches pumps 1 and 2 and sets pump 3's speed. In Auto the station follows either the level band or the schedule, and the valves follow their own mode. Minimum run (2 h) and rest (1 h) are enforced in Auto, except that a tank within 0.15 m of overflow stops its pumps.

### 11.5.1 The Level Band

The level-band rule every utility starts with starts the lead pump when the measured level falls to $h_{low} = 2.0\,\text{m}$, the lag pump at $h_{low} - 0.5\,\text{m}$, and stops both at $h_{high} = 6.0\,\text{m}$, swapping lead and lag at every stop. It reads neither the tariff nor the hill's pressure.

### 11.5.2 The Hourly Schedule

The schedule is a mixed-integer linear program over the next $K = 24$ slots, the first running from now to the next whole hour. Its decisions in slot $k$ are the fixed pumps $u_{1k}, u_{2k} \in \{0,1\}$, pump 3's flow $q_{3k} \in [0, \bar{q}_3]$ and start indicators $s_{ik} \in [0,1]$, and it is stated in Equation (11.4).

$$
\begin{aligned}
\min\ & \textstyle\sum_k \pi_k \Delta_k (P_1 u_{1k} + P_2 u_{2k} + e_3 q_{3k}) \\
& \quad + c_V \textstyle\sum_{k=1}^{K-1} \Delta_k V_k + c_s \textstyle\sum_{i,k} s_{ik} \\
\text{s.t. } & V_k = V_0 + \textstyle\sum_{m<k} (Q_1 u_{1m} + Q_2 u_{2m} + q_{3m} - \hat{D}_m)\Delta_m \\
& A_T h_{fire} \le V_k \le A_T (h_{max} - 0.3), \quad V_K \ge V_{term} \\
& V_{k+1}/A_T + \beta_k (Q_1 u_{1k} + Q_2 u_{2k} + q_{3k}) \ge h_{req,k} \\
& s_{ik} \ge u_{ik} - u_{i,k-1}, \qquad \textstyle\sum_k s_{ik} \le N_s - n_i \\
& u_{ik} - u_{i,k-1} \le u_{i,k+\tau}, \quad \tau = 1, \dots, R_{run} - 1 \\
& u_{i,k-1} - u_{ik} + u_{i,k+\tau} \le 1, \quad \tau = 1, \dots, R_{rest} - 1
\end{aligned}
\tag{11.4}
$$

where $V_0$ is the measured level times $A_T$, $\Delta_k$ the slot length in hours, $\pi_k$ the slot's mean price, $N_s = 4$ the starts limit split at the next midnight with $n_i$ the starts already made today, and $R_{run} = 2$ and

$R_{rest} = 1$ slots, so the rest rows are vacuous at hourly slots. A tripped pump is fixed off, a pump inside its minimum run or rest is fixed for the leading slots, and after a switch from Manual the operator's pumps are fixed for the first slot so that the transfer has no bump. The program is linear because the pump flows and powers are the declared model's values at a nominal head, with the tank at mid-level and three pumps running: $Q_1 = 55.29\,\mathrm{L/s}$ at $P_1 = 61.12\,\mathrm{kW}$, $Q_2 = 40.26\,\mathrm{L/s}$ at 45.65 kW, and pump 3 up to $\bar{q}_3 = 55.29\,\mathrm{L/s}$ at $e_3 = 1.140\,\mathrm{kW}$ per L/s. The leakage term charges the stored volume at $c_V = p_w\,(\partial L/\partial h)/A_T = 3.0 \times 10^{-4}\,\$/(\mathrm{m^3\,h})$, with $\partial L/\partial h = 0.167\,\mathrm{L/s}$ per meter from the declared model. At midnight the program has 120 variables, 48 of them binary, and 144 rows.

Three rows carry the physics. The forecast $\hat{D}_k$ is the declared pattern's mean over the slot times the base demand, the hot-day multiplier when the weather forecast says so, an optional forecast error and a measured correction, plus the leakage the night flow last reported. The floor row asks the level at the end of each slot, plus a secant credit $\beta_k$ per unit of station flow, to reach $h_{req,k}$, the level at which the declared network, with valve 3 wide open and no pump running, holds the critical node at the floor plus 0.5 m at the slot's peak forecast demand. A table of $h_{req}$ and $\beta$ over demand multipliers from 0.2 to 2.6 is built once by bisection on the network. The terminal $V_{term}$ is the level the day's first plan had at the clock hour the horizon now ends, so solving again every hour cannot spend the tank a little lower each time. The first plan uses $V_K \geq V_0$. Section 1.4.8 explains why a plan needs a value or a constraint on what it leaves at the end of its horizon.

The program is solved by a depth-first branch and bound [14], the method of Section 1.4.10, written for this application on the framework's shared bounded-variable simplex and since moved to the framework's core, where the power-aware bed schedule of Volume 6, Chapter 10 uses it too. A node is pruned when its relaxation is infeasible or not below the incumbent by more than \$0.01 or $10^{-4}$ of it, it becomes the incumbent when every binary is within $10^{-6}$ of an integer, and otherwise it branches on the binary whose fraction is nearest one half into a child with the binary at 0 and one at 1, the side nearer the fraction explored first. The previous plan, shifted one slot, is offered as a first incumbent and accepted only if it satisfies every row. The search stops when the tree is exhausted, with the incumbent proved optimal, or at a node limit of 200, where it returns the incumbent with the gap to the best open bound. If the program has no feasible point it is solved again with a slack on every floor, fire-reserve and terminal row priced at \$2000 per meter of level per slot, and the plan is flagged as unable to hold the floor. The slacks are the soft constraints of Section 1.4.6, used here only after the hard program has failed. Only the first slot is carried out, and the program is solved again at the start of every hour from the measured level and at once when a trip relay changes, the receding horizon of Section 1.4.3. An option keeps the plan made before a trip until it runs out, as a teaching comparison and not an operating mode.

### 11.5.3 Pump 3 on the Station Flowmeter

Pump 3's speed follows a proportional-integral loop that holds the slot's planned station flow, $Q_{st,SP} = Q_1 u_{1k} + Q_2 u_{2k} + q_{3k}$, against the metered station flow. It is the velocity-form PID with back-calculation anti-windup of Section 1.6.2, at $K_c = 0.2\,\%$ per L/s and $\tau_I = 3\,\mathrm{min}$, executed every minute. Its lower limit is the larger of the drive's 60 % and the shutoff speed $\sqrt{(H_{S,m} - z_{cw})/a_3}$ at the measured station head, and a stopped pump is started at the speed that gives $q_{3k}$ alone, $\sqrt{(H_{S,m} - z_{cw} + b_3 q_{3k}^2)/a_3}$. When the plan asks the most pump 3 can give the pump runs at full speed with the loop tracking, and a planned flow under 4 L/s leaves it stopped.

### 11.5.4 Valves

In Fixed mode each valve holds an outlet pressure, 34, 34 and 40 m above valves at 110, 122 and 130 m. Time of day takes a night reduction, 2 m by default, off every valve from 00:00 to 05:00. Modulated mode sets each outlet head from its own zone meter so that the zone's highest customer

**Table 11.3** The Cases (Each Starts at Midnight)

| # | Case and settings | Teaching point |
|---|---|---|
| 1 | Manual, every pump off, tank 2.5 m, 12 h | the tank sets the hill |
| 2 | Level band, fixed valves | starts at the wrong hour |
| 3 | Schedule; valves fixed, time of day or modulated | leakage against the floor |
| 4 | Schedule, as case 2's day | relaxation and branching |
| 5 | Schedule once a day, forecast 20 % low, noise 10 % | receding horizon |
| 6 | Hot day, pump 1 trips at 07:00, plan kept | reconfiguration |
| 7 | Main break in zone 3 at 10:00, 30 h | the night flow |
| 8 | Solar tariff, 6 starts allowed, no wear cost | the starts limit |
| 9 | Challenge: hot day, peak, pump 2 trips at 18:00 | the whole problem |
| 10 | Hot day, pumps 1 and 2 lost, tank 3 m | an infeasible floor |

receives what it needs at the metered flow, as shown in Equation (11.5).

$$\begin{aligned} S_j &= z_{hi,j} + p_{floor} + 1 + r_{z,j} Q_{j,m}^{1.852}, \qquad j = 1,2 \\ S_3 &= z_{hi,3} + p_{crit,SP} + r_{z,3} Q_{3,m}^{1.852} + \Delta S_3 \end{aligned} \tag{11.5}$$

where $p_{crit,SP} = 26\,\text{m}$ and the trim $\Delta S_3$, between $-6$ and $+12\,\text{m}$, is a PI on the logger at $K_c = 0.6$ and $\tau_I = 1\,\text{h}$, executed when a report arrives, its integral held while valve 3 lies wide open, since an open valve cannot add head. No separate lift of the setpoints is applied when the floor cannot be met, because in the infeasible case valve 3 already lies open and in the feasible cases the trim already raises it.

### 11.5.5 Two Estimators

Every hour the town's use, the station meter less the tank's gain on its transmitter, is divided by the forecast of the same hour, and the ratio corrects the forecast through an exponential filter of weight 0.5 held between 0.7 and 1.5. The night flow is the mean of the three zone meters over the quietest hour, 03:00 to 04:00, less the customers' use in that hour by the declared pattern, reported at 04:00 as the leakage $\hat{L}$. It sees only leakage downstream of the zone meters, so the trunk main's loss, about 3 L/s at night, is invisible to it. The leak detector of Volume 10, Chapter 12 is a mass balance of the same kind, metered flows against the change in inventory, and like this one it is only as good as its meters.

🕮 **Pause.** Try Exercise 11.2 before reading on.

### 11.5.6 Three Models

The plant is the network of Section 11.3 with its leakage multiplier, its break and its noise. The controller's model is the declared network with neither, reduced to nominal pump flows, a floor table and a leakage slope, and corrected only through the measured forecast ratio and the night flow. The identified quantities are those two estimates. A parameter a learner edits during a run rebuilds both the plant's network and the declared model, so the two never disagree about the network the learner drew.

## 11.6 Ten Cases, Measured

The cases appear in Table 11.3. Each number that follows comes from a headless run of the engine of Section 1.6.1 on the case's own seed, stepped exactly as in the browser, after the last change to the application's behavior. Times are the controller's, so a pump starts at the control step that switches it. A case yields an identical run when loaded first, when reached from the menu after another case, and when restarted with Reset.

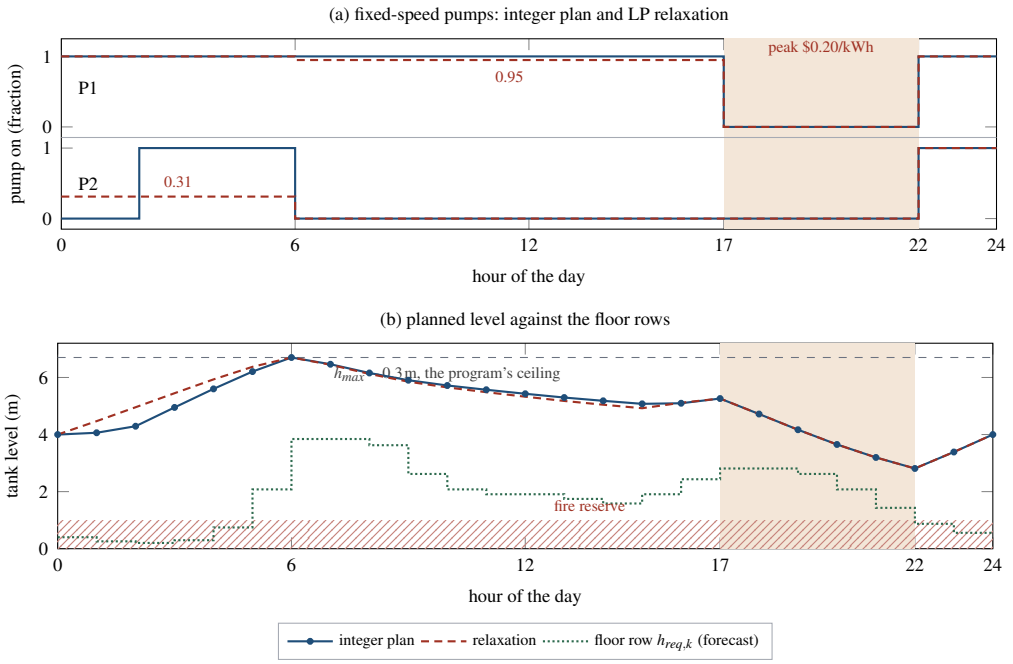



**Figure 11.2** Case 4's first solve at midnight. (a) The fixed-speed pumps of the integer plan (solid) and of the LP relaxation (dashed), slot by slot, with the relaxation's fractions 0.95 and 0.31. (b) The planned tank level of both (integer plan solid with dots, relaxation dashed), which differ by at most 0.67 m, against each slot's floor row before its credit for station flow (dotted). Shaded, the evening peak. Hatched, the fire reserve.

### 11.6.1 The Band Against the Schedule

The band in case 2 waits for the tank to reach 2.0 m, which it does in the morning peak, so the lead pump starts at 06:33, after the hill has dipped under the floor at 06:20 for 12.8 min to 24.43 m. The lag pump follows at 09:03 and both run to midnight, through the evening peak, for $241.9 of energy (1881 kWh). Case 4 runs the same day on the schedule. The first plan fills the tank on night power to 6.70 m by 06:00, coasts through the day on pump 1 with pump 3 topping it up from 15:00, runs nothing in the evening peak and refills from 22:00, as shown in Figure 11.2. The day costs $128.1 (1577 kWh), none of it at the peak price.

**Pause.** Try Exercise 11.1 before reading on.

The relaxation of that first program sets pump 1 at 0.95 from 06:00 to 17:00 and pump 2 at 0.31 through the night, 17 fractional slots in all, at $194.91. Branch and bound needs 29 nodes, 15 of them branched, to a depth of 7, and the incumbent improves three times, to $195.97, 195.95 and finally $195.90, proved optimal $1.00 (0.51 %) above the relaxation. The two are close because pump 3's continuous flow fills most of the gaps the binaries leave, and the search still has to branch to turn the fractions into runs that respect the two-hour minimum. Sixteen of the day's 24 solves branch.

### 11.6.2 Pressure, Valves and the Floor

Case 3's comparison of valve modes is given in Section 11.4. Lowering the valves of zones 1 and 2 by hand saves leakage only until those zones' own highest customers lose the floor, which happens below 32.6 m at valve 1 and 31.9 m at valve 2, so the verdict holds every zone's highest customer to the floor and not only the critical node.

### 11.6.3 Receding Horizon, a Trip and a Break

Case 5 forecasts demand 20 % low with 10 % noise. Solved once at midnight and followed blind, the plan ends the day with the tank 0.67 m below a plan that no longer describes the day, for \$519.8. Solved again every slot, each plan starts from the measured level, the forecast correction reaches $\times 1.32$, the tank stays within 0.11 m of the plan in force and the day costs \$499.7.

Case 6 trips pump 1 at 07:00 on a hot day. Kept, the plan made before the trip counts on pump 1 to fill the tank through the day, so the tank never fills, the critical node falls under the floor at 17:00 for 6 h, lowest 23.23 m, and the tank empties at 22:24. Re-solved at 07:01 with pump 1 fixed off, in 11 nodes with 5 branched, the program moves the refill to pumps 2 and 3, pump 2 running from 07:01 to 17:00, and the floor holds at a lowest 25.70 m for \$559.0, against \$545.0 without the trip.

Case 7 breaks a main in zone 3 at 10:00. The break discharges 557 $m^3$ by 30 h. The forecast correction, $\times 1.05$ when the main breaks, reaches $\times 1.17$ by 16:00, so the schedule pumps for the lost water before anyone knows it is lost, and the floor holds at a lowest 25.23 m. The night flow tells how much of it is loss. The zone meters' 03:00–04:00 mean, less the pattern's 27.5 L/s of night use, reports 11.2 L/s the first night (38.7 metered) and 19.2 L/s the second (46.8 metered). The truth downstream of the meters is 11.8 L/s and then 11.8 plus the break's 7.8 L/s, while the trunk main's 3.0 L/s sits upstream of every meter and is never seen.

> **TRY IT IN THE APP** **7 · A Main Breaks: The Night Flow Finds It**
>
> Load the scenario and run its 30 h. A main breaks in zone 3 at 10:00, and the forecast correction climbs from $\times 1.05$ to $\times 1.17$ by 16:00. Read the night flow reported at 04:00 each morning: 11.2 L/s after the first night and 19.2 L/s after the second. The trunk main's 3.0 L/s never appears.

### 11.6.4 Starts, and an Infeasible Day

On the solar tariff of case 8, six starts a day give each fixed pump three starts for a day of \$462.9, and one start each costs \$469.0. The first program's own cost rises from \$138.39 in 9 nodes to \$148.11 in 7, with a limit of two needing 158 nodes for \$142.52. Only the program's own cost is ordered by the limit. The realized day is not (\$462.4 at three starts, \$464.1 at two), because the day's demand is not the forecast the plan was made on.

Case 10 loses pumps 1 and 2 at midnight on a hot day. Pump 3 alone gives 66 to 68 L/s, and the town's draw of demand and leakage exceeds 67 L/s from 05:16 to 21:31 without a break. Every one of the 24 solves after the trips is the relaxed program, flagged as unable to hold the floor, and the critical node falls under it at 17:12, for 96 min in all, lowest 24.56 m, with the tank under its fire reserve at 19:48. A logged sweep of twelve runs, valves modulated, by time of day and fixed at 30, 40, 50 and 60 m, each scheduled and by hand, breaks the floor in all twelve, for 96 to 117 min, or all day with valves fixed at 30 m, where the valve itself is under the floor. A valve can take head away but cannot add it.

**Pause.** Try Exercise 11.3 before reading on.

### 11.6.5 The Challenge

The challenge runs a hot day with the evening peak and trips pump 2 at 18:00. It passes at 24 h if the critical node and each zone's highest customer never fall below 25 m, the tank never falls under its fire reserve or spills, neither fixed-speed pump starts more than four times and the energy cost plus the leaked water is at most \$530, with every parameter as shipped. Changing a parameter, the floor, the tariff peak, the starting level or the starts limit marks the challenge as not graded. The five policies of Table 11.4 share the plant, the events and the instruments. The shipped level band with fixed valves fails on every seed. It breaks the floor in the morning peak, for 55 min from 06:05 on the case's seed, and buys 384 kWh at the peak price. The suggested schedule with modulated valves passes on all eleven seeds and buys under 1 kWh at the peak price. The trip at 18:00 is absorbed by a

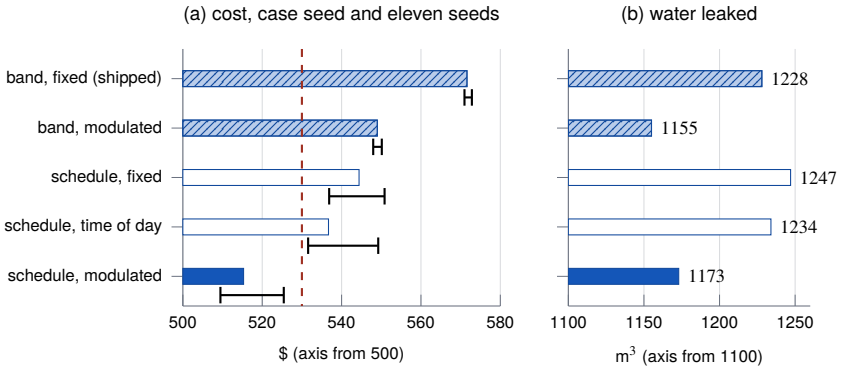



**Figure 11.3** The challenge of Table 11.4: (a) the graded cost on the case's seed (bars) with its range over eleven seeds (the bracket under each bar), and (b) the water leaked on the case's seed, for the level band (hatched bars), the schedule with fixed or time-of-day valves (open bars) and the schedule with modulated valves (dark bar). The red dashed line is the $530 target. Notice that only the dark bar and its whole bracket lie below the target, and that the modulated valves give the least leakage under either pump policy.

**Table 11.4** The Challenge at 24 h: Case Seed and Eleven Seeds

| Policy | Energy ($) | Leaked ($m^3$) | Cost ($) | 11 seeds ($) | Pass |
|---|---|---|---|---|---|
| band, fixed (shipped) | 203.2 | 1228 | 571.6 | 570.7–573.1 | 0/11 |
| band, modulated | 202.4 | 1155 | 549.0 | 547.7–550.4 | 0/11 |
| schedule, fixed | 170.4 | 1247 | 544.4 | 536.6–551.1 | 0/11 |
| schedule, time of day | 166.6 | 1234 | 536.7 | 531.3–549.5 | 0/11 |
| schedule, modulated (sugg.) | 163.5 | 1173 | 515.3 | 509.2–525.7 | 11/11 |

The two band policies break the floor on every seed (lowest 24.08 and 23.98 m). The scheduled ones hold it on every seed.

re-solve at 18:01 whose relaxation is already integral, one node. The two mixed scheduled policies hold the floor on every seed and fail only on cost, so both levers are needed. Figure 11.3 draws the five policies against the target.

The $530 target sits in a narrow gap. Over the case's seed and seeds 1 to 60 the suggested settings cost $509.0 to 527.4 and pass 60 of 61, and the schedule with time-of-day valves costs $531.3 to 549.5 and passes none. On seed 27 the suggested settings put the critical node 8 mm under the floor for one 10 s step at 17:52. The plan had kept the tank 0.54 m above that hour's floor row, but zone 3's own demand ran 18 % over its pattern, a 3.7-standard-deviation swing of the 5 % noise, while valve 3 lay open and no pump ran, and the logger's 15 min samples never saw the dip. That is the residual of an hourly plan made on a town-wide forecast against fifteen-minute noise concentrated in one zone, not a fault of the floor rows, and the scorecard counts it as a hard limit.

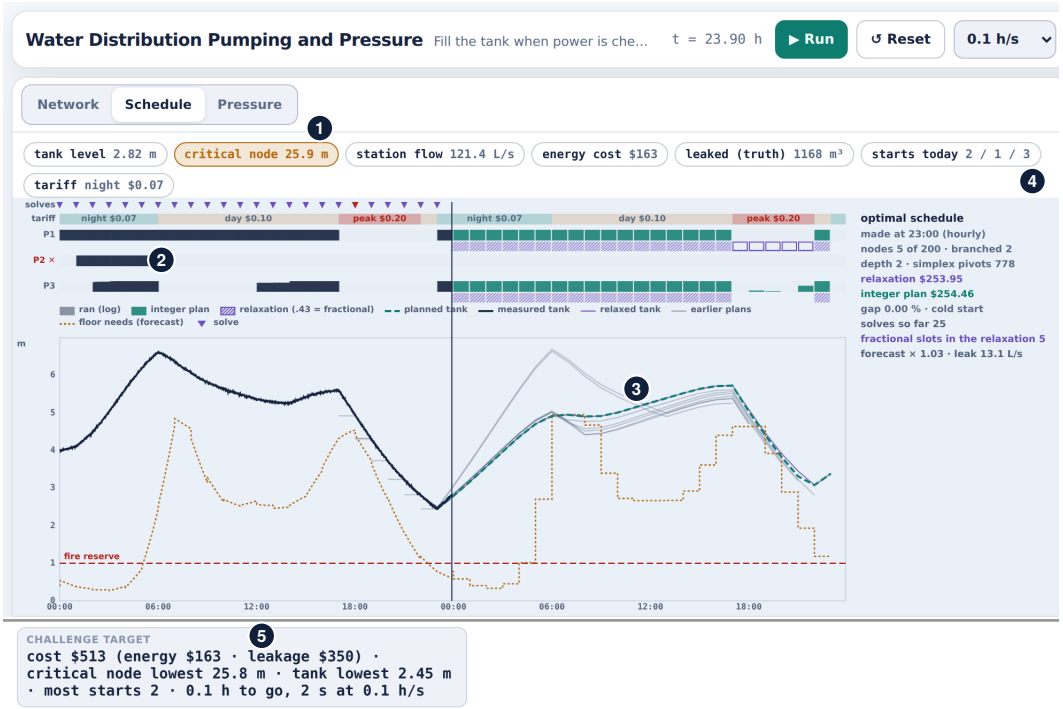



**Figure 11.4** The challenge with the suggested settings (schedule, modulated valves) at $t = 23.9$ h. The figure joins two parts of the screen, the header with the Schedule view (top) and the challenge tile of the Learn panel (bottom), and leaves out the control panel, the trends and the rest of the Learn panel. (1) The badges read the tank at 2.82 m, the logger at 25.9 m (the run's record puts the true critical node at 26.03 m), the station at 121.4 L/s, $163 of energy and 1168 $m^3$ leaked, and the night price of $0.07/kWh. (2) The pump rows show what ran against the tariff strip, nothing in the evening peak, pump 2 crossed after its 18:00 trip and the mark of the trip's re-solve at 18:01 among the hourly solve marks. (3) Ahead of the clock the integer plan made at 23:00 stands over the relaxation's sub-rows, with the planned tank dashed against the measured, the floor rows and the fire reserve. (4) The side panel reports that plan optimal in 5 nodes with 2 branched, a relaxation of $253.95 against $254.46, 25 solves so far, a forecast correction of $\times 1.03$ and a night-flow leakage of 13.1 L/s. (5) The challenge tile reads $513 ($163 of energy and $350 of leakage) with a lowest critical node of 25.8 m and a lowest tank of 2.45 m.

> ✦ **DID YOU KNOW? Valves That Lower the Energy Bill**
>
> From the simulator: in the challenge, the schedule with fixed valves spends $170.4 on energy and leaks 1247 $m^3$, while the same schedule with modulated valves spends $163.5 and leaks 1173 $m^3$. The valves move no pump, yet the energy falls too, and only the modulated policy passes on all eleven seeds (Table 11.4).

The application at that instant of the case's seed is shown in Figure 11.4.

### 11.6.6 Code Verification

Code verification and agreement with selected literature are kept apart here, and neither is a validation against a real network. Over case 7's 30 h the tank's gain of 1300.423 $m^3$ equals the integrated junction balance of the solves to $1.6 \times 10^{-4}$ $m^3$, and the water ledger closes to $2.3 \times 10^{-4}$ $m^3$ of 9059.9 $m^3$ pumped. On every one of 250 solves over eleven runs, 166 of them branched, the relaxation is no higher than the plan, and every plan and every realized day keeps the minimum run, the minimum

rest and the starts limit. The branch and bound matches enumeration on two instances written for the check: a unit commitment with 12 binaries, optimal at $348.10 in 21 nodes against all 4096 assignments solved through a different simplex, and the application's own program over six slots with 9 free binaries, $31.581448 in 17 nodes against 512 assignments. Stopped at six nodes, the first instance returns its incumbent with a bound of $338.70 and a gap of 2.70 % that brackets the optimum. On the shipped application every one of the 286 solves of the suggested challenge over eleven seeds and the 209 solves of the other scheduled cases reached an incumbent and proved it optimal, at most 29 and 106 nodes against the limit of 200, so the two-hour slots held in reserve were not needed.

Halving the step to 5 s with every noise off moves the challenge's day energy by at most 0.009 % for all five policies, with identical pump starts. With the instruments' noise on, a new step is also a new draw of that noise, the hourly plan can move an hour of pumping, and the day's energy moves by up to 2.3 % for the scheduled policies with fixed or time-of-day valves, and by up to 3.2 % as shipped with the 5 % demand noise also redrawn, while no verdict and no ranking changes in either comparison. In agreement with the literature, the affinity laws hold to round-off on the pump curve, the Hazen–Williams resistance matches EPANET's to a ratio of 1.00030, and the leakage law reproduces the FAVAD figures quoted in the previous section. In a desktop browser at the top speed of 0.5 h/s the challenge's day takes under 50 s of wall time (48.0 s in six runs), every one of 531 solves made during play takes under 250 ms (slowest 196.6 ms, median 30.0 ms), and the first solve, made while a case loads and before the clock moves, takes about half a second in case 5 (521 ms at most over three loads), whose first program needs 106 nodes. These are wall-clock bounds on one machine, not properties of the method.

## 11.7 Technology in Practice

### What Measures It

Design standards ask for controls that hold the levels of distribution storage, for level indication at a central location, and for overflow and low-level alarms placed where they are watched around the clock [18]. A common level measurement on an open tank is a differential pressure transmitter connected at the bottom with its low side vented to the air, so that the difference it reads is the hydrostatic head of the water. The lowest level it can measure is set by the point where it joins the tank, and the density of the fluid, the ambient temperature and the humidity affect its accuracy [20]. In this chapter the tank transmitter reads $h$ with 1 cm of colored noise, the station flowmeter $Q_{st}$ with 0.5 % and each zone meter with 1 % (Section 11.3).

> **TECHNOLOGY** **Sensor: Pressure Monitors at the Highest Customers**
>
> A federal fact sheet on pressure management advises placing pressure monitors at the highest and lowest elevations of each pressure zone and wherever pressure tends to run low. The logging interval can be set to seconds, keeping the minimum, maximum and average of each interval, and low-pressure alarms let operators react quickly. A drop in pressure may last a few seconds or many hours, and a surge swings between very high and very low pressure [19]. Many transducers let the pressure strain a fine wire grid whose change in resistance gives the signal, and they lose accuracy when the air pressure around them varies [20]. No open, vendor-neutral source gave an accuracy. The chapter's logger reports every 15 min, and on seed 27 it missed a dip of one 10 s step (Section 11.6).

### What Moves It

Fixed-speed pumps make up the majority of the pumps utilities have installed, while variable-speed pumps are more popular in new installations [4]. Design standards ask each booster pumping station for at least two pumps, sized so that peak demand can be met with the largest one out of service, and for a flow rate indicator and totalizer. Where static pressure exceeds 100 psi, pressure reducing devices are required on mains or on individual service lines [18]. The chapter's station has two fixed

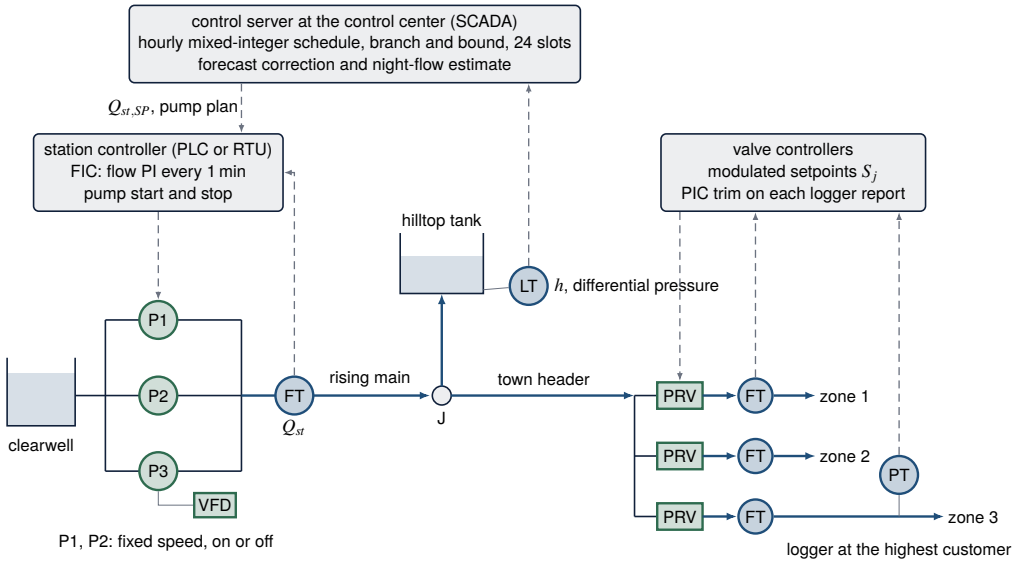



**Figure 11.5** The chapter's control layers drawn with real devices. The control server solves the hourly schedule and sends the station a flow setpoint and a pump plan, and the station controller starts and stops the fixed-speed pumps P1 and P2 and holds the metered flow (FT, $Q_{st}$) with pump 3's variable-frequency drive. The tank level (LT) returns to the server, the three pressure-reducing valves follow their zone meters (FT), and the logger (PT) at the highest customer trims valve 3. Notice that the only measurement of the hill's pressure arrives every 15 min.

pumps and one on a drive, and its three zones each take water through a pressure-reducing valve. Case 10 removes both fixed pumps on a hot day, and pump 3 alone cannot hold the floor.

> **TECHNOLOGY Final Element: Fixed-Speed Pumps in Parallel**
>
> A station can meet a changing demand with several smaller pumps in parallel instead of one large one. Each pump started shifts the operating point to the right along the system curve, and where static head is high, switching pumps on and off keeps each running pump nearer its best efficiency point. Pumps of different sizes need care, because the largest can push the others below their minimum flow. Starts have a price: repeated starting and stopping wears motor controllers, seals and bearings, and large motors draw high starting currents [21]. Soft starts or variable-frequency drives limit the pressure swings a start causes [19]. The chapter's two fixed pumps differ in size, and the schedule prices each start and allows four a day (Table 11.2, Equation (11.4)).

### Where the Controller Runs

Supervisory control and data acquisition systems are used in distribution systems such as water distribution. A control server at a control center stores and processes the data, field sites hold remote terminal units or programmable logic controllers that control the local process, and the server polls the field devices at defined intervals and sends new set points as required [22]. In this chapter the hourly schedule plays the server's part and sends the station a flow setpoint and a pump plan. The flow loop on pump 3 runs every minute and the valve trim at each 15 min logger report, the local loops of Figure 11.5.

> **IN PRACTICE** **A Full Tank Is Not a Fresh Tank**
>
> Storage tanks have often been kept full to be ready for emergencies, and a federal issue paper on finished water storage warns that a full but underused tank lets its water age, which favors microbial growth and chemical change. Operators manage water age by turning the stored water over and letting the level move. A complete turnover in three to five days has been recommended as a starting point, with a goal set for each tank. One large utility let its levels fall a further ten feet in normal operation and cut the mean residence time by two to three days [23]. The schedule of this chapter draws its tank down each day, but its model carries no water quality (Section 11.8).

### What Is Modernizing It

Three developments bear on this chapter's method. The first is exact scheduling. Optimal pump scheduling has been reported to cut a network's energy cost by 10 to 20 percent, and branch and bound can return schedules with a bounded optimality gap where heuristic methods give no bound [4]. The second is the network model kept live. The public-domain solver EPANET can be used to optimize the operation of tanks and pumps, and one utility moved from a static model to a real-time one that imports data from its supervisory system [24]. The third is metering by area. District metered areas, parts of a network that can be isolated and metered separately, support pressure and supply optimization, especially with real-time control [19]. The night flow of Section 11.5 reads the zone meters in the same spirit.

## 11.8 Limitations

The town is five nodes with teaching sizes, curves and prices, so the numbers show a mechanism rather than predict any network. The hydraulics are quasi-static, demand ignores pressure, and transients, water quality, demand charges and pump wear beyond a price per start are absent. The leakage law has one exponent for every zone and a coefficient fitted to give leakage of about a fifth to a quarter of supply, and the floor, the fire reserve and the exponent are teaching numbers, and their real counterparts come from a service standard, a fire authority and a field test.

The schedule's program is linear in nominal pump flows and powers at mid-tank, its floor rows come from the declared network at each hour's forecast peak, and the hourly re-solve from the measured level absorbs what the linearization misses. It forecasts the town as a whole, so a swing of demand in one zone within an hour can still dip the floor, as seed 27 shows. The branch and bound re-solves every relaxation from the start because the shared simplex has no warm start, and the node limit bounds its time. The protection the scorecard applies is an educational verdict on the truth, not a compliance test.

> **OPEN CHALLENGE** **Loops Slow the Exact Search**
>
> Branch and bound gives a pump schedule with a bound on how far it can be from the best, which heuristic searches cannot. A study of such formulations found that loops in the network raise the effort sharply: over 24 time steps, one formulation took 1 s on a network without loops and 43 s on one with two. Smaller time steps improved the computed cost by 5 to 10 percent while the effort grew exponentially. Loops are common in real networks, and the study named two ways forward: solvers that exploit the structure of loops, and reductions to equivalent networks without them [4]. The town of this chapter is five nodes with no loop (Section 11.8).

## 11.9 Conclusion

The simulator of this chapter couples integer pump scheduling, pressure-reducing valves and pressure-dependent leakage on one quasi-static network whose balances close to round-off. On the same day the level band buys 571 kWh at the peak price and lets the hill dip under its floor, while the hourly schedule spends \$128.1 against \$241.9 and buys none. Branch and bound turns a relaxation 0.51 % cheaper than the plan into a plan proved optimal in 29 nodes, and every one of 495 solves on the

shipped cases finishes inside its node limit. Modulated valves save 87 m$^3$ of leakage a day where a uniform night cut of 3 m breaks the floor, a re-solve after a trip holds a floor that a kept plan loses for six hours, and the night flow reveals a break that the daytime forecast correction had already paid for. Two pumps lost on a hot day cannot be planned around, and the program says so from its first solve. The schedule forecasts the town rather than each zone and the network is lumped. A fuller model would add pressure-dependent demand, a looped network with an EPANET-class solver, demand charges and a zone-level forecast. The simulator shows that the cheapest safe day for a water network is decided by when the pumps run, not by how full the tank is kept.

**KEY POINTS**

- The cheapest safe day for a water network is decided by when the pumps run, not by how full the tank is kept.
- On the same day the level band buys 571 kWh at the peak price and lets the hill dip under its floor, while the hourly schedule spends $128.1 against $241.9 and buys none.
- Branch and bound turns a relaxation 0.51 % cheaper than the plan into a plan proved optimal in 29 nodes.
- Modulated valves save 87 m$^3$ of leakage a day, and a re-solve after a trip holds a floor that a kept plan loses for six hours.

**Where else this applies.** The refrigeration pack of Volume 8, Chapter 12 stages two fixed-speed compressors and one on a drive, the same arrangement as the pumps here.

## References

[1] P. W. Jowitt and G. Germanopoulos, “Optimal pump scheduling in water-supply networks,” *Journal of Water Resources Planning and Management*, vol. 118, no. 4, pp. 406–422, 1992.

[2] M. J. H. Sterling and B. Coulbeck, “Technical note. A dynamic programming solution to optimization of pumping costs,” *Proceedings of the Institution of Civil Engineers*, vol. 59, no. 4, pp. 813–818, 1975.

[3] U. Zessler and U. Shamir, “Optimal operation of water distribution systems,” *Journal of Water Resources Planning and Management*, vol. 115, no. 6, pp. 735–752, 1989.

[4] R. Menke, E. Abraham, P. Parpas, and I. Stoianov, “Exploring optimal pump scheduling in water distribution networks with branch and bound methods,” *Water Resources Management*, vol. 30, no. 14, pp. 5333–5349, 2016.

[5] L. E. Ormsbee and K. E. Lansey, “Optimal control of water supply pumping systems,” *Journal of Water Resources Planning and Management*, vol. 120, no. 2, pp. 237–252, 1994.

[6] H. Mala-Jetmarova, N. Sultanova, and D. Savic, “Lost in optimisation of water distribution systems? A literature review of system operation,” *Environmental Modelling & Software*, vol. 93, pp. 209–254, 2017.

[7] G. Germanopoulos, “A technical note on the inclusion of pressure dependent demand and leakage terms in water supply network models,” *Civil Engineering Systems*, vol. 2, no. 3, pp. 171–179, 1985.

[8] J. E. van Zyl and A. M. Cassa, “Modeling elastically deforming leaks in water distribution pipes,” *Journal of Hydraulic Engineering*, vol. 140, no. 2, pp. 182–189, 2014.

[9] G. Germanopoulos and P. W. Jowitt, “Leakage reduction by excess pressure minimization in a water supply network,” *Proceedings of the Institution of Civil Engineers*, vol. 87, no. 2, pp. 195–214, 1989.

[10] K. Vairavamoorthy and J. Lumbers, “Leakage reduction in water distribution systems: Optimal valve control,” *Journal of Hydraulic Engineering*, vol. 124, no. 11, pp. 1146–1154, 1998.

[11] B. Ulanicki, P. L. M. Bounds, J. P. Rance, and L. Reynolds, “Open and closed loop pressure control for leakage reduction,” *Urban Water*, vol. 2, no. 2, pp. 105–114, 2000.

[12] C. Ocampo-Martínez, V. Puig, G. Cembrano, and J. Quevedo, “Application of predictive control strategies to the management of complex networks in the urban water cycle,” *IEEE Control Systems*, vol. 33, no. 1, pp. 15–41, 2013.

[13] E. Creaco, A. Campisano, N. Fontana, G. Marini, P. R. Page, and T. Walski, “Real time control of water distribution networks: A state-of-the-art review,” *Water Research*, vol. 161, pp. 517–530, 2019.

[14] A. H. Land and A. G. Doig, “An automatic method of solving discrete programming problems,” *Econometrica*, vol. 28, no. 3, p. 497, 1960.

[15] L. A. Rossman, "EPANET 2 users manual," U.S. Environmental Protection Agency, Washington, DC, Tech. Rep. EPA/600/R-00/057, 2000. [Online]. Available: https://cfpub.epa.gov/si/si_public_record_report.cfm?Lab=NRMRL&dirEntryId=95662

[16] A. Lambert, "Leakage reductions: The fundamental role of pressure management," WaterWorld (Water & Wastewater International), Mar. 1, 2013, 2013. [Online]. Available: https://www.waterworld.com/drinking-water-treatment/potable-water-quality/article/16201852/leakage-reductions-the-fundamental-role-of-pressure-management

[17] Water Supply Committee of the Great Lakes–Upper Mississippi River Board of State and Provincial Public Health and Environmental Managers, *Recommended Standards for Water Works*, 2022nd ed. Eagan, MN: Minnesota Government Publications, 2022.

[18] ——, *Recommended Standards for Water Works*, 2012th ed. Albany, NY: Health Research, Inc., 2012. [Online]. Available: https://www.health.mn.gov/communities/environment/water/docs/tenstates/waterrev2012.pdf

[19] U.S. Environmental Protection Agency, "Protecting water quality through distribution system pressure management," U.S. Environmental Protection Agency, Office of Water, Washington, DC, Distribution System Toolbox fact sheet EPA 815-F-21-005, 2021. [Online]. Available: https://www.epa.gov/system/files/documents/2021-12/ds-toolbox-fact-sheets_pressure-management.pdf

[20] *DOE Fundamentals Handbook: Instrumentation and Control, Volume 1*, U.S. Department of Energy, Washington, DC, 1992, dOE-HDBK-1013/1-92.

[21] U.S. Department of Energy, "Improving pumping system performance: A sourcebook for industry, second edition," U.S. Department of Energy, Office of Energy Efficiency and Renewable Energy, Industrial Technologies Program, Washington, DC, Tech. Rep. DOE/GO-102006-2079, 2006.

[22] K. Stouffer, M. Pease, C. Tang, T. Zimmerman, V. Pillitteri, S. Lightman, A. Hahn, S. Saravia, A. Sherule, and M. Thompson, "Guide to operational technology (OT) security," National Institute of Standards and Technology, Gaithersburg, MD, NIST Special Publication 800-82r3, 2023.

[23] American Water Works Association and Economic and Engineering Services, Inc., "Finished water storage facilities," U.S. Environmental Protection Agency, Office of Ground Water and Drinking Water, Washington, DC, Distribution System Issue Paper, 2002. [Online]. Available: https://www.epa.gov/sites/default/files/2015-09/documents/2007_05_18_disinfection_tcr_whitepaper_tcr_storage.pdf

[24] U.S. Environmental Protection Agency, "Use of modeling to protect water quality," U.S. Environmental Protection Agency, Office of Water, Washington, DC, Distribution System Toolbox fact sheet EPA 815-F-22-011, 2023. [Online]. Available: https://www.epa.gov/system/files/documents/2023-08/DS%20Toolbox%20Fact%20Sheets_Modeling_508ed.pdf

## 11.10 Exercises

**11.1** Concept check · 2 min

In scenario 2 the level band starts the lead pump at 06:33, thirteen minutes after the critical node first falls under the 25 m floor. Why does the band let the hill dip before it acts?

(a) The fixed-speed pumps take several minutes to run up to speed, so the hill keeps falling after the start command

(b) The band acts on the tank level alone, and in the morning peak the hill needs more level than the 2.0 m start point

(c) The one-hour minimum rest holds the lead pump off after its last stop, which delays the start past the low band

(d) The band waits for the next fifteen-minute logger report before it sees that the critical node has lost its floor

**11.2** Concept check · 2 min

In scenario 7 the night flow reports 11.2 L/s of leakage on the first night, while the trunk main leaks another 3.0 L/s that the estimate never includes. Why does the estimate miss the trunk main?

(a) The trunk main leaks only by day, when the junction head is highest, so its loss is zero in the quiet hour

(b) The estimator subtracts the trunk main's leakage, computed on the declared model, before it reports the night flow

(c) The estimate is built from the three zone meters at the valves, and the trunk main leaks upstream of all of them

(d) The declared pattern counts the trunk main's loss as part of the 27.5 L/s of the customers' night use

**11.3** Concept check · 2 min

In scenario 10 pumps 1 and 2 are out on a hot day. Twelve runs with the valves modulated, on time of day, and fixed at 30, 40, 50 and 60 m, each scheduled and by hand, all break the floor. Why can no valve setting hold the critical node?

(a) The trim on valve 3 winds up while the logger reads low and holds the valve partly shut through the evening

(b) The relaxed program prices the floor slack at $2000 per meter, so the schedule stops pump 3 to save energy

(c) The fire-reserve row forces the schedule to keep water in the tank instead of sending it up to the hill

(d) A pressure-reducing valve can only take head away, and pump 3 alone cannot keep the tank high enough for the hill

**11.4** Theory and calculation · 6 min

Pump 1 has a shutoff head $a_1 = 110\,\mathrm{m}$, a curve coefficient $b_1 = 8264\,\mathrm{s^2/m^5}$, a best-efficiency flow of 55 L/s and a peak efficiency of 0.80, and its motor is 94 % efficient (Table 11.2). Use Equations (11.1) and (11.3) with $\rho = 1000\,\mathrm{kg/m^3}$ and $g = 9.81\,\mathrm{m/s^2}$.

(a) Pump 1 alone fills the tank at 3 m with 65.52 L/s. What head does it add? How much of it is lost to friction between the station and the tank? *Checkpoint:* 74.52 m; 3.52 m above a 71 m static lift

(b) The schedule's program uses pump 1 at its nominal 55.29 L/s. Compute the head, the efficiency and the electrical power there, and the energy per cubic meter lifted. *Checkpoint:* 84.74 m, 0.800, 61.12 kW, 0.307 kWh/m$^3$

(c) Pump 3 has the curve of pump 1 scaled by its speed ratio. Find the least speed at which it can lift water into the empty tank and into the full tank at 7 m, and say whether the drive's 60 % minimum matters. *Checkpoint:* 78.6 % and 82.6 %

**11.5** Theory and calculation · 6 min

The leakage law of Equation (11.2) is $C_j \bar{p}_j^{N_1}$ with $N_1 = 1.18$.

(a) Compute the fraction by which leakage falls when the pressure is halved at $N_1 = 0.5$ and at $N_1 = 1.5$, and when it falls by a tenth at $N_1 = 1.18$. *Checkpoint:* 29.3 %, 64.6 %, 11.7 %

(b) In case 3 the three zones leak 377, 359 and 270 m$^3$ with fixed valves and 331, 327 and 262 m$^3$ with modulated ones. If the modulated valves had lowered every zone's mean pressure by the same fraction at every hour, what fraction would that be? Why does the trunk main's 245 m$^3$ not move? *Checkpoint:* about 7.3 %

(c) The schedule charges the stored volume at $c_V = p_w\,(\partial L/\partial h)/A_T$. Compute it from $p_w = 0.30\,\$/\mathrm{m^3}$, $\partial L/\partial h = 0.167\,\mathrm{L/s}$ per meter and $A_T = 600\,\mathrm{m^2}$. *Checkpoint:* $3.0 \times 10^{-4}$ \$/(m$^3$ h)

**11.6** Theory and calculation · 7 min

The schedule of Equation (11.4) has, in each of its 24 slots, the two pump states $u_{1k}$ and $u_{2k}$, pump 3's flow $q_{3k}$ and two start indicators $s_{ik}$.

(a) How many variables and how many binaries does the program have? How many on-off assignments of the binaries are there, and how many linear relaxations did case 4's first solve need to prove its plan optimal? *Checkpoint:* 120 variables, 48 binaries, $2^{48} \approx 2.8 \times 10^{14}$; 29

(b) The Schedule view's side panel reports that solve as a relaxation of \$194.91, an integer plan of \$195.90 and a gap of 0.00 %. How far is the plan above the relaxation, in percent? Why does the panel show a gap of zero? *Checkpoint:* 0.51 %

(c) With $R_{run} = 2$ and $R_{rest} = 1$, write the minimum-run and minimum-rest rows for pump 1 in slot $k$. If pump 1 is off in slot 4 and on in slot 5, what must it do in slot 6? *Checkpoint:* on in slot 6; no rest rows

**11.7** App challenge · 8 min

Load scenario 2, "The level band across the tariff", at the default seed. Set the speed to 0.25 h/s and press Run. The case pauses at 24 h, about 96 seconds later.

(a) Read the Energy cost · leakage tile and the Critical node · worst tile. What did the day's energy cost, how many kWh did it take, and how long was the critical node below the floor? *Checkpoint:* \$242, 1881 kWh; 13 min, lowest 24.4 m

(b) Load scenario 4, "The schedule: relaxation and branch and bound", and run it to its pause at 24 h in the same way. Read the same tiles. *Checkpoint:* \$128, 1577 kWh; never below, lowest 26.1 m

(c) Export the CSV of each run and read `tank_m` in its last row. Compute the average price each day paid per kWh. Is the difference in kWh between the two days a fair measure of what the schedule saves? *Checkpoint:* 4.55 m and 2.41 m; \$0.129 and \$0.081 per kWh

Give money to \$1, energy to 5 kWh and levels to 0.02 m.

**11.8** App challenge · 8 min

Load scenario 6, "A pump trips at the peak", at the default seed. It loads with On a trip at Keep the plan. Set the speed to 0.5 h/s and press Run.

(a) The run pauses when the tank empties. Read the clock and the Critical node · worst tile. When did the tank empty, and how low did the critical node fall? *Checkpoint:* 22:24; 23.2 m at 19:12

(b) Reload the scenario with the reload button, set On a trip to Re-solve on the Schedule card before pressing Run, and run to the pause at 24 h. Read the Critical node · worst and Energy cost · leakage tiles. In the exported CSV, read `sched_nodes` and `sched_branched` in the rows from 07:02 to 08:00. *Checkpoint:* 25.7 m, never below; \$559; 11 nodes, 5 branched

(c) The run of (a), continued by pressing Run once more after the pause, stops at 24 h with \$485 on the tile, less than the re-solved day. Why is that not a saving? *Checkpoint:* about 678 kWh not pumped

Give pressures to 0.1 m and money to \$1.

**11.9** App challenge · 9 min

Load the challenge, "a hot day, the peak and a trip" (case 9 of Table 11.3), at the default seed, and set the speed to 0.5 h/s. Each run pauses at 24 h, 48 seconds after Run, with the verdict on the Challenge target tile.

(a) Run it as shipped. Read the verdict, the cost and its split, and the time the critical node spent below the floor. *Checkpoint:* not met; \$572 (energy \$203, leakage \$368); 55 min

(b) Set Pump control to Schedule and Valves to Modulated, press Reset, which keeps them, and run again. Read the tile. *Checkpoint:* met; \$515 (energy \$163, leakage \$352)

(c) Set Valves back to Fixed, keep the schedule, press Reset and run. Which part of the target fails now? What do (a) to (c) say about the two levers? *Checkpoint:* not met on cost, \$544; floor held

Give money to \$1.

**11.10** Application area · 4 min

The chapter's pressure floor is 25 m of head at the highest customer.

(a) Convert it to kilopascals and to pounds per square inch (1 psi $= 6.895$ kPa), with $\rho = 1000\,\mathrm{kg/m^3}$ and $g = 9.81\,\mathrm{m/s^2}$. How many meters is the 20 psi minimum of the design standards cited in Section 11.2? *Checkpoint:* 245 kPa, 35.6 psi; 14.1 m

(b) The floor, the 1 m fire reserve and the leakage exponent are teaching values. Where would a utility take each of them from?

**11.11** Application area · 5 min

An operator objects that filling the tank at night only to hold it high all day pays for its margin in leaked water. Use pump 1's nominal 55.29 L/s at 61.12 kW, the tank's $600\,\mathrm{m^2}$ and the program's leakage price $c_V = 3.0 \times 10^{-4}\,\$/(\mathrm{m^3\,h})$.

(a) How long does pump 1 take to lift one meter of tank, how much energy does that take, and what does buying it at the \$0.07 night price instead of the \$0.20 peak price save? *Checkpoint:* 3.01 h, 184 kWh, \$23.95

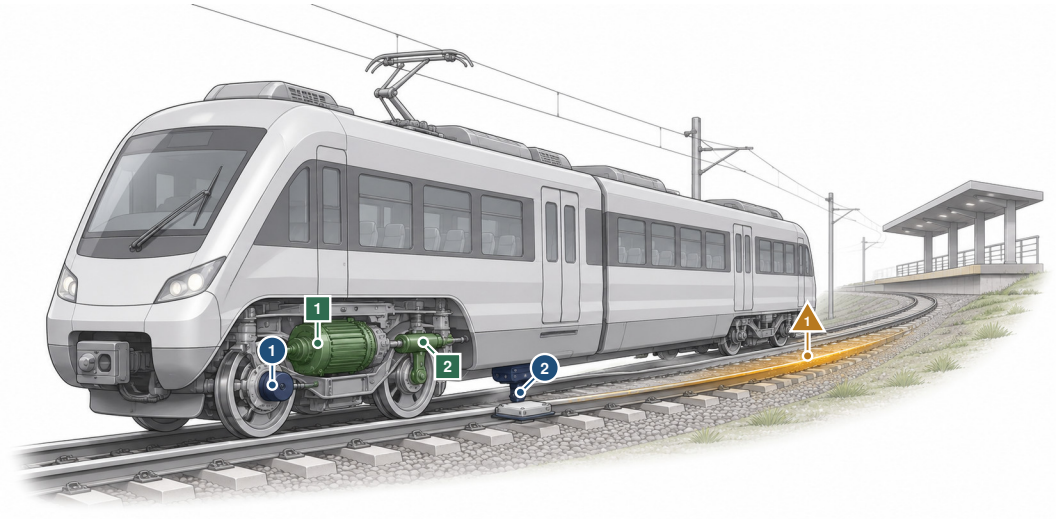
(b) What does holding that extra meter for a whole day cost in leaked water? Is the operator's objection right? *Checkpoint:* \$4.32

**11.12** Review the agent · 5 min

An AI assistant looks at case 4's first solve (Figure 11.2) and writes: "Branch and bound spent 29 relaxations to gain one dollar. Skip it and round the relaxation instead. Pump 1 at 0.95 from 06:00 to 17:00 rounds to on, pump 2 at 0.31 through the six night slots rounds to off, and pump 3 keeps its relaxed flows. The rounded plan keeps the two-hour minimum run, and since rounding moves each slot by less than half a pump, it is as feasible as the relaxation and costs within a few dollars of \$194.91." Find the flaw, and say how you would check the rounded plan before using it.

CHAPTER 12

# Automatic Train Operation Simulator Where the Timetable Decides Where to Coast




**The controller moves** 1 traction motor and drive 2 brake

**The controller reads** 1 wheel speed sensor 2 balise reader under the car

**What disturbs it** 1 wet rail ahead

*Notice where the force acts: the motors push the train, the brakes stop it at the platform, and on the wet rail ahead the wheels can slip. Illustration, not to scale.*

### IN PLAIN TERMS Where a Train Should Stop Pushing

**The main idea.** An electric train must reach the next platform at a set time, and it usually has a few seconds to spare. It can spend them by running slower. It can also switch off its motors early and roll, which saves electric power.

**The challenge.** Speed limits on curves and hills change where rolling pays, so a planner works out the whole run before the train leaves. Wet rails, a heavier load and a late start then push the train off its plan, and the loops that steer it must bring it back.

**Why it matters.** On the chapter's line, one well-placed switch from power to rolling cuts the energy of a run by 31 percent.

### THREE WAYS TO READ THIS CHAPTER

| | | |
|---|---|---|
| **Explore** | 12 min | the first page, the history in Section 12.2, the fact and try-it segments in the Introduction and Section 12.6, and the key points |
| **Study** | 45 min | the whole chapter with its three pause points and the twelve exercises at its end |
| **Practice** | 15 min | the first page, the results (Section 12.6), Technology in Practice (Section 12.7), Limitations (Section 12.8) and the open challenges |

**The control question** An electric train due at the next platform at a set time can spend its spare seconds on speed or on coasting. Where should the force change along the run so that the train arrives on time with the least energy, and what does a plan need from the loop that tracks it?

**Learning objectives**

- Derive from a Hamiltonian linear in the force why an energy-optimal run uses only full power, hold, coast and brake.
- Interpret the price of time $\lambda$ as the energy one more second of budget saves, and relate it to the holding speed.
- Compare one coast point with a dynamic program on a line where speed limits interleave, and explain the reversal of the classical rule.
- Explain how wheel slip, a heavier train and a late departure each defeat a plan, and which part of the stack recovers each.
- Recognize a timetable that no plan can meet from the planner's earliest arrival.

**Prerequisites** Sections 1.4.9, 1.4.7, 1.4.6, 1.6.2 and 1.6.4, and Volume 2, Chapter 2 for a vehicle speed loop in velocity form.

**The app** https://apmonitor.com/apps/train-ato/

**See also** Volume 20, Chapter 9 for a speed limited by the distance needed to stop, Volume 20, Chapter 2 for a vehicle whose loops depend on its speed, and Volume 7, Chapter 4 for slip regulated at the peak of a friction curve.

## 12.1 Introduction

A suburban train leaving station A has 5.0 km and 275 s to reach the platform mark at station B. In the simulator of this chapter the fastest drive that the train and the speed limits allow takes 250.8 s on its plan, and driven flat out the train arrives 21.0 s early with 28.2 kWh on the energy meter at the pantograph. Leaving that drive at one coast point 2.45 km from A, and coasting from there to the final braking curve, the train arrives 3.1 s early with 19.5 kWh, which is 31 % less. A dynamic program that places its own switches stops powering 0.35 km from A at 81 km/h and arrives 1.7 s early with 16.4 kWh. The three runs differ only in where the force changes, and the 24.2 s between the flat-out time and the timetable is what the second and the third spend.

The misconception this plant corrects is that a timetable's slack has one best use whatever the line, a belief that the textbook rule, coast rather than slow down, can leave behind. On a level 5 km test line in the same model, a plan barred from coasting needs 7.2 to 8.8 % more energy than the plan with one coast point, at budgets from 5 to 30 % over the flat-out time, which is the classical result. On the A to B run, where a 60 km/h curve begins 1.5 km out, the order reverses at the 271 s budget. The one-coast-point plan, which holds 98 km/h and brakes into the curve before it coasts, plans 18.76 kWh. The same dynamic program barred from coasting holds about 78 km/h and plans 16.32 kWh, and the unrestricted program, which combines a lower top speed with coasting, plans 15.45 kWh. Where the limits fall decides which rule wins, and that is the argument for planning the whole run.

The problem belongs to a general class, the least-energy transfer between two states of rest in a fixed time, under state constraints that change with position. Energy-efficient driving of road vehicles has been posed as the same optimal control problem [1], and the recursion that solves such a problem on a grid is the one Bellman published in 1957 [2]. A batch brought to a hold temperature by a deadline, or a tank filled by a pump before a tariff changes, has the same shape, with a slack in time that can be spent on rate or on letting the system run. What the train adds is a structure known in advance. Pontryagin's maximum principle limits the energy-optimal drive between stops to full power, speed holding, coasting and full braking [3, 4], so a numerical plan can be checked against the regimes it ought to contain.

**✦ DID YOU KNOW? Energy, a Large Line in a Transit Budget**

A review of how urban rail systems recover braking energy reports that payments to electric utilities for energy and peak power demand reach up to 35 percent of a rail transit operator's operating expense. Studies that time braking trains to feed accelerating ones claimed savings of 4 to 34.5 percent [5].

A course usually meets the two halves of this problem apart. The theory of optimal train control is written for analysis [4, 6, 7], dynamic programming of running profiles is reported as a numerical method [8, 9], and automatic train operation appears as a survey of research and practice [10]. Speed control elsewhere in the series is a regulator on one setpoint, as in the cruise control of Volume 2, Chapter 2, while here a dynamic program plans the whole trajectory. This chapter works with a browser train simulator in which an analytic four-regime plan and a dynamic program plan the same runs, a tracking loop drives those plans through a lagged drive and a wet rail, and the plan, the run, the clock and the meter are logged side by side. Its measurements are made on one declared line, among them the reversal above and the price of a lost minute. The dynamic program has eleven decisions on a 50 m by 1 km/h grid and solves a plan in 156 to 239 ms in a browser. It is checked against brute force, a closed form and two energy balances, behind a measurement boundary that a test enforces. Pontryagin's structure, Davis resistance, the creep law and dynamic programming are all standard, and Section 1.4.9 sets out the same optimal control and recursion on the heated tank of Chapter 1. Each of the ten scenarios carries one completion rule, and the exercises at the end of the chapter are built on them.

Section 12.2 traces the domain from the driver's judgment to planned profiles. Section 12.3 gives the train, the line, the rail and the instruments. Section 12.4 derives the structure of an energy-optimal run and shows how its switching points move with the budget. Section 12.5 describes the planners, the tracking and stopping loops, the slip override and re-planning. Section 12.6 measures the nine cases and the challenge and verifies the code, Section 12.7 describes the instruments, drives and protection of real trains, Section 12.8 states the limitations, and Section 12.9 draws the results together.

## 12.2 From a Driver's Judgment to Planned Profiles

For most of the history of electric traction, how a train was driven between stations was left to the driver, who had a timetable and the feel of the line. A 1968 study that posed the run as an optimal control problem is told in the box below, and a solution of the energetically optimal control of a train by the maximum principle followed in 1985 [11].

**⧖ HISTORY 1968: The Run Between Stations Becomes an Optimal Control Problem**

A paper received in 1967 and published in 1968 asked for the train operation that uses the least energy to take a train from one station to the next, arriving at a specified instant and stopping there. Because the speed limit must be observed strictly, it treated the run as an optimization problem with a bounded state variable. Pontryagin and others had given a jump condition of the costate for such problems, but the paper judged it not easily applied to the synthesis of a controller. Its author used a simple method of his own instead and obtained the optimal control law. He believed the work to be the first scientific research on the economization of train operation [12]. The speed limit of that paper is the state constraint $v \le v_{lim}(s) - \Delta v_m$ of Equation (12.6).

The structure that this simulator teaches was settled over the following two decades. A monograph on energy-efficient train control appeared in 1995 [6]. In 2000 the optimal control of a train was characterized as a sequence of a few control modes [3], and a numerical algorithm for a variable grade under arbitrary speed restrictions was built on the analytical properties of the maximum-principle solution [13], with further work on energy-efficient operation of rail vehicles in 2003 [14]. A systematic account of the key principles, from strategies of optimal type and the location of optimal

**Table 12.1** Roles of the Variables, and Their Values at $t = 470\,\text{s}$ When the Challenge Runs on Its Suggested Settings

| Variable | Role | At 470 s | Range |
|---|---|---|---|
| $\hat{v}$ | CV, measured speed | 67.4 km/h | 0 to $v_{lim}$ |
| $\Delta t$ | CV, projected arrival at C | +3.6 s | ±15 s |
| $u_{cmd}$ | MV, lever | 51 % | −100 to 100 |
| $F_{cap}$ | MV of the override, motor cap | 71.0 kN | 5 kN to $F_t$ |
| $\mu(s)$ | DV, adhesion (hidden) | 0.08 | 0.25, 0.08 |
| $\kappa$ | DV, line receptivity | 100 % | 0 to 100 % |
| — | DV, restriction, door hold, load | on, spent, 0 | events |
| $s$, $\hat{s}$ | state, position, odometer | 6.822, 6.822 km | 0 to 12 |
| $v$ | state, true speed | 67.3 km/h | $\geq 0$ |
| $w$, $\hat{w}$ | state, slip, measured | 0.41, 0.42 m/s | — |
| $F_d$, $F_m$ | state, demand, motor force | 74.7, 71.0 kN | — |
| $E$ | state, meter | 53.0 kWh | — |
| $E_t$, $E_r$ | states, drawn, returned | 69.7, 16.7 kWh | — |
| $\theta$ | grade at the train | +2.0 % | ±2.0 % |

switching points to existence, uniqueness, a local energy minimization principle and computational techniques, was published in two parts in 2016 [4, 7].

Dynamic programming entered the problem as the numerical alternative that accepts whatever the line contains. A discrete dynamic programming algorithm that minimized traction energy on a nonlinear point-mass model, subject to the travel time, the speed limits and the available power, was reported in 2000 with energy savings between 10 and 30 % in simulation and in measurements, and was proposed as the basis of a model predictive controller running in hard real time [8]. A 2005 study applied Bellman's dynamic programming to the running profile with complicated motor and brake characteristics, speed limits, nonlinear resistance and variable grades, found computation time to be its obstacle, and reduced it with a confined state space and an irregular lattice [9]. The 50 m by 1 km/h grid used here is a small case of the same approach, sized so that a plan is solved inside one control step of a browser. A 2017 review covers energy-efficient train control together with its integration into the timetable [15], the connection on which cases 8 and 9 of Section 12.6 turn.

Automatic train operation, the onboard function that drives a train from one platform to the next, has its own research and development literature, which a 2017 survey collects [10]. Two physical questions sit beside the optimization. Regenerated braking energy is worth something only if the line can take it, and strategies and technologies for managing it in urban rail have been reviewed [16]. A traction vehicle at the limit of adhesion needs a creep-force law that holds beyond its maximum [17], which is where a wet rail puts a train that its plan expected to find dry.

## 12.3 The Train, the Line and the Rail as Modeled

The simulated train is a generic four-car electric unit with two motored and two trailer cars, 190 t loaded, on a declared 12 km line, as shown in Figure 12.1. Stations A, B and C stand at 0, 5.0 and 12.0 km. The line speed is 100 km/h, with permanent limits of 60, 80, 80 and 70 km/h over 1.5 to 2.3, 3.1 to 3.8, 7.6 to 8.3 and 10.6 to 11.2 km, a climb of +2.0 % from 5.2 to 7.4 km, a summit 44 m above A, and a descent of −2.0 % from 8.2 to 9.8 km. A temporary restriction of 40 km/h can be imposed from 5.9 to 6.3 km, and the rail can be wet from 6.4 to 7.4 km, the upper part of the climb, or along the whole line. Twelve balises reset the odometer. All of these numbers were chosen for the lesson, and none comes from a railway.

Table 12.1 gives each variable its role. Eight states are integrated, the true position $s$ and speed $v$, the slip velocity $w$ of the motored wheelsets, the drive's lagged force demand $F_d$, and four energy integrals, the meter $E$, the traction energy drawn $E_t$, the regenerated energy returned $E_r$, and the braking work done at the rail $E_b$. The base unit is the second, and the trend axis reads minutes past one minute.

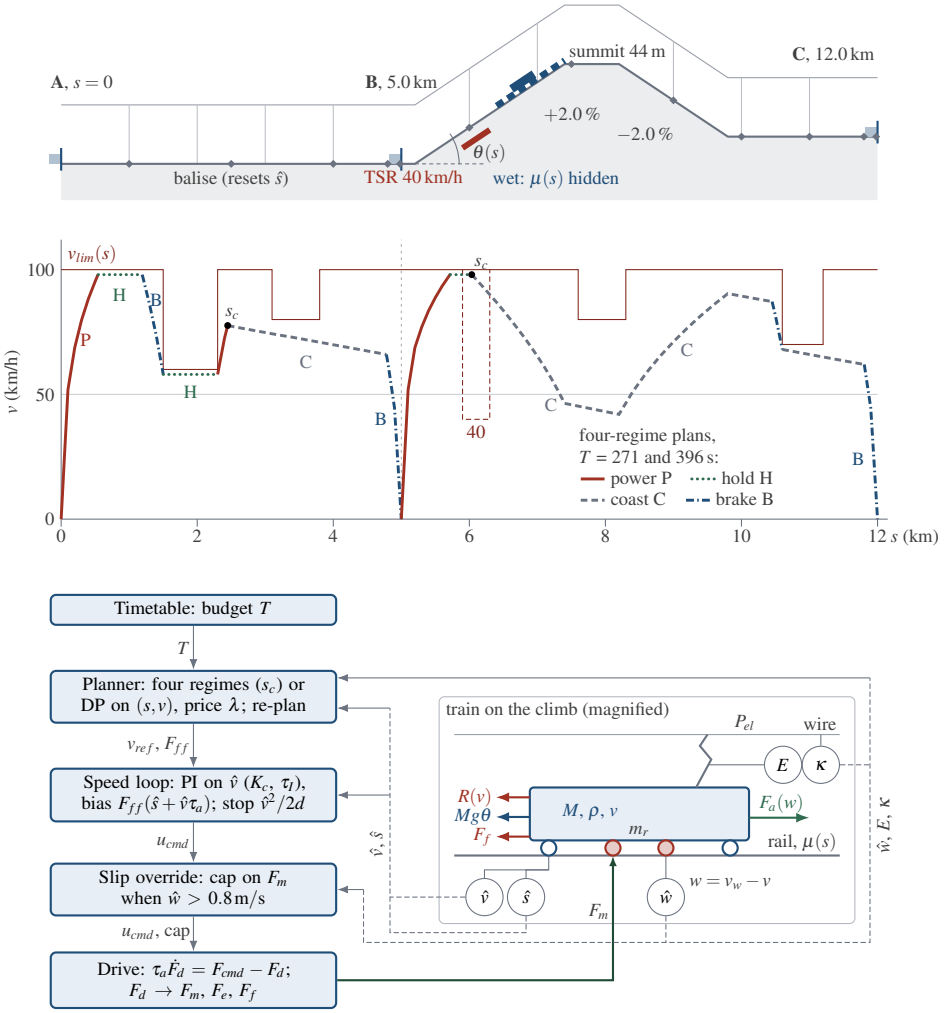



**Figure 12.1** The declared line, the train and its automatic train operation, after the simulator's Line view. Top, the profile with heights exaggerated, the stations, the overhead wire, the balises that reset the odometer $\hat{s}$, the 40 km/h restriction, the grade $\theta(s)$ and the wet stretch where the hidden adhesion $\mu(s)$ falls, with the train of Figure 12.4 on it. Middle, speed against distance with the posted limits $v_{lim}(s)$ and the simulator's four-regime plans for the timetable less the 4 s margin, drawn by regime (power solid, hold dotted, coast dashed, brake dash-dotted), with their coast points $s_c$. Bottom right, the train magnified, with the forces of Equations (12.1) to (12.3), the instruments $\hat{v}$, $\hat{s}$, $\hat{w}$, $E$ and $\kappa$. Bottom left, the control layers from the timetable's budget $T$ down to the drive, with the measurements returning on dashed lines. Not to scale.

### 12.3.1 Motion in Time

The train is one point mass on the route profile, with the motored wheelsets split off so that they can slip, as shown in Equation (12.1).

$$\begin{aligned} \frac{\mathrm{d}s}{\mathrm{d}t} &= v, \qquad R(v) = aM + bMv + cv^2 \\ (\rho M - m_r)\frac{\mathrm{d}v}{\mathrm{d}t} &= F_a(w) - F_f - R(v) - Mg\,\theta(s) \end{aligned} \tag{12.1}$$

where $M$ is the mass aboard, 190 t declared plus any added load, $\rho = 1.08$ is the rotating-mass factor, $m_r = 0.05M$ is the equivalent mass of the motored wheelsets, $F_a$ is the force the rail transmits to them, $F_f$ is the friction brake on the trailer axles, $\theta(s)$ is the grade as a fraction, and $R(v)$ is Davis resistance with $a = 13\,\mathrm{N/t}$, $b = 0.20\,\mathrm{N/(t\,m/s)}$ and $c = 6.0\,\mathrm{N/(m/s)^2}$, the Davis equation whose coefficients the train-resistance literature estimates for suburban and high-speed trains [18]. The resistance is 4.77 kN at 60 km/h and 8.16 kN at 100 km/h, where a coasting train on the level slows at $0.040\,\mathrm{m/s^2}$. Speed is clipped at zero, resistance and friction act only while the train moves, and a train at rest stays there unless the rail pushes it forward.

### 12.3.2 The Motored Wheelset and the Creep Curve

The motored wheelsets turn at $v_w = v + w$, driven by the motor force $F_m$ and held back by the rail, with a grip that rises with the slip velocity to a peak and falls beyond it, as shown in Equation (12.2).

$$\begin{aligned} m_r \frac{\mathrm{d}v_w}{\mathrm{d}t} &= F_m - F_a(w), \qquad w = v_w - v \\ F_a(w) &= \mu(s)\,\varphi M g\,\frac{2\,w/w_{pk}}{1 + (w/w_{pk})^2} \end{aligned} \tag{12.2}$$

where $\varphi = 0.5$ is the share of the mass on motored axles and $w_{pk} = 0.6\,\mathrm{m/s}$ is the slip of peak grip. The rail gives at most 233.0 kN dry at $\mu = 0.25$ and 74.6 kN wet at $\mu = 0.08$. Past the peak more slip means less grip, so a wheelset asked for more than the peak accelerates itself. A braking tire has the same peak, and the anti-lock brake of Volume 7, Chapter 4 regulates the slip at it. The curve keeps only the rising and falling branches that the creep-force literature describes [17], and its shape is a declared teaching choice. Its time constant, $m_r/(2\mu\varphi M g/w_{pk})$, is 12.2 ms on a dry rail and 38.2 ms on a wet one, and it sets the integration step.

### 12.3.3 Drive, Brakes and the Energy Meter

The drive's demand follows the command with a first-order lag, and the force that reaches the wheels is the demand cut by the tractive-effort curve, the motors' top speed and the slip override's cap, as shown in Equation (12.3).

$$\begin{aligned} \tau_a \frac{\mathrm{d}F_d}{\mathrm{d}t} &= F_{cmd} - F_d \\ F_m &= \min\big(F_d,\ \chi(v_w)\,F_{t,\max}(v_w),\ F_{cap}\big) \quad (F_d \ge 0) \\ F_B &= \min(-F_d,\ b_{\max}\rho M) \quad (F_d < 0) \\ F_e &= \min\big(F_B,\ F_{r,\max}(v),\ F_{cap,b}\big), \qquad F_m = -F_e \\ F_f &= \min\big(F_B - F_e,\ \mu(s)(1-\varphi)Mg\big) \\ F_{t,\max}(v) &= \min(F_t,\ P_t/v) \\ F_{r,\max}(v) &= \min(F_r,\ P_r/v)\min(1,\ v/v_f) \end{aligned} \tag{12.3}$$

where $\tau_a = 1.5\,\mathrm{s}$, $F_t = 220\,\mathrm{kN}$ and $P_t = 2.8\,\mathrm{MW}$ give a base speed of 45.8 km/h, the electric brake of $F_r = 200\,\mathrm{kN}$ and $P_r = 3.2\,\mathrm{MW}$ fades below $v_f = 2\,\mathrm{m/s}$, and $b_{\max} = 1.0\,\mathrm{m/s^2}$ is the full service brake. The taper $\chi = \min\big(1, \max(0, (V_w - v_w)/\Delta V_w)\big)$ with $V_w = 125\,\mathrm{km/h}$ and $\Delta V_w = 10\,\mathrm{km/h}$ makes an unchecked spin end at the motors' top speed rather than run away. Braking blends, with the electric brake taking what it can through the motored wheelsets and the friction brake on the trailers taking the rest under ideal wheel-slide protection. The meter integrates the electrical power at the pantograph, as shown in Equation (12.4).

$$\frac{\mathrm{d}E}{\mathrm{d}t} = P_{el} = \frac{F_m^{+}\,v_w}{\eta_t} - \kappa\,\eta_r\,F_e\,v_w \tag{12.4}$$

where $\eta_t = 0.85$ and $\eta_r = 0.80$, so a kilowatt-hour that goes out as traction and comes back through the brake keeps 0.68 of itself, and the receptivity $\kappa$ is the share of regenerated power the line accepts. The line voltage $V_0 - R_f P_{el}/V_0$, with $V_0 = 1500\,\mathrm{V}$ and $R_f = 0.06\,\Omega$, is displayed and drives nothing.

### 12.3.4 Timetable and Instruments

The timetable allows 275 s from A to B, a dwell of 30 s, and 400 s from B to C, with the first departure 10 s after a case starts. The flat-out plans take 250.8 and 311.7 s, so the runs carry 24.2 and 88.3 s of slack, the second large enough to recover a minute. The doors close at the later of the dwell's end and the timetable's departure, plus any door hold. The controllers read only the instruments of Equation (12.5).

$$\begin{aligned} \hat{v} &= q\big[v(1+\varepsilon) + \sigma_v\, n(t)\min(1, v/v_1)\big], \quad \hat{v} = 0 \text{ at rest} \\ \frac{\mathrm{d}\hat{s}}{\mathrm{d}t} &= v(1+\varepsilon), \qquad \hat{s} \leftarrow s_b + (s - s_b)(1+\varepsilon) \text{ at a balise } s_b \\ \hat{w} &= w(1+\varepsilon) \end{aligned} \tag{12.5}$$

where the tachometer on a trailer axle reads high by the hidden scale error $\varepsilon = 0.3\,\%$, $\sigma_v = 0.03\,\mathrm{m/s}$, $n(t)$ is colored unit noise with a correlation time of 0.5 s drawn from the case's seed, $v_1 = 1\,\mathrm{m/s}$, and $q$ rounds to 0.01 km/h. Balises stand at 1.0, 2.5, 4.0, 6.0, 7.5, 9.0, 10.0 and 11.0 km and 200 and 20 m before the marks at B and C. The meter reads to 0.01 kWh, and the drive's force, the line's receptivity, the signaled restriction and a door hold arrive as signals. The adhesion, the true position and speed and the mass aboard belong to the simulator alone. They are written to log columns whose names begin with `truth_` and drawn dashed as teaching truth, and a test that scans the source fails if the controller or the planner names one of them.

### 12.3.5 Constants, Starting State and Integration

Table 12.2 collects the constants. A case starts with the train at rest at its first station's mark, with the doors open, $w = F_d = 0$ and every energy integral at zero, and the controller holds the brake at –40 % of the lever while the doors are open. Fixed-step fourth-order Runge-Kutta integration advances the states every 10 ms, the controllers run every 0.1 s, and the log samples at 5 Hz, the three clocks of Section 1.6.1. Hard events, a balise passed, the wet stretch entered, an overspeed or a slip begun and a platform stop, are each handled once, after the integration step in which they occur, and marked on the trends.

### 12.3.6 Assumptions

The four cars are one rigid mass, so coupler forces and the train's length on a change of grade are absent, and a train 80 m long that meets a grade change over 3.6 s at 80 km/h feels it at once. Resistance has constant coefficients, with no curving term, tunnel factor or wind, and curves appear only as speed limits. One lumped wheelset with a declared creep curve stands for eight motored axles, so neither the cleaning of the rail by the leading axles nor the variation of adhesion along the rail is represented, and a long spin is where the model should be trusted least. The line is a receptivity and a feeder resistance, with no substations, no voltage limit on regeneration and no other trains except through $\kappa$. The drive's lag is a first-order teaching value, and the traction and regeneration efficiencies do not change with the operating point. Every speed limit, the 2 km/h overspeed margin and the $\pm 0.5$ m stopping accuracy are teaching values. On a railway they are set by the infrastructure manager's line data, the train protection system and the operator's platform specification.

## 12.4 The Structure of an Energy-Optimal Run

Written in distance, because the route is, the planning problem asks for the applied force $F(s)$ that spends the least electrical energy between a start and the next stop within a time budget, as shown in

**Table 12.2** Parameters. All Are Teaching Values, Adjustable in the Advanced Panel Unless Marked Fixed

| Symbol | Meaning | Value | Unit |
|---|---|---|---|
| $M$, $\rho$ | declared mass, rotating-mass factor | 190, 1.08 | t, — |
| $\varphi$, $m_r/M$ | motored share, wheelset inertia | 0.5, 0.05 | — |
| $F_t$, $P_t$ | starting effort, traction power | 220, 2.8 | kN, MW |
| $F_r$, $P_r$, $v_f$ | electric brake, power, fade | 200, 3.2, 2 | kN, MW, m/s |
| $b_{plan}$, $b_{\max}$ | planning, full service brake | 0.75, 1.0 | m/s$^2$ |
| $\tau_a$ | drive lag | 1.5 | s |
| $\eta_t$, $\eta_r$ | traction, regeneration efficiency | 0.85, 0.80 | — |
| $a$, $b$, $c$ | Davis coefficients | 13, 0.20, 6.0 | see text |
| $\mu_{dry}$, $\mu_{wet}$ | adhesion (hidden) | 0.25, 0.08 | — |
| $w_{pk}$ | slip of peak grip (hidden) | 0.6 | m/s |
| $V_0$, $R_f$ | line voltage, feeder resistance | 1500, 0.06 | V, Ω |
| $\varepsilon$, $\sigma_v$ | tachometer error (hidden), noise | 0.3, 0.03 | %, m/s |
| $\Delta v_m$ | plan margin under each limit, fixed | 2 | km/h |
| $K_c$, $\tau_I$ | speed loop (shipped default) | 10, 20 | %/(km/h), s |
| $\Delta t_m$ | punctuality margin | 4 | s |
| $\Delta t_{rp}$ | re-plan threshold | 5 | s |
| $\Delta s$, $\Delta v$ | DP grid | 50, 1 | m, km/h |
| $\Delta t$, $\Delta t_c$ | integration step, control period, fixed | 0.01, 0.1 | s |

Equation (12.6).

$$\begin{aligned} &\min_{F(s)} \int_{s_0}^{s_N} \left( \frac{F^+}{\eta_t} - \hat{\kappa}\,\eta_r F_e \right) ds \\ &\text{s.t. } \tfrac{1}{2}\rho M \frac{\mathrm{d}(v^2)}{\mathrm{d}s} = F - R(v) - Mg\,\theta(s) \\ &\qquad -F_b \le F \le \min\!\left(F_{t,\max}(v),\ \hat{\mu}(s)\varphi Mg\right) \\ &\qquad v \le v_{lim}(s) - \Delta v_m, \qquad \int_{s_0}^{s_N} \frac{ds}{v} \le T \end{aligned} \tag{12.6}$$

where $F_b = b_{plan}\rho M = 153.9\,\mathrm{kN}$ is the planning brake, $\hat{\mu}$ is the adhesion the planner is told, dry unless the wet-rail report says otherwise, $\hat{\kappa}$ is the receptivity it assumes, 1 unless the forecast is on, and $T$ is the timetable's running time less the punctuality margin, or, with re-planning, the time left to the scheduled arrival. The planner never reads the rail's true adhesion, the mass aboard or the model's state.

Attaching a price $\lambda$ to time and a co-state $p$ to the speed gives the Hamiltonian of Equation (12.7), which is linear in the force as the tank's is in Equation (1.10).

$$H = -\left( \frac{F^+}{\eta_t} - \hat{\kappa}\eta_r F_e \right) - \frac{\lambda}{v} + \frac{p}{\rho M v}\left( F - R(v) - Mg\theta \right) \tag{12.7}$$

The force that maximizes $H$ therefore sits at a bound or at zero, full power when $p/(\rho M v) > 1/\eta_t$, coasting between $\hat{\kappa}\eta_r$ and $1/\eta_t$, and full braking below $\hat{\kappa}\eta_r$, and speed holding is the singular arc on which $p/(\rho M v)$ stays at $1/\eta_t$. On a level line the arc requires $\lambda\eta_t = V^2 R'(V)$, so the price of time sets the holding speed, 77.8 kW for 60 km/h, 177.0 kW for 80 km/h and 317.9 kW for 98 km/h with the parameters of Table 12.2. This is the four-regime structure of the literature [3, 4]. It says what kind of drive to look for and leaves the switching points to be found.

🕮 **Pause.** Try Exercise 12.1 before reading on.

The switching points move with the budget, as shown in Figure 12.2. On the A to B run the one-coast-point plan coasts from 3.99 km at a budget of 252 s, from 2.45 km at 271 s and from 1.91 km at 330 s, and its planned energy falls from 26.81 to 15.46 kWh over that range, against 28.49 kWh planned flat out. The dynamic program plans 25.22, 15.45 and 9.60 kWh at the same three budgets, so its advantage grows from 6 to 38 % as the slack grows. Between 268 and 275 s its energy falls

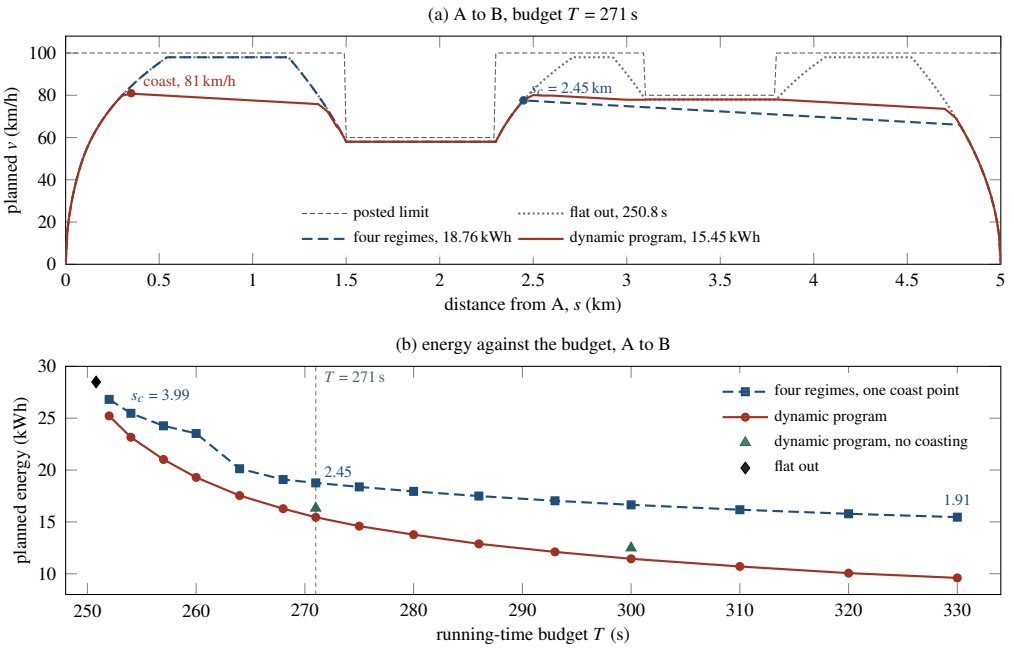



**Figure 12.2** Planned drives on the A to B run, from the simulator's own planners. (a) Speed against distance at the 271 s budget. The four-regime plan (dashed) holds 98 km/h, brakes into the 60 km/h curve and coasts from $s_c$, while the dynamic program (solid) coasts from 0.35 km into the curve. The flat-out plan is dotted and the posted limit thin dashed. (b) Planned energy against the budget, the four-regime plan dashed with squares and its coast point written beside three of its points, the dynamic program solid with circles, the same dynamic program barred from coasting at 271 and 300 s as triangles, and the flat-out plan as a diamond.

by 0.241 kWh per second of budget, 867 kW, against a converged price of 925 kW at 271 s, which is the meaning of $\lambda$ as the marginal value of time. The gap between the two planners is not a defect of either. The analytic plan holds the line limit and coasts once, while the true optimum lowers its hold speed as $\lambda$ falls and coasts wherever a limit or the platform would otherwise take the speed away. The no-coasting plans of Figure 12.2(b) were computed by calling the simulator's planner outside its interface, which offers the three planners and not a program barred from coasting. The flat 5 km line shows the other side. There the two planners agree within 0.21 % at 5, 10 and 15 % over the flat-out time, and a plan barred from coasting needs 7.2 to 8.8 % more energy than the four-regime plan.

🕮 **Pause.** Try Exercise 12.2 before reading on.

## 12.5 Planners and Tracking as Implemented

A plan is made at each departure for the run to the next station, and again whenever the supervisor re-plans. Each plan is a profile on a 5 m grid of speed, force, regime, time and energy that the tracking loop follows.

### 12.5.1 Flat Out and Four Regimes

The fastest drive applies full traction until the speed meets the braking envelope $v_b(s)$, the highest speed from which the planning brake still meets every lower limit ahead and the stop, and then

follows it, as shown in Equation (12.8).

$$\begin{aligned} \tfrac{1}{2}\rho M \frac{\mathrm{d}(v_b^2)}{\mathrm{d}s} &= -\big(F_b + R(v_b) + Mg\theta\big) \\ v_b(s_N) &= 0, \qquad v_b \le v_{lim} - \Delta v_m \end{aligned} \tag{12.8}$$

integrated backward from the mark by the fourth-order Runge-Kutta method in $v^2$ with 5 m steps. Following the envelope means holding a limit, or braking along the curve into a lower limit and into the platform. A speed bounded by the distance needed to stop is also how the warehouse robot of Volume 20, Chapter 9 limits its speed. The four-regime plan is the same drive left at a coast point $s_c$, after which the force is zero except where the envelope caps the speed, and it never powers again. A later coast point always arrives sooner, so $s_c$ is found by bisection until the arrival is within 1 ms of $T$ or after 60 halvings. On both runs of the line at their timetables the worst miss is 0.99 ms.

### 12.5.2 The Dynamic Program

The run is cut into stages of $\Delta s = 50\,\mathrm{m}$ and the speed into nodes $\Delta v = 1\,\mathrm{km/h}$ apart, 100 stages by 99 nodes from A to B and 140 by 99 from B to C. The cost-to-go is computed backward from the platform, as shown in Equation (12.9).

$$\begin{aligned} V_N(0) &= 0 \\ V_i(v_j) &= \min_{d \in \mathscr{D}} \Big[ e_i(v_j, d) + \lambda\, t_i(v_j, d) + V_{i+1}\big(v_i^{+}(v_j, d)\big) \Big] \end{aligned} \tag{12.9}$$

where $e_i$ and $t_i$ are the energy and the time of stage $i$ under decision $d$, $v_i^{+}$ is the speed at its end, and $V_{i+1}$ is interpolated linearly between nodes and infinite above a limit or above the braking envelope. It is the recursion of Equation (1.11) written in distance instead of time, with time priced by $\lambda$, and the same backward recursion made stochastic gives the water value of Chapter 7. The decision set $\mathscr{D}$ holds eleven moves, full traction, hold, coast and the planning brake, traction at three quarters, a half and a quarter, a half brake, power and coast each capped by the braking envelope, and a terminal stop. Every move is integrated over its stage in ten 5 m substeps of the fourth-order Runge-Kutta method in $v^2$, energy per substep is $Fh/\eta_t$ for $F \ge 0$ and $-\hat{\kappa}\eta_r \min(-F, F_{r,\max})h$ for braking, and a capped move that would have to brake harder than the planning brake to follow the envelope is refused. The terminal stop is the constant braking force $F_{end} = \tfrac{1}{2}\rho M v^2/d - aM - \tfrac{2}{3}bMv - \tfrac{1}{2}cv^2 - \bar{G}$ that stops the train at the mark, where $\bar{G}$ is the mean grade force over the distance $d$, allowed between $0.3F_b$ and $F_b$. The part forces let a switch fall inside a stage, so the energy is not quantized by the stage length. The four regimes are still the only regimes the plans contain, but the program is told neither their order nor how many switches to make.

The budget enters through the price. A forward pass from the measured start takes at each stage the decision that minimizes the bracket of Equation (12.9) at the exact speed reached, and $\ln\lambda$ is bisected between 10 W and 1 GW until the interval is below $10^{-4}$, which takes 18 bisections, keeping the lowest price whose plan meets $T$. Because the decisions are discrete, the plan at the bracketing higher price arrives early and the one at the lower price late, so at the first stage where the two differ the force is blended as $(1-\alpha)F_{hi} + \alpha F_{lo}$, with $\alpha$ bisected 30 times so that the arrival lands on the budget. If even the highest price misses $T$, the run cannot be made in time. The controller then drives the flat-out profile and reports the timetable impossible only if flat out also misses it. Transitions do not depend on $\lambda$ and are computed once per run and condition, 81 008 feasible ones from A to B and 119 789 from B to C, so a re-plan reuses them.

### 12.5.3 Tracking, Stopping and the Slip Override

The speed loop is the velocity-form proportional-integral law of Section 1.6.2 with the plan's force fed forward as its bias, run every $\Delta t_c = 0.1\,\mathrm{s}$ on the measured speed in km/h, as shown in Equation (12.10).

$$\begin{aligned} e_k &= v^*_k - \hat{v}_k, \qquad P_k = K_c e_k \\ \sigma_k &= \sigma_{k-1} + P_k - P_{k-1}, \qquad \hat{u}_k = L_{ff,k} + \sigma_k + I_{k-1} \\ u_{cmd,k} &= \min\bigl(100, \max(-100, \hat{u}_k)\bigr) \\ I_k &= I_{k-1} + \Bigl[\frac{K_c}{\tau_I} e_k + \frac{u_{cmd,k} - \hat{u}_k}{\tau_I}\Bigr]\Delta t_c \end{aligned} \tag{12.10}$$

where the reference $v^*$ is the plan's speed $v_{ref}(\hat{s})$, capped by the braking envelope of the limits in force now, and on a coasting stretch it equals $\hat{v}$ while the error is within 2 km/h, so that a coast stays a coast. The bias $L_{ff}$ is the plan's force at $\hat{s} + \hat{v}\tau_a$, read one drive lag ahead and expressed as a lever, and the lever maps to the command force as $F_{cmd} = (u_{cmd}/100)F_{t,\max}(\hat{v})$ in power and $(u_{cmd}/100)\,b_{\max}\rho M$ in braking, with the declared mass. The loop is reverse acting, so a measured speed below the reference raises the lever. The running sum and the back-calculation are those of Equations (1.15) to (1.17), with the tracking time $\tau_I$, and $\tau_I = 0$ removes integral action.

The stopping controller takes over where the plan's final braking begins, or earlier once $\hat{v}^2/(2d)$ reaches 85 % of $b_{plan}$ within 1500 m of the mark. It feeds forward the braking force $F_S = \rho M \hat{v}^2/(2d) - R(\hat{v}) - Mg\theta$ through the declared mass and runs the same loop against the curve $\sqrt{2 b_{plan} d}$, where $d = s_{mark} - \hat{s}$. A train at rest short of the mark by more than 10 m creeps in at 15 % of the lever, and the doors open after 0.4 s at rest within $\pm 10$ m.

The slip override acts on the drive's cap every control period, as shown in Equation (12.11).

$$F_{cap,k} = \begin{cases} \max\bigl(F_{lo},\ 0.6\min(|F_m|, F_{cap,k-1})\bigr) & \hat{w} > w_{on} \\ \min\bigl(F_t,\ F_{cap,k-1} + r\Delta t_c\bigr) & \hat{w} < w_{off} \end{cases} \tag{12.11}$$

where $w_{on} = 0.8\,\mathrm{m/s}$, $w_{off} = 0.4\,\mathrm{m/s}$, $F_{lo} = 5\,\mathrm{kN}$, $r = 10\,\mathrm{kN/s}$, and the cap is removed once it recovers to $F_t$, with the mirror rule on $F_{cap,b}$ for negative slip in electric braking. The cap acts on the motor force directly, not through the lag, and while it binds the speed loop tracks the capped lever so that its integral does not wind up. The force therefore hunts about the rail's peak.

### 12.5.4 Re-Planning and the Three Models

Every control period the supervisor projects the arrival as the time now plus the plan's time still to run from $\hat{s}$, and compares it with the timetable's arrival less the margin. When the difference $\Delta t$ exceeds the threshold $\Delta t_{rp}$, or the restriction changes, it plans again from the next 50 m node with the measured speed and the time left, at most once every 20 s, not within 400 m of the mark and not when the plan in force is already flagged impossible and late. The note it writes states the energy the new plan spends beyond the old plan's remainder. Three models stay apart. The plant is everything in Section 12.3. The planner's model is Equation (12.6) with the declared train, a dry rail unless told and no drive lag. The stopping controller uses the declared mass, so a heavier train than declared is corrected only by feedback.

## 12.6 Nine Cases and a Challenge, Measured

Table 12.3 lists the cases with the rule that completes each. Every result below is a headless replay of the train engine on its default seed, in the 10 ms steps a browser tab would take. A setting that a sentence names is applied before the first step, which primes the run exactly as Reset does (Section 1.6.1). A case is credited for the reader's actions together with the state the simulated train reaches, a scripted event earns no credit, and a parameter, wet-rail or receptivity change during the challenge voids the attempt. The three ways of reaching a case (loading it first, picking it after another

**Table 12.3** The Nine Cases and the Challenge, With What Completes Each

| # | Case, settings and events | Teaching point | Required to complete |
|---|---|---|---|
| 1 | By hand, A to B, Manual, lever in full power | the clock and the meter | Leave A in Manual, stop within 20 m of the B mark, and confirm reading the arrival and the energy. |
| 2 | Flat out, A to B | the fastest run and its cost | Run flat out to B and confirm reading drawn, returned and net energy. |
| 3 | Four regimes, A to B | one coast point | Run the four-regime plan to B, choose Flat out and Reset, and run it to B. |
| 4 | Four regimes, B to C on a 330 s express path | hold uphill, coast over the top | Watch the plan hold up the climb, switch to coasting between 7.0 and 8.2 km, and run to C. |
| 5 | Dynamic program, A to B, Grid view | where limits interleave | Run the program to B, choose Four regimes and Reset, and run it to B. |
| 6 | Dynamic program, $K_c = 1$, no integral, 40 t unplanned load | tracking and the platform mark | Raise $K_c$ to 8 or more with integral action and stop within $\pm 0.5$ m of the B mark. |
| 7 | B to C, override off, wet top of the climb, restriction | slip and what the plan knew | Spin the wheels past 5 m/s with the override off, turn it on, report the wet stretch, Reset, and reach C within $\pm 15$ s with the wet slip under 2 m/s. |
| 8 | B to C, doors held 60 s, re-planning off | the price of a minute | Arrive late by more than 15 s, turn re-planning on, Reset, and arrive within $\pm 15$ s. |
| 9 | B to C, restriction, doors held 90 s | a timetable no plan can meet | Watch the planner report the timetable impossible while running, run to C, and confirm the late arrival. |
| 10 | Challenge, A to C, wet top of the climb, doors held 25 s at B, restriction announced at 330 s | the whole stack | Both arrivals within $\pm 15$ s, both stops within $\pm 0.5$ m, no overspeed, and the meter at C below 75 % of the flat-out run on this case, 55.1 kWh of 73.4 kWh, with the case's train, line and events.[a] |

[a] Shipped with the four-regime plan, $K_c = 1$, no integral, no re-planning and no slip override, it stalls on the wet climb. The suggested settings are the dynamic program, $K_c = 10$, $\tau_I = 20$ s, re-planning, the override and the wet-rail report for the top of the climb.

case, and Reset part-way through) give logs with identical hashes for all ten, the application's suite passes 43 of 43 checks, and a run paced at 1 × and at 30 × by wall-clock frames ends bit-identical.

### 12.6.1 Coasting, Limits and a Grade

Driven by hand with the lever left in full power, the train of case 1 enters the 60 km/h curve 63.0 km/h over its limit. A cautious drive, full power to 25 s, the lever at 3 % to 320 s and then −60 %, never exceeds 58.6 km/h and arrives 62.9 s late and 7.7 m short with 10.5 kWh. It saves energy by spending the passengers' time. Flat out in case 2 draws 58.9 kWh and returns 30.7 kWh, 75 % of the braking work at the rail, for a net 28.2 kWh, and the 21.0 s it gains buy nothing because the train waits at B for the timetable. The coast point of case 3 and the dynamic program of case 5 reproduce the comparison of the Introduction, 19.5 kWh against 16.4 kWh when driven, 16 % less, as listed in Table 12.4.

**Table 12.4** Measured Case Results. Arrival Against the Timetable (+ Late), True Stopping Error (+ Past the Mark) and Energy at the Mark

| # | Configuration | Arrival (s) | Stop (m) | Energy (kWh) |
|---|---|---|---|---|
| 1 | cautious hand drive | B +62.9 | −7.7 | 10.5 |
| 2 | flat out | B −21.0 | −0.05 | 28.2 |
| 3 | four regimes, $s_c$ 2.45 km | B −3.1 | −0.02 | 19.5 |
| 4 | four regimes, $s_c$ 6.96 km | C +0.6 | −0.02 | 32.7 |
| | flat out | C −14.7 | — | 43.6 |
| | margin 8 s, $s_c$ 8.35 km | C −3.3 | — | 35.9 |
| 5 | dynamic program | B −1.7 | −0.01 | 16.4 |
| 6 | $K_c = 1$, no integral | B +6.6 | +6.86 | 17.6 |
| | $K_c = 10$, $\tau_I = 20$ s | B −0.5 | +0.11 | 19.6 |
| | $K_c = 10$, no integral | — | +0.56 | — |
| 7 | override off | stalls at 7.10 km, slip 34.6 m/s | | |
| | override on | C −0.6 | 0.00 | 30.8 |
| | override and wet report | C −1.2 | +0.02 | 29.2 |
| 8 | no re-planning | C +60.8 | −0.06 | 26.1 |
| | re-planning | C −0.6 | −0.00 | 30.5 |
| 9 | any planner, margin or gain | C +42.4 | −0.03 | 47.2 |

**TRY IT IN THE APP** **3 · The Coast Point: One Decision, and What It Saves**

Load the scenario and press Run. The train leaves A under full power on the four-regime plan. Watch the regime change to coasting 2.45 km from A, after which the train never powers again before B. It arrives 3.1 s early with 19.5 kWh on the meter.

Case 4 runs B to C on a 330 s express path. Holding 98 km/h up the 2.0 % grade takes 45.2 kN of the 102.9 kN the motors give at that speed, and the plan coasts from 6.96 km, 443 m below the summit, so the train crosses it at 84 km/h and the descent restores the speed with no power. Coasting at 80 km/h the train loses 0.212 m/s$^2$ up the grade and gains 0.151 m/s$^2$ down it. It arrives 0.6 s late with 32.7 kWh against 43.6 kWh flat out. An 8 s margin shrinks the budget by 4 s and moves the coast point past the summit to 8.35 km, for 35.9 kWh, while on the ordinary 400 s timetable the coast point falls at 6.04 km, near the foot of the climb. Figure 12.3 draws the energies and arrivals of Table 12.4 side by side.

### 12.6.2 Tracking a Heavier Train Than Planned

In case 6, 40 t more than the declared mass boards at the first second. With $K_c = 1$ % per km/h and no integral action the train stops 6.86 m past the B mark and 6.6 s late, because the stopping feedforward brakes for 190 t with 230 t aboard. With $K_c = 10$ and $\tau_I = 20$ s it stops 0.11 m past and 0.5 s early, and with $K_c = 10$ and no integral action 0.56 m past, outside the declared accuracy. Proportional action leaves an error in proportion to the missing force, and the integral removes it.

**Pause.** Try Exercise 12.3 before reading on.

### 12.6.3 A Wet Climb

Case 7 leaves B with the restriction in force and the rail wet over the upper climb, which the plan does not know. Accelerating out of the restriction with no override, the motors can deliver 168 kN at 60 km/h where the rail gives 74.6 kN, the wheels spin up to 34.6 m/s of slip at the motors' top speed, and the train stalls at 7.10 km. With the override on, the largest slip on the wet stretch is 1.89 m/s, the train climbs at the rail's limit, falls behind its dry-rail plan, re-plans three times and arrives 0.6 s early with 30.8 kWh. With the wet-rail report as well the plan expects 74.6 kN there, the slip peaks at 1.57 m/s, one re-plan suffices and the arrival is 1.2 s early with 29.2 kWh. Over six seeds the override alone holds the wet slip between 1.38 and 2.37 m/s, above the declared 2 m/s at one seed, and with the report at or below 1.72 m/s. The peak comes within one 0.1 s control period of the wet entry.

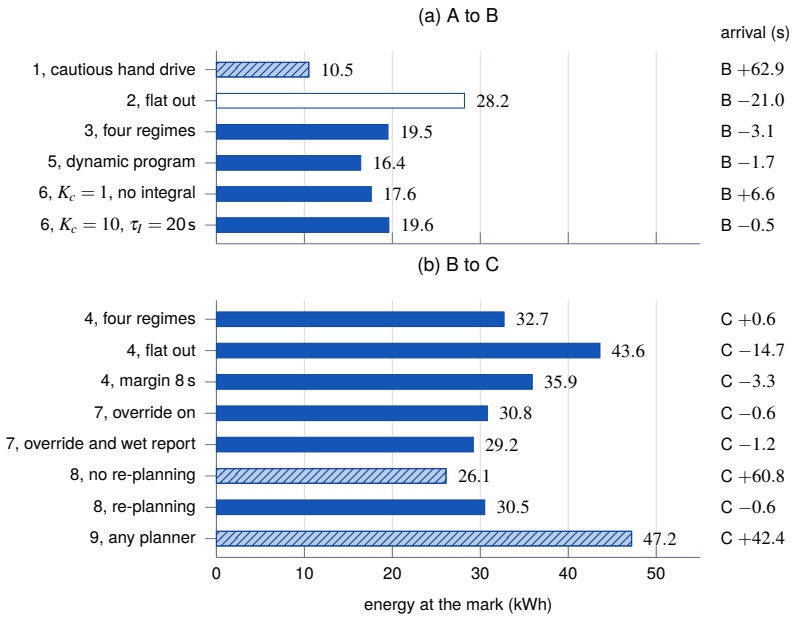



**Figure 12.3** The measured cases of Table 12.4: energy at the mark (bars, with values) and arrival against the timetable (right, + late) for (a) the runs from A to B and (b) the runs from B to C. Dark bars arrive within 15 s of the timetable, the open bar more than 15 s early and the hatched bars more than 15 s late. Notice that on each leg the cheapest run is a late one, and that among the runs on time from A to B the dynamic program spends the least.

> **TRY IT IN THE APP** **7 · The Wet Rail: Slip, the Override, and What the Plan Knew**
>
> Load the scenario and press Run. The slip override is off, and the top of the climb is wet, which the plan does not know. Watch the wheel slip as the train accelerates out of the restriction: the wheels spin up to 34.6 m/s of slip, and the train stalls at 7.10 km.

### 12.6.4 A Minute Late, and a Timetable No Plan Can Meet

In case 8 the doors at B are held 60 s. Without re-planning the train runs its on-time plan and arrives 60.8 s late with 26.1 kWh. Re-planned at departure with a 336 s budget it arrives 0.6 s early with 30.5 kWh, against 26.1 kWh for the same run leaving on time, so the minute costs 4.4 kWh, 17 % more. Case 9 adds the restriction to a 90 s hold. The fastest plan from B to C with the restriction takes 348.5 s, the timetable leaves 310 s, and the planner reports the timetable impossible and hands the loop the flat-out profile. The train arrives 42.4 s late with 47.2 kWh, and a logged sweep over the three planners, margins of 0 and 10 s and gains of 10 and 20 finds no arrival earlier than 42.4 s late. The flag switches between door holds of 47 and 48 s, where flat out stops meeting the budget.

### 12.6.5 The Challenge

The challenge runs A to C with the wet top of the climb, a 25 s door hold at B and the restriction announced at 330 s, while the train stands at B. Of the 88.3 s of slack on the B to C timetable, the hold takes 25 s, the restriction 36.8 s and the wet rail, once the plan knows it, 8.4 s more. As shipped the train reaches B 0.6 s early and stalls on the wet climb at 7.10 km on six of six seeds. The suggested strategy passes on six of six, as listed in Table 12.5, with C reached between 2.8 and 1.4 s early, 51.7 to 52.0 kWh and both stops within 0.08 m. Removing any one element fails the run. The four-regime plan meets the clock and spends 58.8 kWh, the run without the wet-rail report arrives 15.1 s late with

**Table 12.5** The Challenge, Default Seed. Arrival (s) and Stopping Error (m) at Each Station. Energy at C, or at 760 s for Stalled Runs

| Configuration | B | C | kWh | Verdict |
|---|---|---|---|---|
| shipped | −0.6, −0.07 | stall 7.10 km | 61.3 | fails |
| suggested | −1.7, −0.01 | −1.4, −0.04 | 51.8 | passes |
| flat out, rest suggested | −21.0, −0.05 | −10.6, −0.06 | 73.4 | fails |
| four regimes, rest suggested | −3.1, −0.02 | +0.9, −0.03 | 58.8 | fails |
| suggested, no wet report | −1.7, −0.01 | +15.1, −0.00 | 57.9 | fails |
| suggested, no re-planning | −1.7, −0.01 | +28.7, −0.01 | 44.6 | fails |
| suggested, no override | −1.7, −0.01 | stall 7.08 km | 55.4 | fails |

**Table 12.6** Code Verification

| Check | Result |
|---|---|
| meter against $\int P_{el}\,dt$, every 10 ms, A to B | $1.7 \times 10^{-7}$ |
| work balance, challenge suggested, wet, hold, restriction | $1.3 \times 10^{-5}$ |
| work balance, case 7 spin and stall | $2.6 \times 10^{-5}$ |
| free rolling against $R/\rho M$, 40 to 100 km/h | $1.6 \times 10^{-9}$ |
| coasting times against the closed form, three bands | $\leq 1.3 \times 10^{-4}$ |
| halving $\Delta t$, A to C energy | 0.007 % |
| grid 50 m by 1 km/h against 25 m by 0.5 km/h | 0.05 % |
| priced DP against enumeration, three prices | $\leq$ 0.316 % |
| DP without blend against constrained enumeration | 0 to 5.42 % |
| DP against four regimes, flat 5 km line | $\leq$ 0.21 % |
| DP fastest plan against flat out | $\leq$ 0.19 s |
| four-regime arrival against $T$ | $\leq$ 0.99 ms |

57.9 kWh, the run without re-planning arrives 28.7 s late, and the run without the override stalls at 7.08 km. The suggested run on the wet climb is shown in Figure 12.4.

> ✦ DID YOU KNOW? **The Cheapest Challenge Run Fails**
>
> From the simulator: in the challenge, the suggested settings without re-planning reach C with 44.6 kWh, the least energy of every run in Table 12.5, and still fail, because the train arrives 28.7 s late. The suggested run spends 51.8 kWh and arrives 1.4 s early. The challenge accepts an arrival only within ±15 s (Table 12.3).

### 12.6.6 Code Verification

The checks of Table 12.6 test the implementation against itself and against closed forms, and none of them compares the model with a railway. The meter equals the integral of the logged power to $1.7 \times 10^{-7}$. Over the suggested challenge the work the motors do at the wheels, 27.6 kWh, equals the change of kinetic energy plus the slip loss of 1.85 kWh, the work against resistance and friction, and the gravity term of 6.213 kWh, which matches $Mg\Delta h$, to a residual of $1.3 \times 10^{-5}$. Halving the step to 5 ms moves a whole A to C trip's energy by 0.007 % and keeps every challenge verdict, the largest case change being an arrival of case 7 that moves from 0.6 to 1.6 s early. Refining the grid from 100 m by 2 km/h to 50 m by 1 km/h and to 25 m by 0.5 km/h gives 15.589, 15.449 and 15.441 kWh for the A to B plan at 271 s, changes of 0.91 and 0.05 %. On a 600 m instance with a grade and a limit, the priced policy matches an enumeration of all 44 244 sequences of the four regimes and the stop exactly at two prices and within 0.316 % at a third. On a 300 m instance with all eleven decisions and the budget as a hard constraint, the program without its blend lies 0.00 to 5.42 % above the enumerated optimum over ten budgets, and with the blend at or below it, because the blend uses a force the enumeration lacks. A plan with its transitions takes 163 to 224 ms in Node.js, and two timed runs of the challenge at 30 × in a desktop browser took 24.2 s of wall time each, with plans of 156 to 239 ms and re-plans of 65 to 84 ms.

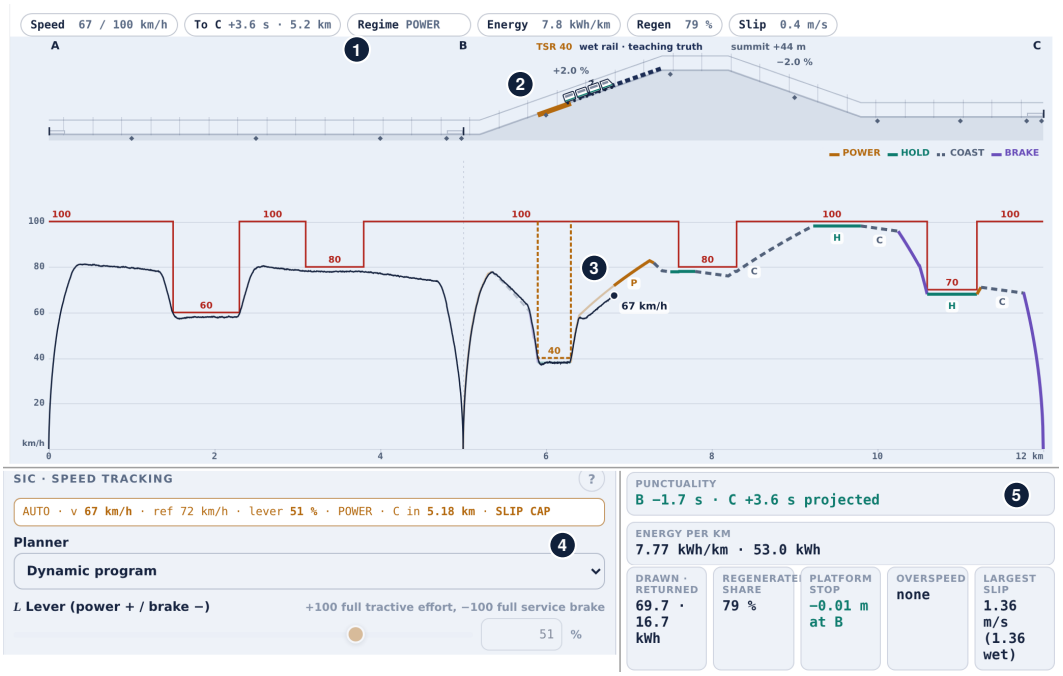



**Figure 12.4** Suggested settings in the challenge, $t = 470$ s, on the Line view. The figure joins three parts of the screen, the Line view with its badges (top), the speed-tracking card (bottom left) and four tiles of the Learn panel (bottom right), and leaves out the header, the rest of the control panel, the trends and the other tiles. In this run the train left B 26.1 s after the timetable. (1) The badges read 67 km/h, C projected 3.6 s late, the regime and the regenerated share. (2) The train ran through the 40 km/h restriction from 5.9 to 6.3 km and climbs the wet stretch. (3) The dynamic program's plan to C lies ahead under the limits, lettered by regime, and the run lies behind. (4) The speed-tracking card reads 67 km/h against a reference of 72 km/h, the lever at 51 % and SLIP CAP. At 6.82 km the slip override holds the motor force at 71.0 kN under the wet rail's 74.6 kN, with a measured slip of 0.42 m/s (Table 12.1). (5) The tiles read B 1.7 s early and 0.01 m short, C projected 3.6 s late, 53.0 kWh on the meter with 69.7 drawn and 16.7 kWh returned, 79 % of the braking work returned, and a largest slip of 1.36 m/s. A re-plan follows at 492.7 s, and the train reaches C 1.4 s early with 51.8 kWh (Table 12.5).

### 12.6.7 Agreement With the Literature

The agreement claimed with published work is structural only. Every plan the dynamic program returns is made of the four regimes that the maximum principle predicts [3, 4], and on a flat 5 km line a direct search over power, hold at 50 to 98 km/h, coast and a final brake of 0.3 to 1.0 times the planning brake finds plans within 0.26 % of both planners at 5, 10 and 15 % over the flat-out time. No number in the model was fitted to measured trains, and the Davis coefficients are teaching values in the quadratic form [18], not a measured curve. The energy savings of 10 to 30 % reported for planned driving in [8] are against a baseline that is not the flat-out run used here, so the 31 % of case 3 is not compared with them.

## 12.7 Technology in Practice

### What Measures It

Odometry, the onboard function that estimates a train's speed and travelled distance by dead reckoning, is the base of automatic train protection and control. Besides wheel angular speed sensors it can use

Doppler radar, accelerometers and gyroscopes, and a single sensor is reliable only over a limited range of conditions. Satellite positioning is held back on railways by lines that are not georeferenced and by signal outages, in tunnels for example. European train control requirements set the speed error and the distance error that an odometry algorithm may make [19]. In this chapter the controllers read only the tachometer, the odometer with its balise resets, the measured slip and the signals of the line (Equation (12.5)), and the energy meter is logged.

**TECHNOLOGY Sensor: Tachometers and Balises for Speed and Position**

Onboard odometry estimates a train's speed and travelled distance by dead reckoning, usually from wheel angular speed sensors that count pulses per revolution. In pure rolling the speed is the wheel radius times the angular speed, but wheels wear and are periodically returned, so the radius may not be reliable. When adhesion is degraded and the train accelerates or brakes, the wheels slide: a driven wheel runs faster than the train and a braked one slower, and the error may become very high. The estimate of distance is reset each time the train passes a balise [19]. The chapter's tachometer sits on a trailer axle and reads 0.3 % high, and twelve balises reset its odometer (Equation (12.5)).

### What Moves It

Electric train operation divides into four modes, acceleration, cruising, coasting and braking. On urban lines, where stations are close together, cruising is typically omitted: trains accelerate rapidly to their top speed and brake shortly afterwards for the next stop [5]. The same wheel angular speed sensors serve the wheel slide protection and anti-skid systems that act at the wheels when adhesion is poor [19]. In this chapter the drive answers the lever with a 1.5 s lag, the electric brake works through the motored wheelsets, and the friction brake on the trailers takes the rest (Equation (12.3)).

**TECHNOLOGY Final Element: Traction Motors That Also Brake**

In regenerative braking, common in electric rail systems, a train slows by reversing the operation of its motors, which then act as generators. The energy they produce supplies the train's own auxiliary loads, and the surplus is fed back to the line that supplies the trains, where a nearby train that is accelerating can use it. With no train near, the line voltage rises toward an overvoltage limit, the braking train may be unable to return its energy, and the excess is burned as heat in onboard or wayside resistors [5]. The receptivity $\kappa$ of Equation (12.4) is the share the line accepts, and the friction brake on the trailers takes what the electric brake cannot.

### Where the Controller Runs

An automatic train operation system first generates a target speed curve from the condition of the train and of the line, and then sends speed control commands so that the train tracks it [20]. Under it a train control system enforces the speeds that protection allows [21]. No open source read for this chapter gives the scan period of an onboard computer. In this chapter the planner solves a plan in 156 to 239 ms in a browser, and the supervisor, the speed loop, the stopping controller and the slip override run every 0.1 s (Figure 12.5), where a real drive's slip control runs in milliseconds (Section 12.8).

**IN PRACTICE Protection Sits Under the Plan**

On railroads in the United States, a federal rule requires each positive train control system to reliably prevent overspeed derailments, including those related to civil engineering speed restrictions and slow orders. An automatic train control system must apply the brakes automatically when the train's speed exceeds the rate its speed control mechanism sets [21]. Protection depends on odometry, and an error in position may overestimate the distance available for braking. One national protection system is forced to overestimate speed for safety, so its emergency braking occurs more often than required [19]. The plans of this chapter keep 2 km/h below each limit (Table 12.2).

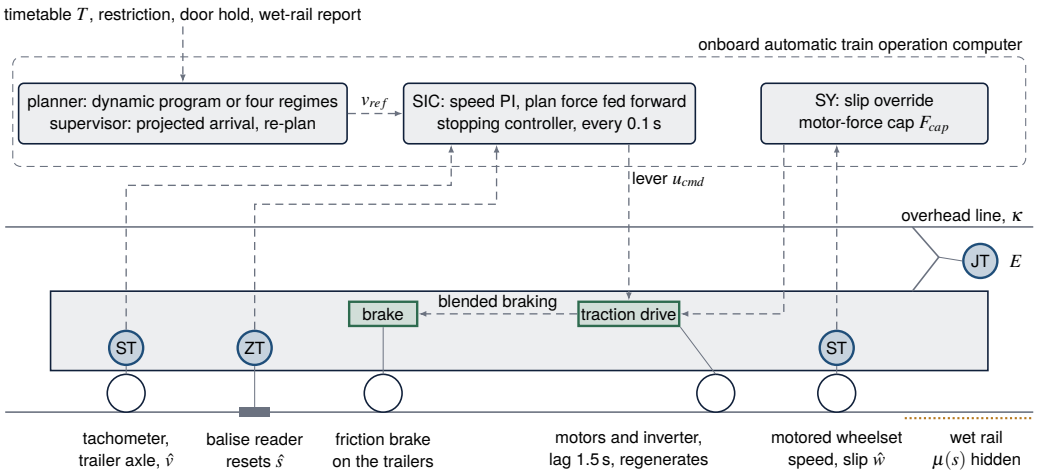



**Figure 12.5** The chapter's control layers drawn with real devices. The onboard computer plans the run from the timetable and the signals it is given, the speed loop (SIC) tracks the plan with the lever of the traction drive, and the slip override (SY) caps the motor force when the motored wheelsets slip. The tachometer (ST) on a trailer axle gives $\hat{v}$, the balise reader (ZT) resets the odometer $\hat{s}$, a speed sensor on the motored wheelsets gives the slip $\hat{w}$, and the energy meter (JT) at the pantograph integrates the power drawn and returned. Notice that no instrument reads the rail's adhesion: the plan learns of a wet rail only from a report.

### What Is Modernizing It

Three developments bear on this chapter's method. The first is odometry that fuses tachometers with inertial measurement units through Kalman filters, aimed at accuracy in critical adhesion conditions. Its preliminary results estimated position and speed markedly better than an algorithm in service [19]. The second is the recovery of braking energy that the line cannot take. Wayside energy storage has reported savings of up to 30 percent in service, and reversible substations feed regenerated energy back to the AC grid [5]. The third is learning. Algorithms built from the records of experienced subway drivers and refined by reinforcement learning used less energy than manual driving and existing automatic operation in simulations with field data from one line [20]. In this chapter the receptivity forecast and the dynamic program stand where such storage and such learning would enter.

## 12.8 Limitations

The train, the line and the timetable are teaching values that make a consistent run, not data from any railway, and a simulated pass of the challenge is not evidence about any real train's punctuality or energy. The point mass, the single lumped wheelset and the constant-coefficient resistance leave out coupler forces, the train's length on a grade, curving resistance and the way adhesion recovers along a train, and the line has no substations and no other trains except through the receptivity.

> ⚠ **OPEN CHALLENGE** **Automatic Operation Still Hands Back to the Driver**
>
> A study of automatic train operation on subways listed problems that remain unsolved. Most work weighs one or two objectives, although safety, punctuality, energy and riding comfort all count. The automatic system tracks a speed profile optimized before the run, which limits its flexibility, and when the trip time or the track's resistance changes, the subway transfers to manual driving. The study's answer adapts to such changes without an offline speed profile [20]. The supervisor of this chapter re-plans for a late departure (Section 12.5).

The slip override runs at the controller's 10 Hz, where a real drive's slip control runs in milliseconds, and on case 7 without the wet-rail report its peak slip reaches 2.37 m/s at one of six seeds, above the declared 2 m/s. The dynamic program is exact on its grid only. Its switches fall on 50 m stages, its cost-to-go is linear between nodes, and a budget that falls between two prices is met by blending one stage rather than by an exact constrained optimum, a duality gap that the blend closes only approximately. The four-regime planner holds each line limit and coasts once, which is a restricted family rather than the full Pontryagin solution with a hold speed set by $\lambda$, and its distance from the dynamic program on the A to B run measures that restriction as much as the line. In the challenge the restriction is announced while the train stands at B, so the plan to C is solved twice, at the arrival and at the departure, and each solve costs a visible frame of about 0.2 s. The receptivity forecast and the punctuality margin are the only economics, with no price of energy that varies in time.

**OPEN CHALLENGE** **Knowing the Rail Before the Wheels Reach It**

Most slip control on trains acts after a wheel has begun to slide, so damage to wheels and track cannot be avoided. Predictive slip controls have assumed constant traction over the prediction horizon and fail to avoid sliding when traction changes fast. Rain mixing with dust into a muddy film on the rails is one cause of poor traction. One proposal records traction conditions in a map along the line and lets a predictive controller read them ahead, and it names the online adaptation of that map from measurements as future work [22]. In case 7 the plan learns of the wet rail only from a report.

## 12.9 Conclusion

This chapter studied a four-car electric train on a declared 12 km line, a plant on which the energy of a run is decided by where the regime changes. Flat out between A and B costs 28.2 kWh for an arrival 21.0 s earlier than the timetable needs. One coast point brings it to 19.5 kWh on time, and a dynamic program on a 50 m by 1 km/h grid, which coasts from 0.35 km into a 60 km/h curve, to 16.4 kWh. The program's advantage over the one-coast-point plan grows from 6 to 38 % of planned energy as the budget grows from 252 to 330 s, and on that run a drive barred from coasting beats the one-coast-point plan by 13 %, while on a level line it loses by 7.2 to 8.8 %. A minute's delay recovered by re-planning costs 4.4 kWh, a slip override at 10 Hz keeps a train climbing a wet rail that otherwise stalls it at 7.10 km, and a timetable left 38.5 s short of the fastest possible run is reported impossible rather than planned. The model is a point mass with a lumped wheelset and every constant is a teaching value. A fuller model would add a train of several masses, a substation network that decides the receptivity, an energy price that varies in time, a drive-level slip controller, and a plan over several stations. The simulator shows that a timetable's slack is an energy budget whose best use depends on the line.

**KEY POINTS**

- A timetable's spare seconds are an energy budget, and the best way to spend them depends on where the limits of the line fall.
- From A to B, flat out costs 28.2 kWh, one coast point 19.5 kWh and a dynamic program 16.4 kWh, and the last two arrive a few seconds early.
- Pontryagin's principle limits an energy-optimal run to full power, speed holding, coasting and full braking, and the price of time sets the holding speed.
- Feedback finishes the plan: integral action stops a heavier train at the mark, a slip override keeps it climbing a wet rail, and re-planning recovers a minute for 4.4 kWh.

**Where else this applies.** The anti-lock brake of Volume 7, Chapter 4 regulates the slip of a braking tire at the same kind of peak that the slip override here hunts about.

## References


[1] A. Sciarretta, G. De Nunzio, and L. L. Ojeda, "Optimal ecodriving control: Energy-efficient driving of road vehicles as an optimal control problem," *IEEE Control Systems*, vol. 35, no. 5, pp. 71–90, 2015.

[2] R. E. Bellman, *Dynamic Programming*. Princeton, NJ: Princeton University Press, 1957, reprinted 2010.

[3] P. Howlett, "The optimal control of a train," *Annals of Operations Research*, vol. 98, no. 1–4, pp. 65–87, 2000.

[4] A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou, "The key principles of optimal train control—Part 1: Formulation of the model, strategies of optimal type, evolutionary lines, location of optimal switching points," *Transportation Research Part B: Methodological*, vol. 94, pp. 482–508, 2016.

[5] M. Khodaparastan, A. A. Mohamed, and W. Brandauer, "Recuperation of regenerative braking energy in electric rail transit systems," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 2831–2847, 2019.

[6] P. G. Howlett and P. J. Pudney, *Energy-Efficient Train Control*, ser. Advances in Industrial Control. London: Springer, 1995.

[7] A. Albrecht, P. Howlett, P. Pudney, X. Vu, and P. Zhou, "The key principles of optimal train control—Part 2: Existence of an optimal strategy, the local energy minimization principle, uniqueness, computational techniques," *Transportation Research Part B: Methodological*, vol. 94, pp. 509–538, 2016.

[8] R. Franke, P. Terwiesch, and M. Meyer, "An algorithm for the optimal control of the driving of trains," in *Proceedings of the 39th IEEE Conference on Decision and Control*, vol. 3, Sydney, NSW, Australia, 2000, pp. 2123–2128.

[9] H. Ko, T. Koseki, and M. Miyatake, "Numerical study on dynamic programming applied to optimization of running profile of a train," *IEEJ Transactions on Industry Applications*, vol. 125, no. 12, pp. 1084–1092, 2005.

[10] J. Yin, T. Tang, L. Yang, J. Xun, Y. Huang, and Z. Gao, "Research and development of automatic train operation for railway transportation systems: A survey," *Transportation Research Part C: Emerging Technologies*, vol. 85, pp. 548–572, 2017.

[11] I. A. Asnis, A. V. Dmitruk, and N. P. Osmolovskii, "Solution of the problem of the energetically optimal control of the motion of a train by the maximum principle," *USSR Computational Mathematics and Mathematical Physics*, vol. 25, no. 6, pp. 37–44, 1985.

[12] K. Ichikawa, "Application of optimization theory for bounded state variable problems to the operation of train," *Bulletin of JSME*, vol. 11, no. 47, pp. 857–865, 1968.

[13] E. Khmelnitsky, "On an optimal control problem of train operation," *IEEE Transactions on Automatic Control*, vol. 45, no. 7, pp. 1257–1266, 2000.

[14] R. Liu and I. M. Golovitcher, "Energy-efficient operation of rail vehicles," *Transportation Research Part A: Policy and Practice*, vol. 37, no. 10, pp. 917–932, 2003.

[15] G. M. Scheepmaker, R. M. P. Goverde, and L. G. Kroon, "Review of energy-efficient train control and timetabling," *European Journal of Operational Research*, vol. 257, no. 2, pp. 355–376, 2017.

[16] A. González-Gil, R. Palacin, and P. Batty, "Sustainable urban rail systems: Strategies and technologies for optimal management of regenerative braking energy," *Energy Conversion and Management*, vol. 75, pp. 374–388, 2013.

[17] O. Polach, "Creep forces in simulations of traction vehicles running on adhesion limit," *Wear*, vol. 258, no. 7–8, pp. 992–1000, 2005.

[18] B. P. Rochard and F. Schmid, "A review of methods to measure and calculate train resistances," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 214, no. 4, pp. 185–199, 2000.

[19] M. Malvezzi, G. Vettori, B. Allotta, L. Pugi, A. Ridolfi, and A. Rindi, "A localization algorithm for railway vehicles based on sensor fusion between tachometers and inertial measurement units," *Proceedings of the Institution of Mechanical Engineers, Part F: Journal of Rail and Rapid Transit*, vol. 228, no. 4, pp. 431–448, 2014.

[20] K. Zhou, S. Song, A. Xue, K. You, and H. Wu, "Smart train operation algorithms based on expert knowledge and reinforcement learning," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 52, no. 2, pp. 716–727, 2022.

[21] U.S. Federal Railroad Administration, "Rules, standards, and instructions governing the installation, inspection, maintenance, and repair of signal and train control systems, devices, and appliances," Code of Federal Regulations, Title 49, Part 236 (October 1, 2024 edition), U.S. Government Publishing Office, 2024. [Online]. Available: https://www.govinfo.gov/content/pkg/CFR-2024-title49-vol4/pdf/CFR-2024-title49-vol4-part236.pdf

[22] M. Hauck, P. Schmidt, A. Kobelski, and S. Streif, "A map-based model predictive control approach for train operation," in *2023 European Control Conference (ECC)*, Bucharest, Romania, 2023, pp. 1–6.

## 12.10 Exercises

**12.1** Concept check · 2 min

The dynamic program of this chapter may choose among eleven moves at every stage, including traction at three quarters, a half and a quarter. Yet every plan it returns is made of full power, speed holding, coasting and braking. Why?

(a) The Hamiltonian is linear in the force, so the best force sits at a bound or at zero except on the arc that holds a speed

(b) The grid of 1 km/h is too coarse to resolve a part force, so the program rounds each one to a full regime

(c) The tractive-effort curve is flat below the base speed, so a part force costs the same energy as full power there

(d) The drive's lag of 1.5 s filters out a part force, so the tracking loop can only follow the four regimes

**12.2** Concept check · 1 min

At the 271 s budget from A to B, the one-coast-point plan holds 98 km/h and plans 18.76 kWh, while the dynamic program barred from coasting holds about 78 km/h and plans 16.32 kWh. On the level 5 km test line the no-coasting plan loses by 7.2 to 8.8 %. What reverses the order on the A to B run?

(a) The dynamic program's grid is finer than the bisection that places the single coast point

(b) The fast plan must brake into the 60 km/h curve at 1.5 km and throws away speed it paid for

(c) Regenerative braking returns a larger share of the energy at low speed than at high speed

(d) Davis resistance is lower on the A to B run because that run has no grade

**12.3** Concept check · 2 min

In Scenario 6, 40 t more than the declared 190 t boards at the first second. With $K_c = 10$ and no integral action the train stops 0.56 m past the B mark, outside the 0.5 m accuracy. With $\tau_I = 20$ s added it stops 0.11 m past. Why does the gain alone not suffice?

(a) The tachometer reads 0.3 % high between balises, and only an integral can correct a scale error

(b) The drive's lag of 1.5 s delays the brake, and only an integral can anticipate a lag

(c) The stopping feedforward brakes for 190 t, and a proportional loop needs a standing error to supply the missing force

(d) The gain of 10 makes the loop oscillate about the braking curve, and the integral adds the damping that stops it

**12.4** Theory and calculation · 6 min

Use Table 12.2 with $M = 190\,\text{t}$ and $g = 9.81\,\text{m/s}^2$. Davis resistance is $R(v) = aM + bMv + cv^2$ with $v$ in m/s, and the motors give at most $\min(F_t, P_t/v)$.

(a) Compute $R$ at 60 and 100 km/h, and the deceleration of a train coasting on the level at 100 km/h, with the inertia $\rho M$. *Checkpoint:* 4.77 kN, 8.16 kN, $0.040\,\text{m/s}^2$.

(b) Compute the force that holds 98 km/h up the +2.0 % climb and the most the motors give at that speed. *Checkpoint:* 45.2 kN of 102.9 kN.

(c) Compute the deceleration of a train coasting at 80 km/h up the climb and its acceleration coasting down the −2.0 % descent. Why does case 4 coast before the summit rather than after it? *Checkpoint:* $0.212\,\text{m/s}^2$ up, $0.151\,\text{m/s}^2$ down.

**12.5** Theory and calculation · 6 min

The creep curve of Equation (12.2) peaks at $\mu\varphi Mg$ with $\varphi = 0.5$, $\mu_{dry} = 0.25$, $\mu_{wet} = 0.08$, $w_{pk} = 0.6\,\text{m/s}$ and $m_r = 0.05M$.

(a) Compute the largest force the rail transmits to the motored wheelsets, dry and wet. *Checkpoint:* 233.0 kN and 74.6 kN.

(b) Compute the base speed $P_t/F_t$ and the motor force available at 60 km/h, where case 7 accelerates out of the restriction. What happens on the wet rail if nothing limits the motors? *Checkpoint:* 45.8 km/h; 168 kN, over twice the wet peak.

(c) Compute the creep time constant $m_r/(2\mu\varphi Mg/w_{pk})$ on each rail, and how many of the wet ones pass in one 0.1 s control period. *Checkpoint:* 12.2 ms and 38.2 ms; about 2.6.

**12.6** Theory and calculation · 7 min

On the singular arc of a level line the price of time sets the holding speed through $\lambda\eta_t = V^2R'(V)$, with $R'(V) = bM + 2cV$ and $\eta_t = 0.85$.

(a) Compute $\lambda$ for holding 60, 80 and 98 km/h. *Checkpoint:* 77.8, 177.0 and 317.9 kW.

(b) Between budgets of 268 and 275 s from A to B the dynamic program's planned energy falls by 0.241 kWh per second of budget. Express that in kilowatts and compare it with the converged price of 925 kW at 271 s. *Checkpoint:* 867 kW.

(c) In case 8 the lost minute cost 4.4 kWh driven. The plans log $\lambda = 165\,\text{kW}$ at the 396 s budget and 444 kW at 336 s. What is the average price of the minute, and why does it lie between the two? *Checkpoint:* About 264 kW.

**12.7** App challenge · 8 min

Load Scenario 3, *3 · The coast point: one decision, and what it saves*, at the default seed and the shipped speed of 10 (each run to B takes about half a minute).

(a) Before Run, read the coast point on the *Timetable and plan* card. Run to B and read *Punctuality*, *Energy per km* (its kWh), *Drawn · returned* and *Regenerated share*. *Checkpoint:* 2.45 km; B −3.1 s, 19.5 kWh, 37.1 and 17.6 kWh, 76 %.

(b) Choose *Flat out* as the Planner, press Reset and run to B. Read the same tiles, and compute the saving of the coast point in net energy and in energy drawn. *Checkpoint:* B −21.0 s, 28.2 kWh, 58.9 and 30.7 kWh, 75 %. 31 % net, 37 % drawn.

(c) Load Scenario 5, *5 · The dynamic program against the analytic plan*, and run it to B. Read *Punctuality* and the energy, and say where on the run the program saves over the coast point. *Checkpoint:* B −1.7 s, 16.4 kWh, about 16 % less, before the 60 km/h curve.

**12.8** App challenge · 8 min

Load Scenario 8, *8 · Running late: re-planning, and the energy a minute costs*, at the default seed. Choose a speed of 30 (the run to C takes about 16 s). The doors at B are held 60 s and *Re-plan when late* is Off.

(a) Run to C. Read *Punctuality* and the energy, and export the CSV. In any row with `phase` RUN, read `plan_T_s` and `plan_lambda_kW`. *Checkpoint:* C +60.8 s, 26.1 kWh; 396 s and 165.3 kW.

(b) Turn *Re-plan when late* On, press Reset and repeat. *Checkpoint:* C −0.6 s, 30.5 kWh; 336 s and 443.9 kW.

(c) What did the minute cost, and why does the price $\lambda$ of the new plan rise? *Checkpoint:* 4.4 kWh, 17 %.

**12.9** App challenge · 9 min

Load the challenge, *Challenge · the full route: wet rail, a restriction, a late departure*, at the default seed and choose a speed of 30 (the route takes about 24 s).

(a) Run the shipped settings. The run pauses itself when the attempt cannot finish. Read the *Challenge target* tile and *Largest slip*. *Checkpoint:* Not met, B −0.6 s −0.07 m, stalled at 7.10 km; slip 34.63 m/s.

(b) Press *Reveal suggested tuning*, then Reset, and run to C. Read the tile, *Largest slip* and *Re-plans*. *Checkpoint:* Met, B −1.7 s −0.01 m, C −1.4 s −0.04 m, 51.8 of 55.1 kWh; 1.36 m/s; 1.

(c) Set *Wet-rail report* back to None, press Reset and run to C. Explain why one setting fails both the clock and the energy. *Checkpoint:* C +15.1 s, 57.9 of 55.1 kWh, slip 2.64 m/s, *CANNOT MAKE IT* on the plan card.

**12.10** Application area · 4 min

The meter of Equation (12.4) credits regenerated power only in the share $\kappa$ that the line accepts, and the line has no substations or other trains in the model except through $\kappa$.

(a) Flat out from A to B draws 58.9 kWh and returns 30.7 kWh at $\kappa = 1$, and the four-regime plan draws 37.1 kWh. What does each meter read with *Line receptivity* at 0 %, and what is the coast point's saving then? Check one in the app. *Checkpoint:* 58.9 kWh and 37.1 kWh; 37 %.

(b) What on a real line decides $\kappa$, and what does that mean for the value of coasting? *Checkpoint:* Another train drawing nearby, or equipment that takes the power. Coasting is worth more where the line cannot receive.

**12.11** Application area · 4 min

The slip override runs at the controller's 10 Hz. In case 7 with no wet-rail report its peak slip on the wet stretch ranges from 1.38 to 2.37 m/s over six seeds, above the declared 2 m/s at one, and at most 1.72 m/s with the report.

(a) Why does the peak vary from seed to seed, and why does the report lower it? *Checkpoint:* The wet entry falls at a random point in the control period; the report lowers the force demanded on entry.

(b) As the engineer specifying the traction drive for this line, what would you require, and how would you check it in service? *Checkpoint:* Slip control inside the drive at millisecond rates, and adhesion reports passed to the planner.

**12.12** Review the agent · 5 min

An AI assistant reviews the challenge and writes: *"Pontryagin's principle proves that the energy-optimal drive is full power, hold, coast and brake. The Four regimes planner builds exactly that drive, so it is the optimum. Keep the other suggested settings, choose Four regimes, and the run will meet*

*the clock and the energy target with room to spare.”* Find the flaw and say how you would check it in the app.

CHAPTER 13

# Summary and Outlook




**Open gaps** 1 models 2 baselines 3 margins 4 forecasts 5 coordination 6 infeasibility

*Notice that the gap of keeping the model right joins four plants, the balloon, the weld, the boiler and the data hall, more than any other gap. Illustration, not to scale.*

**IN PLAIN TERMS** **What Eleven Plans Taught**

**The main idea.** Each chapter of this book put a plan above a plant: a controller that looks ahead, or an optimizer that picks the best place to run. Each plan was then set against a simpler rule on the same day.

**The challenge.** The plan won where it knew something the rule did not, such as a forecast, an order or a price. It lost when its picture of the plant was wrong, and it used up every spare margin it was given.

**Why it matters.** So a plan is worth its cost only after a fair test against the simple rule, tuned once, with every saving taken apart.

Eleven chapters put a model and an optimizer above eleven plants and measured each against the simpler control that Table 1.3 names for it. The results ran from a weld that a step test from the wrong regime left with millimeters of scrap to a reservoir plan that kept the town under its limit through a water year in which the rule curve flooded it. This chapter sets those results side by side, draws a short guide for deciding when a plan is worth its model, looks at where predictive control and economic optimization are heading, reports how the volume itself was made with AI agents and what they got wrong, and lists the problems that remain open.

## 13.1 What the Chapters Showed

Tables 13.1 and 13.2 complete Table 1.3 with the measurements. The first holds the four chapters that use prediction to regulate, and the second the seven that use it to decide what the regulators hold or which whole decisions to take. Each row names the simpler control, where the predictive or optimizing layer paid, where it did not, and the numbers with the table or section that reports them. Figure 13.1 sets one like pair of numbers from each row side by side.

The columns of the two tables make five points.

**The plan paid for what it could see.** Every clear win in the tables traces to information a feedback loop does not have. The canal's plan sees an offtake in all three reaches at once. The heating network's plan sees the CHP's spare heat in the evening, the reservoir's tree sees seven futures, and the building's plan sees the curtailment notice. Take the information away and the win goes with it. The aeration plan with its preview off used 8622 kWh over two days at a worst mean of 2.607 g N/m$^3$, against 8596 kWh at 2.205 g N/m$^3$ for the supervisor (Table 10.7). The building's plan with its preview off did on the notified day what the supervisor does (Section 9.6.6). This is the closing argument of Section 1.4.13 measured on eleven plants.

**Table 13.1** Where Prediction Paid in Regulation, Chapters 2 to 5

| Chapter | Simpler control | Where the plan paid | Where it did not | Measured result |
|---|---|---|---|---|
| 2, hot-air balloon | Altitude PID. A three-deep cascade. | A smaller overshoot on a level-off. | The landing, where the cascade now touches down softer. The cruise, where the cascade swung less. A 120 s horizon, and a constant bias on an integrating output. | Touchdown 7.85, 1.27 and 1.39 m/s for PID, cascade and DMC (Table 2.6). Late swing 201.1, 0.8 and 6.8 m (Table 2.5). Hold offset 132.9 with the bias alone against 3.9 m with the drift ramp (Table 2.9). |
| 3, friction stir weld | Two PI loops, ITAE-tuned for setpoint or for load. | DMC built on a traverse step test, gentler on the spindle. | The same code built on a plunge step test. Both DMCs on an unmeasured thin patch. | Sound weld 98.3, 96.9 and 91.6 % for the traverse DMC, the load-tuned PI and the plunge DMC (Table 3.4). Thin patch 315 and 410 against 109 °C s (Section 3.6). |
| 4, drum boiler and turbine | Boiler-following, turbine-following and coordinated PI. | Offset-free MPC on sliding pressure, the only setting to pass the challenge. | The nominal MPC, offset under a fuel it cannot see. The tube, which holds the limit and gives up range. | Challenge passed on 11 of 11 seeds, every other setting on none. Final pressure gap 0.12 against 1.81 kg/cm$^2$ for coordinated PI (Table 4.5). Nominal offset 1.25 kg/cm$^2$ (Section 4.6). |
| 5, irrigation canal | Distant downstream PI with order feedforward. | An unannounced offtake, spread over three reaches by one plan. Plans per gate that exchange. | Published orders, which feedforward timed to the wave meets better. Plans per gate that do not exchange. | Farm closure IAE 406 against 1194 (m s), worst 5.84 against 11.46 cm. Published order 396 against 383, 2.88 against 2.15 cm (Table 5.8). Worst after a surprise 2.90, 3.10 and 6.68 cm (Table 5.9). |

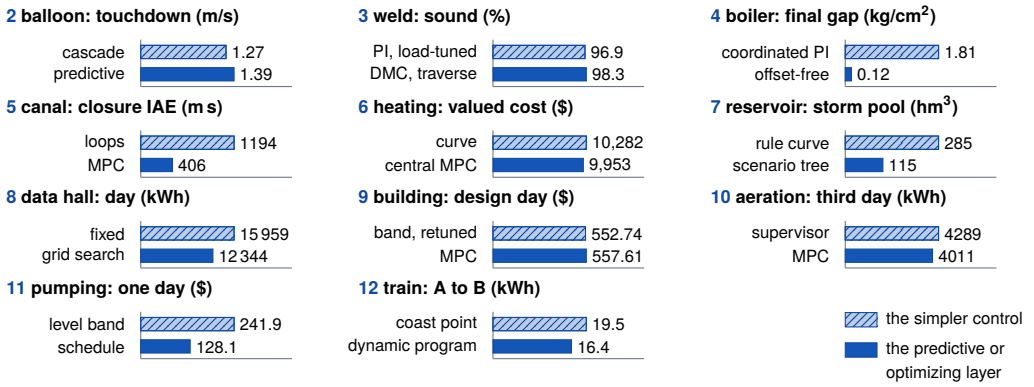



**Figure 13.1** One like pair of numbers from each row of Tables 13.1 and 13.2, digit for digit, each panel on its own scale. The hatched bar is the simpler control of the row, the solid bar the predictive or optimizing layer. Higher is better for the weld, lower for the other ten. Notice that the plan lost on the balloon's landing and to the building's band retuned once, and that the largest gaps come where a first optimizer replaced a fixed rule.

**Table 13.2** Where Optimization Paid in Targets and Whole Decisions, Chapters 6 to 12

| Chapter | Simpler control | Where the plan paid | Where it did not | Measured result |
|---|---|---|---|---|
| 6, district heating | Outdoor-compensated supply curve, a boost by hand. | Charging pipes and tank on the CHP's spare evening heat. | A boost timed by hand. A price that cannot settle when capacity binds. | Valued cost \$10,282 curve, \$10,484 boost, \$9,953 central MPC (Table 6.4). Cold snap passed by the curve on 0 and the distributed MPC on 5 of 7 weathers (Table 6.5). |
| 7, flood reservoir | Rule curve, with pre-release on the forecast. | A scenario tree on the seven-member ensemble. | A season's value table, which neither routes the river nor sees the ramp. | Storm pool 115 against 285 $hm^3$, 123.7 against 93.6 GWh, SDP town 530 $m^3/s$ (Table 7.4). Water year 7 of 7 against 0 of 7 seeds (Table 7.5). |
| 8, data hall cooling | Fixed setpoints. A grid-search supervisor. | The supervisor against fixed setpoints. Continuous moves over the grid. | A horizon over the supervisor. A lag model identified at one valve position. | Day 15 959 against 12 344 kWh (Table 8.5). MPC \$1409.40, grid \$1415.88, retuned grid \$1411.08, cheapest on none of four cases (Table 8.8). |
| 9, office building | Fixed setpoint. A priced band supervisor. | The band against the fixed setpoint. The plan on a curtailment day and on the challenge's peak. | Against the supervisor retuned once, on the design day and the heat wave. A flat tariff. | Design day \$615 fixed, \$561.19 band, \$557.61 MPC, \$552.74 band retuned (Tables 9.4, 9.5). Comfort 100 against 97.8 % with notice. Peak 204.4 against 234.8 kW. |
| 10, aeration basin | Fixed oxygen. An ammonia PI supervisor. | A day-ahead plan of the eight-hour mean on the typical load curve. | The same plan with no preview. Nitrogen discharged. The graded day. | Third day 4011 against 4289 kWh (Table 10.6). Nitrogen 35.9 against 29.0 kg. No preview 8622 kWh at a worst mean of 2.607 (Table 10.7). |
| 11, water network | Level-band pump rule with fixed valves. | An hourly integer schedule with modulated valves. | The schedule with fixed valves. Two pumps lost on a hot day. | \$128.1 against \$241.9 on one day (Section 11.6). Challenge \$515.3 against \$571.6, 11 of 11 against 0 of 11 seeds (Table 11.4). |
| 12, electric train | Flat out. Four regimes with one coast point. | A dynamic program, with re-planning and the slip override. | A timetable shorter than the fastest run. Any element of the stack removed. | 16.4, 19.5 and 28.2 kWh for DP, coast point and flat out (Table 12.4). Challenge 51.8 kWh passes, 58.8 with four regimes fails (Table 12.5). |

**The honest baseline is the simpler control retuned once.** In the cruise the balloon's cascade swung 0.8 m against the predictive controller's 6.8 m, with less burner travel. On the design day the building's supervisor, retuned once, closed at \$552.74 against the plan's \$557.61, and it won the heat wave as well. The data hall's plan was the cheapest on none of four cases, and a grid sixteen times denser matched or beat it on every case but the job step (Section 8.6.8). The canal's loops with feedforward timed to the wave beat the plan on published orders. A comparison against the shipped tuning alone would have credited the plan with all of these.

**The model decided more than the optimizer.** The weld's two predictive controllers are the same code with three different numbers in the step response, and the one identified in the plunge finished the weld 91.6 % sound with 18.2 mm of flash. The balloon's linear step response over-predicts a large move by 29 % at 90 s, so a 120 s horizon cycled at 214.4 m whatever the move suppression (Table 2.8). The boiler's nominal MPC held fuel back against a pressure rise that never came, 1.25 kg/cm$^2$ below its setpoint, until the estimator carried a fuel bias. The data hall's lag model, identified at a 27 % valve, let its valve saturate for 5.0 min in the hot-weather case, where the grid search's did not (Table 8.9). In every case the program was solved correctly. The prediction was wrong.

**Most of the money was in the target, not the trajectory.** Against a fixed setpoint the data hall's grid search saved 22.7 % of a day's cooling energy, and the building's priced band took the design day from \$615 to \$561. The hourly pump schedule spent \$128.1 where the level band spent \$241.9, and the dynamic program ran the train on 16.4 kWh where flat out took 28.2 kWh. The dynamic plan above an optimizing supervisor added far less. It saved \$6.48 on the data hall's day and \$3.58 on the building's design day, and 6.5 % of the aeration energy where the plan could schedule against the permit's own eight-hour mean. The first optimizer above a simple rule is where the large savings sat.

**An optimizer spends every margin it is given.** The aeration plan rode its ceiling and exceeded 2.20 g N/m$^3$ on all 30 seed-days by at most 0.022 g N/m$^3$, and it discharged 24 % more nitrogen than the supervisor, because nothing in its objective priced the nitrogen below the ceiling. The heating network's plan passed the cold snap with its stores 14.9 MWh lower at the end. A continuous search of the data hall's cost, free of the plan's margins, was the cheapest on both days and rode the water budget for 180.4 and 305.5 min. Each saving in the tables has to be read beside what was spent to earn it.

## 13.2 Choosing a Method

The chapters suggest an order of questions before a model and an optimizer are put above a plant. Each names the chapter where its answer decided the result.

1. *What will the plan know that a loop does not?* A preview, a published order, a price, a forecast or a limit ahead. If the answer is nothing, retune the loop and add feedforward first. The weld's thin patch and the canal's published orders went to the simpler control (Chapter 3, Chapter 5).

2. *Is the value in where to operate or in how to get there?* When an active constraint moves with the weather or the tariff, a steady-state optimizer or a priced rule takes most of the saving (Section 1.4.7, Chapter 8, Chapter 9). A dynamic plan adds where storage lets energy be bought early or where the measure is itself a mean over hours (Chapter 6, Chapter 10).

3. *Is the model right in the regime the plan will act in?* Take the step test in that regime, keep the horizon inside the range where the linear response holds, and give an integrating output a drift term (Chapter 3, Chapter 2).

4. *What does the plan do about what it cannot measure?* A predictive controller has no integral term of its own. Without a disturbance model it keeps an offset (Section 1.4.5, Chapter 4).

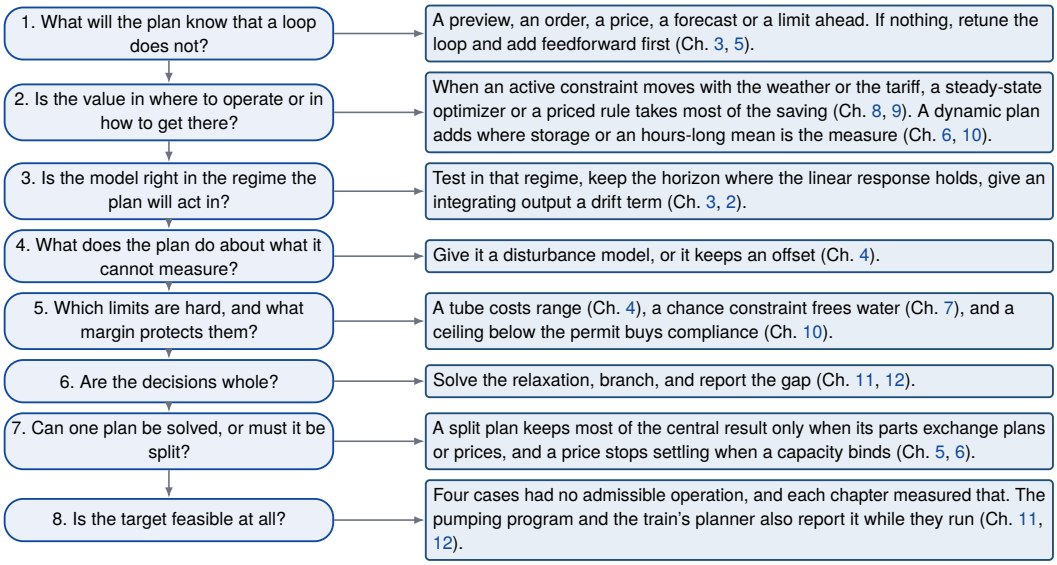



**Figure 13.2** The eight questions of this section in order, with the list's answer and the chapters where it decided the result. Notice that what the plan will know and whether its model holds are settled before any optimizer is chosen.

5. *Which limits are hard, and what margin protects them?* A tube costs range, 70.53 kg/cm$^2$ of 75 usable on the boiler. A chance constraint frees water, 10.4 GWh in the reservoir's spring storm. A ceiling set below the permit costs a little energy and buys compliance on every seed (Chapter 10).
6. *Are the decisions whole?* Pumps and regimes are on or off. Solve the relaxation, branch, and report the gap (Section 1.4.10, Chapter 11, Chapter 12).
7. *Can one plan be solved, or must it be split?* A split plan keeps most of the central result only when its parts exchange plans or prices, and a price stops settling when a capacity binds (Chapter 5, Chapter 6).
8. *Is the target feasible at all?* The boiler's nameplate, a flood with the smallest pool, two pumps lost and a short timetable had no admissible operation. Each chapter measured that, and the pumping program and the train's planner report it while they run.

The comparison that decides whether to keep the plan is then run against the simpler control retuned once, on the same seed, instruments and forecast, as Chapters 5, 8, 9 and 10 do. The aeration plan's saving survives that test at 4.9 % below the best supervisor found with the day known. The building's design-day saving does not. Figure 13.2 draws the eight questions in order.

## 13.3 Where Practice Is Heading

**Predictive control is established, and the work is upkeep.** In a 2018 survey of the IFAC Industry Committee, 85 % of the members who answered named model predictive control as a technology with high impact in the future, against 78 % for PID [1]. A review of industrial practice reports that after decades of application, practitioners now focus on ease of commissioning, monitoring and automated maintenance, that many industries need better usability of existing technology more than better algorithms, and that continuous performance monitoring and automated re-identification are being used toward adaptive MPC [2]. The weld of Chapter 3 shows why. A constant output bias absorbed a gain error of 1.7 to 1.9 as an offset that decayed over about a minute, which is how a badly

identified controller can stay in service unnoticed. In the editor's view the next decade of industrial MPC is about keeping models right, not about new formulations.

**The objective moves from tracking to cost.** Economic MPC, which optimizes an operating cost directly instead of tracking a setpoint chosen elsewhere, has a tutorial review of its methods [3]. For buildings, a review states that MPC has been shown to reduce energy use notably while practical applications remain in early stages, and it calls for standardized performance assessment and methods for comparing building control algorithms [4]. The building of Chapter 9 is a small case of that call. Its plan beat the shipped supervisor and lost to the same supervisor retuned once.

**Uncertainty, structure and schedules enter the plan.** Stochastic MPC, which plans against a distribution of futures, has an overview of its algorithms and of the theoretical challenges still open [5]. Distributed MPC has a tutorial review of its architectures and future directions [6], and the integration of production scheduling with process control has a systematic review [7]. The reservoir's scenario tree, the canal's exchanged plans, the heating network's price and the pump schedule are small instances of each. In the editor's view these will reach practice first where a plant already publishes a forecast, an order book or a tariff, as these four do.

**Computation stops being the limit, and learning enters.** Custom methods that exploit the structure of the MPC problem compute the control action on the order of 100 times faster than a generic optimizer, solving a quadratic program with 450 variables in about 5 ms [8]. An open-source operator-splitting solver for quadratic programs detects infeasible problems, supports warm starts and is suited to embedded use [9]. In this volume the slowest boiler solve was 4.8 ms and the building's plan 15 ms. Learning-based MPC, reviewed in three categories, learns the prediction model from recorded data, infers the cost and constraints that give the best closed-loop performance, or uses MPC to add constraint satisfaction to a learned controller [10]. In the editor's view the model, not the solver, will remain the scarce resource, and learning will be judged by whether it keeps the model right in the regime where the plan acts.

**Forecasts enter the operating rule.** A viability assessment of forecast-informed reservoir operations reports that approved major deviations let the operating agency apply an ensemble forecast model in real-world operations in the water years 2019 and 2020. The model adds a variable buffer pool to the guide curve and recommends flood releases from a 15-day streamflow ensemble forecast. In 2020, the third driest year of a 127-year record, it gave 19 percent more water in storage by the end of winter. The assessment recommends writing the concept into the reservoir's control manual and moving to a further phase only when a defined set of forecast skill metrics improves [11]. The scenario tree of Chapter 7 plans on a seven-member ensemble in the same spirit, and the canal of Chapter 5 on published orders, the older form of the same idea.

**Buildings demonstrate the plan but rarely adopt it.** A technical report on grid-interactive buildings finds supervisory building control typically implemented as rules, such as setpoint schedules and economizer resets, with many tuning parameters set building by building. It reports early experiments at building sites in which predictive control saved up to 17 to 65 percent of energy against the installed control sequences, and finds it not widely adopted despite several field demonstrations, held back by legacy hardware, the cost of accurate models and operators' acceptance. It also calls for work on how much thermal flexibility can be used without predictive optimization at all [12]. Chapter 9 measured that question on one design day, where a band supervisor retuned once beat the plan.

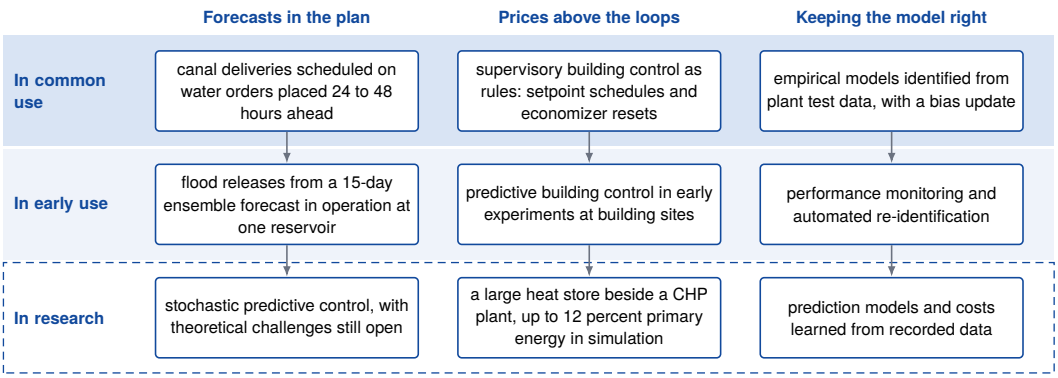



**Figure 13.3** Developments in three threads. The top lane holds what the cited source calls typical or the way most industrial controllers work, the middle lane what it reports in operation at one site, in early experiments at sites or being used in practice, and the bottom lane what it reports from simulation or names as a need, in the editor's reading. Sources by thread: forecasts in the plan [5, 11, 14], prices above the loops [12, 15], keeping the model right [2, 10, 16]. No date is drawn, because the sources give none for these steps.

**Networks split the plan and keep it working.** A paper on automated irrigation reports pilot projects in which sensors, actuators, a communication network and identified models raised the distribution efficiency to better than 90 percent, against 70 percent under manual operation, and met better than 90 percent of water orders on demand. In its simulations of larger canals, decentralized control with feedforward could in most scenarios not be told apart from centralized control, which the paper notes cannot be implemented [13]. The canal of Chapter 5 found the same order on three reaches. After a surprise, plans per gate that exchange held the worst error to 3.10 cm against the central plan's 2.90 cm, and plans that did not exchange to 6.68 cm. Figure 13.3 sets these developments and those above in three threads.

> ⧖ HISTORY **1999 to 2003: Prediction Spreads Beyond the Refinery**
>
> The history boxes of Chapter 1 end in 1986, when limits became rows of a quadratic program. A survey published in 2003, from a snapshot of vendors' figures taken in mid-1999, estimated 4542 linear applications. Refining held 67 percent of those classified, utilities 14 and air and gas 10, and the count had roughly doubled from 1995 to 1999. Industry's research needs were mostly control engineering, not algorithms: which variables to control, when a model is good enough to stop the plant test, why a controller performs poorly, and how to estimate the benefits of a control system [16]. The plants of this volume sit in the small rows of that table, and in the editor's reading its questions are still theirs.

## 13.4 Engineering With AI Agents

An AI agent is a language model that works toward a goal in a loop. It plans, calls tools such as a compiler, a simulator or a browser, reads the results and continues [17]. In the editor's view, agents can now carry out much of the work that used to fill the hours of a control project. They derive a model from balances or fit one to plant data. They write the simulator and its tests, code the controller, tune it against stated criteria, script commissioning checks, write the documentation and watch running loops for degradation. The apps and chapters of this series were produced that way under the editor's direction, and the account that follows this section reports what the agents got wrong in this volume and how it was caught.

In the editor's view, this changes where an engineer's value lies. When a model, a controller and a report can each be produced in minutes, producing them is no longer the scarce skill. The work moves to five things.

**Quality control.** An agent's output arrives as a claim: this model fits, this tuning settles in four minutes, this code implements the equation. The engineer checks claims against the plant or the simulator. A reported number is measured again on a fresh run at the stated seed. The code is read for what it clamps, what it initializes and what it reads, because a display that shows the controller's model in place of the plant looks correct until the two differ. A tuning is tried on a disturbance the agent did not test. The most useful checks are independent of the author, made by a second agent or a person who has not seen the first answer, and they are checks that can fail.

**Supervision.** The engineer sets the scope of each task and decides what an agent may touch. Reading a historian and proposing a tuning is one level of access. Writing a setpoint, downloading a configuration or changing an interlock is another, and it passes through the plant's management of change with a named approver. Supervision also fixes the order of work: a baseline before a change, one change at a time and a comparison that shows what moved.

**Creativity.** An agent answers the question it is given. Choosing the question remains the engineer's task: which loop is worth improving, whether a disturbance calls for feedforward, a cascade or a change to the process, which step test the operation can tolerate, and which comparison would show that a predictive controller earns its place over a retuned PID loop. The experiment that separates two explanations is designed by someone who can imagine both.

**Process insight.** A model contains what its author put in. The equipment does more. A valve sticks, the dead time of an analyzer grows when its sample flow drops, a thermowell lags, a heat exchanger fouls and an operator intervenes at the same hour each night. The engineer who knows the unit recognizes a result that is too good, asks what the model left out and knows which measurement to distrust first.

**Accountability.** Responsibility does not transfer to a tool. Functional safety standards assign the activities of the safety lifecycle to persons and organizations of stated competence, and they require an assessment that includes someone independent of the design team [18]. An agent may draft the hazard analysis or the test procedure. A person owns the decision to accept it, to start up and to keep running.

These five are the professional form of the habit introduced in Section 1.7: check before trusting, and know enough to check.

### 13.4.1 How This Volume Was Made

Each of the eleven apps was built and tested by agents. An agent then wrote the paper on it, keeping verification notes that trace every number to a run, and a referee report was written on each paper. At least five of the eleven reports are self-reviews by the agent that wrote the paper. Five apps passed usability, design and technical-accuracy reviews before their papers, and four papers had a further referee pass on predictive material added later. Defects went into a log for each app with a severity, a cause, a suggested fix and a status. For this book one agent per chapter reran the numbers its exercises use. The editor set the scope of each round and the rules on what each agent could change, and statements that runs contradicted were left for the editor to decide.

**The defect logs.** The eleven logs written beside the papers hold 87 entries with a severity. Of these, 24 are wrong behavior or output of an app, 46 are documentation, 13 are cosmetic and 4 are other kinds, such as a test or a measurement that is not a defect. The six apps without a review before the paper account for 70 entries, and the five with one account for 17, because their reviews had already logged and fixed findings in each app's own record. Five cases show the errors that matter for a volume about plans and their payoff.

**Savings credited to the wrong cause.** The data hall's draft risked crediting the plan's horizon with its saving. The referee pass on the predictive layer had the section state that the plan is the cheapest on none of four cases and that the horizon's only content is the lag of its own moves. The building's Conclusion said half of the precooling saving from \$615 to \$561 was demand charge. It was \$22 of \$54, and the half-and-half split belongs to the \$24 between the floating band and precooling. The same paper split a price ratio of 7.5 as about three from the tariff and the rest from the chiller, a sentence inherited from the app's guide. The tariff gives 4.3 and the chiller 1.8. The canal's headline on an unannounced offtake could be read as a win in total error, and a per-reach copy of the plan scored 388 against the plan's 406 (Table 5.8). The text now credits the joint plan only with the worst error. Each was caught by taking a saving apart into its sources.

**Numbers taken from the app instead of measured.** The weld paper's first draft quoted the first-order-plus-dead-time constants stored in the app as the plant's. Differencing two runs, as the app says it does, gave a plunge time constant of 1.33 s and no dead time where the stored values said 1.92 s and 2.10 s. The chapter's own least-squares fit now reads 1.17 s and no dead time (Table 3.3). The data hall's draft said the binding constraint changes through the diurnal day, following the guide. Rerunning the search found none binding at any of 145 samples. The balloon paper said a higher gain diverges and puts the balloon in a field. Measured, the swing nearly triples and the loop keeps hunting without reaching the ground. All three came from the app's own prose, and the referee of each caught it by measuring.

**A horizon result that could have been the terminal weight.** The balloon's 214.4 m cycle at a 120 s horizon was first argued from a sweep of move suppression. The referee pointed out that the terminal weight sat on the last points of the horizon, exactly where the linear model is worst. With the terminal weight removed the swing at 120 s still ran from 142.8 to 293.7 m, and at 52 s a fortyfold

terminal weight moved it only from 1.6 to 4.4 m. The conclusion of Table 2.8 now rests on two factors, not one.

**A cost booked twice.** The aeration draft claimed that below the blower's surge line a higher oxygen setpoint costs less power. The referee read the source and found the vented air billed twice, hardest at low demand, which reversed the sign. The app was fixed and the claim withdrawn. An optimizer given that cost would have been rewarded for raising the oxygen at low load. The fixed-setpoint frontier did not move, because none of those days vents air.

**Controllers that read the plant's truth.** Five log entries in four apps record a controller or an operator aid that read values a real plant would not give it. The balloon's step test integrated the run's own parameters, air and true state, so that 17 of 34 parameters scaled by 1.25 before Run moved the identified response. The heating network's planners and observers were built on the run's hidden building data. The canal's gate inversion, lead and turnout readings used edited values, and the data hall's start-up registers and a manual helper read plant states. Only the balloon's and the heating network's fixes moved a control result. The balloon's is the one that matters for this volume. Before the fix the predictive controller landed at 0.43 m/s and after it at 0.71 m/s. Part of its margin over the cascade had come from knowing the plant it was landing.

**The book round.** The eleven chapter records list 197 rows of numeric verification, each a stated value or group of values set beside its reproduction. Nine records report that every rerun number reproduced to its printed digit or its rounding. The aeration record found two values that did not follow from the model and a flow of 4.93 printed as 5.0. All three are now corrected. The pumping record found the last digits of the challenge row moved by cents. The book round's chapter passes logged nine defect entries for this volume. The reservoir's challenge could be passed by raising the channel limit, with the town at 980 $m^3$/s, and the app now voids the verdict. The balloon's descent envelope never bound near the ground at the shipped slope, so the flare setting the chapter credits did not set the touchdown. After a repair was approved the landing guidance uses the profile directly, and the suggested landings became 1.27 and 1.39 m/s for the cascade and the predictive controller (Table 2.6). Both still stand up, and the predictive controller's earlier margin has reversed. The agent writing Chapter 1 first priced a soft limit from the steady multiplier, 0.04, and the linear program showed the last hour's multiplier to be the full price, 0.20.

Every error above was caught by a check that could fail: a saving taken apart, a number measured instead of copied, a second factor varied, a source line read against the balance, a hidden value moved to see what else moved. Most of them flattered the plan. The headlines that survived are the ones in Tables 13.1 and 13.2. Four working rules follow. Compare every plan against the simpler control retuned once. Take every saving apart into its sources before it is printed. Read what the plan reads and what it spends. Let a person decide what changes.

## 13.5 Gaps in Predictive Control and Economic Optimization

Six gaps belong to this volume's topic, and the gaps common to agent-era practice follow them in Section 13.5.1. In the editor's view none is closed. Each item gives the chapters' evidence, then the state of the literature with its sources. Why each gap is hard and what progress would look like are the editor's view. Figure 13.4 places the gaps by the chapters that met them and by the maturity of the method available, and the same numbers mark the plants in the picture that opens this chapter.

1. *Keeping the prediction model right in service.* The weld's plunge model, the data hall's lags identified at one valve position and the balloon's step response beyond its linear range each failed without any error in the optimizer, and the boiler's nominal model held an offset under a fuel it could not see until the estimator carried a bias. A routine is needed that detects from closed-loop

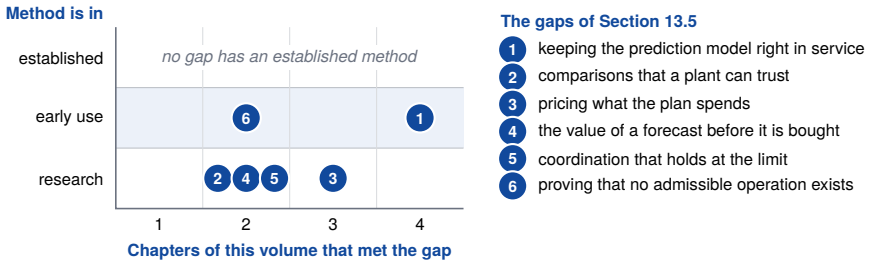



**Figure 13.4** The six gaps of this section. Across is the number of chapters that each item names as evidence: four for gap 1 (Chapters 3, 8, 2 and 4), two for gap 2 (9 and 5), three for gap 3 (10, 6 and 8), two for gap 4 (10 and 7), two for gap 5 (6 and 5) and two for gap 6 (7 and 11). Up is the maturity of the method available, the editor's reading of the sources cited in each item and in the Open challenge box. Notice that the gap met by the most chapters, keeping the model right, is also the one closest to practice.

records which part of the model has drifted, re-identifies it in the regime where the plan acts, and does so without a step test the plant cannot tolerate. The Open challenge box at the end of this section treats this gap.

2. *Comparisons that a plant can trust.* The building's plan won against the shipped supervisor and lost to it retuned once, and the canal's best fixed loop setting could be chosen only after the run. A report on building controls cites savings measured against the installed control sequences and calls for work on how much can be had without predictive optimization [12], and the 2003 survey lists how to estimate the benefits of a control system among industry's open questions [16]. The gap is hard because the baseline a plant runs today is rarely the best it could run, and retuning it costs engineering that nobody budgets. Practice needs a standard protocol for an MPC business case, with the baseline retuned by a stated rule, the same information and weather, several seeds, and the saving taken apart into its sources.

3. *Pricing what the plan spends.* The aeration plan saved energy and discharged more nitrogen, the heating network ended with emptier stores, and a margin-free search rode the data hall's water budget for hours. A review of thermal network control warns that more frequent temperature cycling may speed fatigue in steel pipes, and finds a simple, accurate and concise way to state the flexibility available still lacking [15]. The gap is hard because the margins are owned by different people in different units, a permit, a pipe's life, a comfort band. An economic objective that prices the margins a plant cares about, and says which of them a saving consumed, is not yet routine.

4. *The value of a forecast before it is bought.* The aeration plan needed the typical load curve, and the reservoir's chance constraint let the wettest member go with nothing to say about the year in which that member comes. The forecast-informed reservoir assessment ties each further phase of operation to a set of forecast skill metrics still to be defined, and lists their development among its future research [11]. The gap is hard because a forecast's worth depends on the plant's own storage and limits, not on its skill alone. A method that states in advance what a given forecast is worth, and what a wrong one costs, would decide where a plan belongs.

5. *Coordination that holds at the limit.* The heating network's price did not settle in any of the 55 steps after the trip, and the canal's exchanged plans form a Nash point that costs 1.13 times the central optimum on average (Section 5.4.4). The irrigation paper finds that a global controller must cope with varying information rates and delays across the network, and that substantial further research is necessary [13]. The gap is hard because a coordinating price carries no meaning

once no plan fits the shared capacity. Distributed schemes that keep converging when a shared capacity binds, and that say how far from the central plan they are, remain open.

6. *Proving that no admissible operation exists.* The reservoir's infeasible flood rests on a search, not a proof, and the pumping program could only flag its relaxed plan as unable to hold the floor. A study of pump scheduling finds heuristics that cannot guarantee global optimality often applied, and branch and bound slowed from 1 to 43 s by two loops in a test network of 24 time steps, and it asks for tailored solvers [19]. The gap is hard because the hydraulics and the on-off decisions make the program nonconvex. Certificates of infeasibility for nonlinear plants with integer decisions, delivered fast enough to warn an operator, are not available.

> ⚠ **OPEN CHALLENGE** **Keeping the Model Right Where the Plan Acts**
>
> Gap 1 runs through four chapters, each with a model that failed in service while the optimizer solved correctly. The 2003 survey found most industrial controllers built on empirical models identified from plant tests, and industry asking when a model is good enough to stop the test and how to find the source of a problem when a controller performs poorly [16]. A report on building controls names the development and calibration of models accurate and robust enough as a significant challenge to wider use [12]. In the editor's view the gap is hard because the plan's own moves mask the drift it should detect. Progress would be a monitor that names the drifting part of the model from closed-loop records.

### 13.5.1 Gaps Common to Agent-Era Practice

The final chapter of each volume lists gaps specific to its topic. Seven more are common to every topic in the series. In the editor's view none of them is closed.

1. *Verifying agent-written control code and models before they reach a plant.* Tests show that a controller behaves on the cases tested. For controllers learned from data, guarantees of stability and constraint satisfaction rest on assumptions about the model and its uncertainty, and open problems remain [20]. In the editor's view, code written by an agent has no comparable body of method yet. Practice needs a routine path from an agent's draft to code a plant can trust: independent derivation, tests of what the code reads and clamps, and trials with hardware in the loop.

2. *Trustworthy models from sparse or poor plant data.* Plant records are short, taken in closed loop, noisy and marked by unmeasured disturbances. A fitting routine returns a model from any record. What is missing is a dependable statement, delivered with the model, of where it can be trusted, and a method that says so when the data cannot identify a parameter.

3. *Safety cases and standards for agents in commissioning and operation.* Functional safety standards define a lifecycle of specified requirements, verification and assigned human roles [18]. In the editor's view, how an agent that writes configuration, runs commissioning checks or advises an operator is qualified, limited and audited under them is not yet settled.

4. *Security of agent access to control systems.* An agent with credentials to a control system is a new path into it. Industrial security standards state the requirements a control system must meet at each security level [21]. An agent that reads documents, messages or web pages can also be steered by text placed there. In the editor's view, least privilege, read-only access by default and human approval of every write are a starting point and not a solution.

5. *Benchmarks and reproducibility.* An agent asked twice may deliver two different controllers, and a result reported without its plant, seed and scoring cannot be checked. Control needs shared benchmark plants with fixed disturbances and published measures, on which agent-built and conventional designs are compared like for like. The apps of this series, with a default seed and an exported record of every run, are a start.

6. *Transfer from simulation to hardware.* A controller developed on a simulator inherits what the simulator omits. The robotics literature documents the gap between simulated and real performance and the methods used to narrow it, such as randomizing the simulated plant and adapting on the real one [22]. Process plants add slow dynamics, rare upsets and little freedom to experiment, so each transfer still needs a staged plan from the simulator to a bench rig to the unit.

7. *Keeping engineers' judgment sharp when agents do the routine work.* Automation removes the routine work through which skill is built and leaves the person the hard cases, an irony described for process operators in 1983 [23]. The same now applies to engineers. An engineer who has never tuned a loop by hand or closed a balance on paper is poorly placed to review an agent that has. Deliberate practice, of the kind the exercises in this book provide, has to be designed into education and into work.

**KEY POINTS**

- A plan paid for what it could see that a loop could not: a forecast, an order, a price or a limit ahead.
- Most of the saving sat in the first optimizer above a simple rule. A dynamic plan above that optimizer added much less.
- The model, not the solver, decided the result, and keeping it right in service is the open problem.
- An optimizer spends every margin it is given, so each saving is read beside what it spent and against the simpler control retuned once.

## References


[1] T. Samad, M. Bauer, S. Bortoff, S. Di Cairano, L. Fagiano, P. F. Odgaard, R. R. Rhinehart, R. Sánchez-Peña, A. Serbezov, F. Ankersen, P. Goupil, B. Grosman, M. Heertjes, I. Mareels, and R. Sosseh, "Industry engagement with control research: Perspective and messages," *Annual Reviews in Control*, vol. 49, pp. 1–14, 2020.

[2] M. G. Forbes, R. S. Patwardhan, H. Hamadah, and R. B. Gopaluni, "Model predictive control in industry: Challenges and opportunities," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 531–538, 2015.

[3] M. Ellis, H. Durand, and P. D. Christofides, "A tutorial review of economic model predictive control methods," *Journal of Process Control*, vol. 24, no. 8, pp. 1156–1178, 2014.

[4] J. Drgoňa, J. Arroyo, I. Cupeiro Figueroa, D. Blum, K. Arendt, D. Kim, E. P. Ollé, J. Oravec, M. Wetter, D. L. Vrabie, and L. Helsen, "All you need to know about model predictive control for buildings," *Annual Reviews in Control*, vol. 50, pp. 190–232, 2020.

[5] A. Mesbah, "Stochastic model predictive control: An overview and perspectives for future research," *IEEE Control Systems*, vol. 36, no. 6, pp. 30–44, 2016.

[6] P. D. Christofides, R. Scattolini, D. Muñoz de la Peña, and J. Liu, "Distributed model predictive control: A tutorial review and future research directions," *Computers & Chemical Engineering*, vol. 51, pp. 21–41, 2013.

[7] M. Baldea and I. Harjunkoski, "Integrated production scheduling and process control: A systematic review," *Computers & Chemical Engineering*, vol. 71, pp. 377–390, 2014.

[8] Y. Wang and S. Boyd, "Fast model predictive control using online optimization," *IEEE Transactions on Control Systems Technology*, vol. 18, no. 2, pp. 267–278, 2010.

[9] B. Stellato, G. Banjac, P. Goulart, A. Bemporad, and S. Boyd, "OSQP: An operator splitting solver for quadratic programs," *Mathematical Programming Computation*, vol. 12, no. 4, pp. 637–672, 2020.

[10] L. Hewing, K. P. Wabersich, M. Menner, and M. N. Zeilinger, "Learning-based model predictive control: Toward safe learning in control," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 3, no. 1, pp. 269–296, 2020.

[11] Lake Mendocino FIRO Steering Committee, "Lake Mendocino forecast informed reservoir operations final viability assessment," UC San Diego, Tech. Rep., Dec. 2020. [Online]. Available: https://cw3e.ucsd.edu/FIRO_docs/LakeMendocino_FIRO_FVA.pdf

[12] A. Roth and J. Reyna, "Grid-interactive efficient buildings technical report series: Whole-building controls, sensors, modeling, and analytics," U.S. Department of Energy, Washington, DC, Tech. Rep. NREL/TP-5500-75478, 2019.

[13] I. Mareels, E. Weyer, S. K. Ooi, M. Cantoni, Y. Li, and G. Nair, "Systems engineering for irrigation systems: Successes and challenges," in *IFAC Proceedings Volumes (16th IFAC World Congress)*, vol. 38, no. 1, 2005, pp. 1–16.

[14] U.S. Bureau of Reclamation, *Canal Systems Automation Manual, Volume 1*, 1st ed. Denver, CO: U.S. Department of the Interior, Bureau of Reclamation, 1991. [Online]. Available: https://www.usbr.gov/tsc/techreferences/mands/mands-pdfs/CanalSysAuto1.pdf

[15] A. Vandermeulen, B. van der Heijde, and L. Helsen, "Controlling district heating and cooling networks to unlock flexibility: A review," *Energy*, vol. 151, pp. 103–115, 2018.

[16] S. J. Qin and T. A. Badgwell, "A survey of industrial model predictive control technology," *Control Engineering Practice*, vol. 11, no. 7, pp. 733–764, 2003.

[17] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen, "A survey on large language model based autonomous agents," *Frontiers of Computer Science*, vol. 18, no. 6, p. 186345, 2024.

[18] International Electrotechnical Commission, "Functional safety – Safety instrumented systems for the process industry sector – Part 1: Framework, definitions, system, hardware and application programming requirements," IEC 61511-1:2016, Edition 2.0, 2016.

[19] R. Menke, E. Abraham, P. Parpas, and I. Stoianov, "Exploring optimal pump scheduling in water distribution networks with branch and bound methods," *Water Resources Management*, vol. 30, no. 14, pp. 5333–5349, 2016.

[20] L. Brunke, M. Greeff, A. W. Hall, Z. Yuan, S. Zhou, J. Panerati, and A. P. Schoellig, "Safe learning in robotics: From learning-based control to safe reinforcement learning," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, pp. 411–444, 2022.

[21] International Electrotechnical Commission, "Industrial communication networks – Network and system security – Part 3-3: System security requirements and security levels," IEC 62443-3-3:2013, Edition 1.0, 2013.

[22] W. Zhao, J. P. Queralta, and T. Westerlund, "Sim-to-real transfer in deep reinforcement learning for robotics: A survey," in *Proc. 2020 IEEE Symposium Series on Computational Intelligence (SSCI)*, 2020, pp. 737–744.

[23] L. Bainbridge, "Ironies of automation," *Automatica*, vol. 19, no. 6, pp. 775–779, 1983.

## 13.6 Exercises

**13.1** Discussion · 8 min

A vendor offers a day-ahead MPC for the office building of Chapter 9 and quotes the design day, \$557.61 against \$561.19 for the shipped supervisor. Using that chapter and Chapter 8, say what comparison you would ask for before buying, and what the two chapters suggest it would show.

**13.2** Discussion · 7 min

The aeration plan of Chapter 10 used 277 kWh a day less than the ammonia supervisor at nearly the same worst eight-hour mean, and discharged 24 % more nitrogen. Is that a better way to run the works? Say what you would change in the plan's objective or limits, and what each change would cost.

**13.3** Discussion · 8 min

An agent hands you a step-response MPC for a new plant. Its model came from one step test taken during start-up, its horizon is twice the settling time of the step, and its disturbance model is a constant output bias. Using Chapter 3, Chapter 2 and Chapter 4, list what you would check before the controller is switched on, and say which chapter shows the failure each check prevents.

**13.4** Discussion · 7 min

The heating network of Chapter 6 and the canal of Chapter 5 each split one plan among several agents. Compare how the two schemes behaved, say when a split plan is worth having, and propose what a split plan should do when its coordination fails.

**13.5** Discussion · 9 min

Section 13.4.1 lists errors the agents made while building this volume. Pick three and name the check that caught each. Then write a procedure of five to eight steps for an agent that will propose an

economic MPC for a running plant, stating what it may read, what it may write and what a person must approve.

**13.6** Discussion · 6 min

The reservoir of Chapter 7, the water network of Chapter 11 and the train of Chapter 12 each meet a case that no operation can satisfy. Discuss what an optimizing controller should tell the operator in such a case, and why a plan that quietly does its best is not enough.

# Answers to Selected Exercises

Short answers are given for the odd-numbered exercises. Numeric answers from an app are for its default seed and settings unless the exercise says otherwise.

### Chapter 1. Introduction

**1.1** (a) 2.952 K/kW. (b) 3.30 kW. (c) 1 kW.

**1.3** (a) $u = 48 - 0.8\,T$. (b) 48 °C, an offset of −2 K.

**1.5** (a). The largest multiplier sits at the horizon's last hour, where a kelvin saved is a full kilowatt-hour, $0.20.

**1.7** $0.244, $0.195 and $0.156. The earliest worth buying is 04:00.

**1.9** (d). Rounding drops the tank to 45.1 °C. No feasible schedule can cost less than the relaxation.

### Chapter 2. Hot-Air Balloon Landing Behind a Dead Time

**2.1** (a). The trim valve matches the fabric loss, so it freezes the envelope temperature and with it the climb rate the burn bought.

**2.3** (c). A cycle that no move suppression and no terminal weight removes is a model error, and the identified step response over-predicts the plant past about 52 s.

**2.5** (a) 2.972 MJ/K, 646 s. (b) 12.65 %, 8.30 g/s (29.9 kg/h). (c) +0.82 and −1.41 K/s, ratio 1.72.

**2.7** (a) 4.0 s, the transport delay alone. (b) 24.2 s and 33.4 s. (c) 39.6 m against $0.20 \times 210.3 = 42.1$ m: the linear model over-predicts.

**2.9** (a) 7.85 m/s, hard, 4.2 m/s across the ground: both clauses fail. (b) 1.39 m/s, 3.5 m/s, 202 s, 2.44 kg. (c) 1.27 m/s, 3.5 m/s, 229 s, 2.44 kg. Both land faster than the requested 0.70 m/s, so a landing passes on its measured touchdown, not on the flare setting.

**2.11** (a) 105 °C. The 2, 15 and 24 °C rows float, the last with only 0.4 K; 28 and 33 °C do not. (b) The crown temperature. The model is right about when the limit is approached and optimistic about where.

### Chapter 3. Friction Stir Weld Regime-Dependent Control

**3.1** (d). Tool temperature is the controlled variable and spindle speed the manipulated variable in every architecture.

**3.3** (b). The constant bias absorbs the gain error as an offset that decays over about a minute, long enough to make 18.2 mm of flash.

**3.5** (a) 0.2275, 16.77 rpm/°C and 11.87 s. (b) 26.91 rpm/°C and 2.90 s, a gain 1.60 times and a reset rate 4.1 times the servo tuning. (c) 2.28 on the traverse model and 4.29 on the plunge model.

**3.7** (a) 12.2 s (first `DWELL` row 12.3 s), 20.0 kN, 372 °C. (b) 9.0 s, still 20.0 kN, 405 °C. (c) 3.8 s and 18.1 kN, below the limit, with the tool at 319 °C: the stall is gone.

**3.9** (a) 91.6 %, 21 mm, 507 °C. (b) 82.5 %, 43 mm, 513 °C: worse. (c) 94.7 %, 13 mm, 506 °C: better. The model gain is too high, so the controller already moves too little, and suppression makes its moves smaller still.

**3.11** (a) The middle of a long weld is forgiving, and all the defects of the five complete welds lie in the first 52 mm. The scrap is at the two ends. (b) About 34.7 °C for wear to half life and 14.3 °C for an anvil contact of 0.30. A closed loop turns either into a move of the spindle speed, so the trend to watch is the manipulated variable.

### Chapter 4. Boiler-Turbine Offset-Free Predictive Control

**4.1** (c). The swell is steam in the risers, not water, and a loop that sees only the level cuts the feedwater at the moment the drum needs more.

**4.3** (a). The tube plans its nominal unit under $145 - h_p$, which is $142.76\,\mathrm{kg/cm^2}$ at the end of the run, so a setpoint of 144 is out of reach.

**4.5** (a) $-0.002\,51\,\mathrm{s^{-1}}$, a time constant of 399 s. (b) $0.0175\,\mathrm{kg/cm^2}$ per second and about $1.54\,\mathrm{kg/cm^2}$ after 100 s, against 1.50 in case 1. (c) 1.94 points, the two points case 1 adds.

**4.7** (a) Boiler-following 2.69 MW and $3.51\,\mathrm{kg/cm^2}$, turbine-following 9.67 MW and $0.10\,\mathrm{kg/cm^2}$, coordinated 0.81 MW and $1.21\,\mathrm{kg/cm^2}$. (b) $2.96\,\mathrm{kg/cm^2}$ with the lead and $4.09\,\mathrm{kg/cm^2}$ without it. (c) Each structure decides who supplies the ramp's steam first. Under sliding pressure the fire must also fill the drum, which the lead supplies.

**4.9** (a) Missed on the pressure band ($8.4\,\mathrm{kg/cm^2}$), the final pressure ($1.07\,\mathrm{kg/cm^2}$) and the credit (0.6 %). (b) Missed on the final pressure only, $1.81\,\mathrm{kg/cm^2}$. (c) Met. Sliding pressure earns the credit, and the fuel-bias state removes the offset that the poorer fuel leaves.

**4.11** (a) From the drum's design pressure and its safety-valve settings under the boiler code. (b) It keeps the plan away from the limit in normal operation. It is not protection, which belongs to independent safety valves and trips. (c) Little: reading gauge moves the steam density by 1.1 to 1.6 %, less than the fitted line's own misfit.

### Chapter 5. Irrigation Canal: Releasing Ahead of Demand

**5.1** (b). Each gate passes whatever arrives to hold its own upstream level, and the head does not know the farm closed.

**5.3** (d). The tail plan is final after round one, gate 2's after round two and the head's after round three.

**5.5** (a) $n_r = 0, 0, 1$ and $f_r = 0.717, 0.983, 0.100$. (b) 1.57 mm in the first period, 5.55 mm in each period after. (c) Shares in proportion to the surfaces, 24.4 %, 33.0 % and 42.6 %, and a common fall of about 1.35 cm.

**5.7** (a) About 10, 29.5 and 92.5 min. (b) Gate 3 passes about $0.05\,\mathrm{m^3/s}$ of the extra $0.5\,\mathrm{m^3/s}$ and reach 3 has risen about 0.4 cm. The delays grow faster than the summed travel times (11, 26 and 45 min) because each fixed gate passes only part of the change.

**5.9** (a) 2.18 cm, then 3.10, 3.09, 2.89 cm. (b) 4.99 cm, then 6.68, 4.55, 2.89 cm. (c) 1.93 cm, then 2.01, 2.28, 2.90 cm. (d) Exchanged plans release the order ahead gate by gate. Only the central plan lets reach 3 take the surprise to spare reaches 1 and 2.

**5.11** The exchange here is instantaneous and lossless and never reaches its cap of ten rounds, and an unsettled decision would be accepted without a mark. Test degraded exchange (one round per decision gives 4.88 cm against 3.10 cm after the surprise) and the fall-back to no exchange, which is less safe than the loops under a large unannounced demand.

### Chapter 6. Pricing Heat in a District Heating Network

**6.1** (b). The flow falls at each junction (35, 19 and 10 kg/s at 13:05), and each segment's volume passes at its own flow.

**6.3** (d). The two-capacity model is the plant's own model, so its error is the floor that any model learned from data is measured against.

**6.5** (a) 224 h. (b) 0.69 K, the same by either equation. (c) 144 kW and 168 kW. The 24 kW more amounts to about 0.07 MWh in three hours.

**6.7** (a) 4.48 h at 50 kg/s, 2.83 h at 78 kg/s. (b) 1.58 against 1.56, and 2.87 h predicted against 2.83 h. (c) The flow divides among the consumers almost the same way at both hours, so every segment's flow scales with the plant's.

**6.9** (a) Not met: \$31,952 against the \$31,553 target, a bill of \$31,580 with the stores 5.0 MWh short. (b) Met at \$31,445, a bill of \$30,334 with the stores 14.9 MWh short. (c) $31\,580 + 74.5 \times 5.0 = \$31{,}952$ and

$30\,334 + 74.5 \times 14.9 = \$31{,}444$. The bill falls 3.9 % and the graded cost 1.59 %. The target grades the cost with the stores valued.

**6.11** A larger boiler (at 17 MW the shipped distributed MPC holds both, at 0.45 and 0 K), a second source, or a contract that lets a consumer be cut. The sweep covers fourteen settings at the shipped boiler and tank. It is not a proof over every setting.

### Chapter 7. Reservoir Flood Control on Ensemble Forecasts

**7.1** (a). The local catchment peaks a day before the inflow, and the rule reads the tributary only at the town gauge, a day too late.

**7.3** (c). With seven members $k/M$ is a declared fraction of the ensemble, not a measured probability.

**7.5** (a) $C_0 = C_2 = 0.231$, $C_1 = 0.538$. (b) $O(0) = -0.25$, $O(1\,\mathrm{d}) = 0.642$, zero at 0.18 d. (c) A dip of 25 $\mathrm{m}^3/\mathrm{s}$, which a ramp spreads over the day.

**7.7** (a) 552 $\mathrm{m}^3/\mathrm{s}$, 4 days over, 322 $\mathrm{hm}^3$ of pool used. (b) 525, 489 and 425 $\mathrm{m}^3/\mathrm{s}$, with 4, 3 and 0 days over: five days of lead. (c) The pool used falls only to 307 $\mathrm{hm}^3$: the lead moves the release earlier, while the channel is still empty.

**7.9** (a) Not met: 8 days over, 75.6 % of 529.6 GWh, levee failed. (b) Not met: 9 days over, 89.5 % of 586.2 GWh, ending at 341 $\mathrm{hm}^3$. (c) Met: no day over, 89.5 % of 590.8 GWh, ending at 305 $\mathrm{hm}^3$, town peak 425 $\mathrm{m}^3/\mathrm{s}$. (d) The SDP's season table puts the town over the limit, and each bound is computed for its own run's end storage.

**7.11** (a) The policy learns a milder winter than the plant has, so its storm releases overshoot the channel: 530 $\mathrm{m}^3/\mathrm{s}$ in case 5 and 9 days over the limit in the challenge. (b) It made the water value, and so every plan priced with it, worse. A utility would check the record's storms against the basin's own and test the policy on the storms it must survive.

### Chapter 8. Data Center Cooling With a Moving Optimum

**8.1** (d). The spray sweep runs \$87.95, 88.24, 88.49, 88.67 and then \$86.18, so a slope method started at the dry end stops at the local minimum.

**8.3** (b). Equation (8.11) prices cooling electricity and water, IT power excluded, subject to the limits of (8.12) and the coil holding the setpoint.

**8.5** (a) 8.52. (b) $COP = 6.06$, 482 kW. (c) 6.29, or 3.8 % per degree. The supervisor chooses warm chilled water and a small lift, where the Carnot term climbs without bound while the ceiling of 21 does not move.

**8.7** (a) Shipped 2.59, $+0.27$, $-0.26$. Gain 0: 5.39, $+0.78$, $-0.77$. A cut of about 65 %. (b) 4.98, $-0.79$ as the job starts and $+1.02$ a minute after it ends: the valve moves further than the load needs in both directions. (c) \$98.08 to \$98.10 at 60 min and \$98.01 to \$98.02 at 65 min, the three runs within \$0.02 at the same minute: the feedforward changes when the valve moves, not the chilled water setpoint.

**8.9** (a) MPC 2322 kWh, 16.71 $\mathrm{m}^3$, \$265.65. Optimizer 2335 kWh, 15.88 $\mathrm{m}^3$, \$265.29, cheaper by \$0.36. (b) 24.38 °C at 95.8 % against 24.78 °C at 90 %, 0.40 °C cooler under the plan. (c) The plan's walk from 7 °C arrives later, and after the trip it buys the cooler hall with spray at \$0.73 per hour more. The Optimizer holds its setpoints because no grid point is legal.

**8.11** A cold-aisle transmitter reading high. The loop holds the hall about 2 °C too cold, at a 6 % electricity penalty, and a second sensor at another rack or a calibrated portable instrument finds it.

### Chapter 9. Building HVAC Control With a Comfort Band

**9.1** (c). Floating at the ceiling saves energy, because less heat crosses the envelope, but no cooling is bought early.

**9.3** (a). From the first scan of the event the terminals sit on their minimum, so stored cooling buys comfort, not kilowatts.

**9.5** (a) 5.39 min and 333 min, a ratio of about 62 (61 for the measured 5.45 and 335 min). (b) 8.43 min: the airflow conductance leaves the air balance. (c) 30 % and 76 %.

**9.7** (a) \$602, 222 kW, 100.0 %. (b) \$585 and 202 kW at 0, \$620 and 242 kW at 180, \$561 and 188 kW at 480 min. (c) The energy bill stays between \$401 and \$414 while the demand charge runs from \$160 to \$206: the horizon is a demand-charge knob.

**9.9** (a) \$558, 183 kW, 100.0 %, energy \$402.47 and demand \$155.14. (b) \$561, 188 kW, 100.0 %, \$401.46 and \$159.73: the plan pays \$4.59 less demand charge and about \$1 more energy. (c) \$579 and 200 kW, \$41 better than the supervisor at the same horizon.

**9.11** (a) On a day three degrees above design the outdoor air stays at or above 20.0 °C, never under the 18 °C high limit, so the damper never opens past its ventilation minimum. (b) Cool but humid outdoor air brings a latent load the coil must remove, so a dry-bulb changeover can cost money where an enthalpy changeover would not.

### Chapter 10. Wastewater Aeration Setpoint Selection

**10.1** (c). By Equation (10.4) the steady growth rate is $b_A(T) + 1/\mathrm{SRT}$ whatever the temperature, so the effluent ammonia carries the loss.

**10.3** (a). The supervisor is seeded bumplessly from the 2.0 mg/L setpoint in force, and an eight-hour integral takes about a day to walk away from it.

**10.5** (a) 0.2411/d. (b) $X_A = 101.6\,\mathrm{g\,VSS/m^3}$ from 520.4 kg N/d. (c) 0.2982/d and 82.1 g VSS/m$^3$ at the same nitrified mass. The model's 79.5 g VSS/m$^3$ nitrifies 503.8 kg N/d, so the effluent rises to about 1.21 g N/m$^3$.

**10.7** (a) 4335 kWh, 2.275 g N/m$^3$, 3.27 g N/m$^3$ at 11:44 to 11:45. (b) 3813 kWh, 3.784 g N/m$^3$, peak 4.79 g N/m$^3$, 603 min over the permit: 522 kWh or 12 % saved and the permit broken. (c) Scenario 6 opens at its own 1.2 mg/L equilibrium, the frontier at the 2.0 mg/L one.

**10.9** (a) 4018, 4010 and 4011 kWh at 2.210, 2.218 and 2.206 g N/m$^3$, peak 3.22 g N/m$^3$ at 20:37. (b) 4306, 4309 and 4306 kWh at 2.595, 2.628 and 2.602 g N/m$^3$, 362 min over the permit. (c) 278 kWh or 6.5 % for 6.9 kg more nitrogen, about 40 kWh per kilogram.

**10.11** (a) $0.48 \times 82 = 39.4\,\mathrm{Nm^3/min}$, and $2(40.0 - 37.1) = 5.8\,\mathrm{Nm^3/min}$ vented, which closes $37.1 - 31.3$. (b) The blower cannot turn down below its surge line, so the air saved at the tanks goes out of the vent. A second smaller blower, inlet guide vanes for wider turndown or a basin out of service at night, each sized on the vented volume.

### Chapter 11. Integer Pump Scheduling for a Water Network

**11.1** (b). With no pump running, the hill holds its floor only above a level that rises with the draw, 3.73 m near the morning peak.

**11.3** (d). Pump 3 gives 66 to 68 L/s against a draw above 67 L/s from 05:16 to 21:31, so the tank falls until the junction cannot supply the hill.

**11.5** (a) 29.3 %, 64.6 %, 11.7 %. (b) About 7.3 %. (c) $3.0 \times 10^{-4}$ \$/(m$^3$ h).

**11.7** (a) \$242, 1881 kWh, 13 min below, lowest 24.4 m at 06:33. (b) \$128, 1577 kWh, never below, lowest 26.1 m. (c) 4.55 m against 2.41 m; \$0.129 against \$0.081 per kWh.

**11.9** (a) Not met: \$572 (energy \$203, leakage \$368), 55 min below the floor, lowest 24.6 m. (b) Met: \$515 (\$163 and \$352), lowest 25.8 m, most starts 2. (c) Not met on cost alone, \$544 above \$530, with the floor held.

**11.11** (a) 3.01 h, 184 kWh, \$12.90 against \$36.85, a saving of \$23.95. (b) \$4.32 a day; the night fill pays more than five times over.

### Chapter 12. Automatic Train Operation: Where to Coast

**12.1** (a). With $H$ linear in $F$, the maximizing force is at a bound or zero, and holding speed is the singular arc, so part forces only place a switch inside a stage.

**12.3** (c). The feedforward uses the declared mass, so a proportional loop must hold an error to make up the force for 40 t, and the integral removes that error.

**12.5** (a) 233.0 kN dry, 74.6 kN wet. (b) 45.8 km/h; 168 kN at 60 km/h, so the wheels spin up past the peak and the grip falls. (c) 12.2 ms and 38.2 ms; about 2.6 wet time constants pass in one control period.

**12.7** (a) Coast at 2.45 km; B 3.1 s early, 19.5 kWh, 37.1 drawn and 17.6 kWh returned, 76 %. (b) B 21.0 s early, 28.2 kWh, 58.9 drawn and 30.7 kWh returned, 75 %: the coast point saves 31 % net and 37 % drawn. (c) B 1.7 s early with 16.4 kWh, about 16 % less, by coasting from 0.35 km into the curve.

**12.9** (a) Not met: B 0.6 s early, then a stall at 7.10 km with 34.63 m/s of slip. (b) Met: B 1.7 s and C 1.4 s early, stops within 0.04 m, 51.8 kWh of the 55.1 kWh allowed, largest slip 1.36 m/s, one re-plan. (c) C 15.1 s late with 57.9 kWh: the plan expects a dry rail, falls behind on the wet climb, and the re-plans buy back time with energy until no plan can meet the clock.

**12.11** (a) The peak comes within one 0.1 s period of the wet entry, so it depends on where in the period the entry falls, which the noise moves. With the report the plan already asks for no more than the wet rail gives. (b) A drive-level slip controller acting in milliseconds, with the override as a backstop, and adhesion reports from earlier trains fed to the planner; check with logged slip peaks against the declared limit over many runs.

### Chapter 13. Summary and Outlook

**13.1** A good answer asks for the baseline retuned once by a stated rule, on the same seed, instruments, forecast and tariff, over several kinds of day, with the saving split into energy and demand charge and with comfort and travel reported. On the building the supervisor retuned once closed the design day at \$552.74 and won the heat wave, \$668.77 against \$682.12, while the plan won both curtailment days at full comfort and the challenge's peak. On the data hall the plan was the cheapest on none of four cases and its gain over the shipped grid search came from continuous decisions, not the horizon. The purchase is justified by the days the plan wins for a reason the rule cannot copy.

**13.3** A good answer checks that the step test was taken in the regime the controller will run in (the weld's plunge model finished 91.6 % sound against 98.3 %), that the linear response holds over the horizon for the moves the controller will make (the balloon's response over-predicts a large move by 29 % at 90 s and a 120 s horizon cycled at 214.4 m), that an integrating output has a drift term (offset 132.9 against 3.9 m), and that an unmeasured disturbance such as fuel quality is estimated so that the loop is offset-free (the boiler's nominal MPC held 1.25 kg/cm$^2$ off). It also asks that the model come from the design, not from the run's own parameters.

**13.5** A good answer pairs the savings credited to the wrong cause with taking each saving apart into its sources, the numbers copied from the app's prose with measuring them, the horizon result with varying the terminal weight as a second factor, the cost booked twice with reading the source against the balance, and the truth reads with moving a hidden value to see whether the controller moves. The procedure gives read-only access by default, a baseline of the existing control retuned once, a model identified in the operating regime and checked over the horizon, a comparison on identical inputs and several day types with each saving decomposed, a statement of which margins the plan spends, a person's approval under management of change for any write, protections left above the plan, and a logged record with a rollback.

# Glossary

Each term is given as this volume uses it, in one plain sentence, followed by the chapter where it is first used in earnest. The two indexes give every page.

**activated sludge** A mixed culture of bacteria kept in suspension in aerated tanks, where it eats the waste in sewage and turns ammonia into nitrate. *Chapter 10.*

**active constraint** A limit that the best operating point sits on, so that relaxing the limit a little would make the result better. *Chapter 1.*

**air handling unit** The box of a building system that mixes outdoor and return air, filters and cools it, and blows it into the ducts. *Chapter 9.*

**ammonium analyzer** An instrument that samples treated water and reports its ammonia content, late and only once every few minutes. *Chapter 10.*

**anti-windup** A guard that stops the integral of a controller from growing while its actuator sits at a limit. *Chapter 1.*

**automatic train operation** A system on board a train that drives it from station to station along a planned speed curve, under train protection. *Chapter 12.*

**back-off** A margin kept between a limit and the value the controller aims for, so that normal variation rarely breaks the limit. *Chapter 1.*

**boiler-following** A way of running a power unit in which the turbine valve sets the output and the fuel follows to restore steam pressure. *Chapter 4.*

**branch and bound** A search for the best whole-number plan that splits the problem in two and drops any branch that cannot beat the best plan found. *Chapter 1.*

**cascade control** An arrangement in which one controller sets the setpoint of another, faster one beneath it. *Chapter 2.*

**chance constraint** A limit that a plan must meet with a stated probability, rather than in every possible future. *Chapter 1.*

**coasting** Running a train with the motors off, so that it rolls on its own momentum and slowly loses speed. *Chapter 12.*

**combined heat and power** A plant that makes electricity and uses the heat left over for a heating network, abbreviated CHP. *Chapter 6.*

**comfort band** A range of room temperatures that occupants accept, which a controller may use freely instead of holding one setpoint. *Chapter 9.*

**constraint tightening** Planning against limits drawn a little inside the true ones, so that errors around the plan still keep the plant inside. *Chapter 4.*

**costate** A price that optimal control attaches to each state, telling how much one more unit of it is worth at that moment. *Chapter 1.*

**critical node** The point of a water network with the lowest pressure, which decides whether every customer is served. *Chapter 11.*

**DCS** Distributed control system: the plant computer network on which the basic control loops of a large plant run. *Chapter 1.*

**dead time** A pure delay between a move of the actuator and the first sign of it in the measurement. *Chapter 2.*

**demand charge** A part of an electricity bill set by the highest power drawn over a short window of the month, not by the energy used. *Chapter 9.*

**dissolved oxygen** The oxygen held in solution in water, which the bacteria of an aeration basin need to break down waste. *Chapter 10.*

**distant downstream control** Canal control in which each gate answers to the water level at the far, downstream end of its reach. *Chapter 5.*

**distributed MPC** Several predictive controllers, each planning its own part of a plant, that exchange plans or prices to agree. *Chapter 1.*

**disturbance model** The part of a controller's model that says how an unmeasured upset is assumed to act, so that it can be estimated. *Chapter 1.*

**DMC** Dynamic matrix control: the first widely used predictive controller, which plans moves on a measured step response. *Chapter 1.*

**droop** The amount a governor lets the speed or frequency fall as the load rises, so that several units share a change. *Chapter 4.*

**dual decomposition** Splitting a shared plan among several decision makers who each answer a price, which is adjusted until supply meets demand. *Chapter 1.*

**duct static pressure** The air pressure inside a supply duct, held by the fan so that every room terminal can draw the air it needs. *Chapter 9.*

**dynamic programming** Solving a long decision problem backward, stage by stage, by keeping for each state the best cost still to come. *Chapter 1.*

**economic MPC** A predictive controller whose plan

minimizes an operating cost, such as money spent on energy, instead of a distance from a setpoint. *Chapter 1*.

**economizer** A damper arrangement that cools a building with outdoor air whenever that air is cool enough. *Chapter 9*.

**elevated tank** A water tank on a tower or hill whose height gives the pressure in the pipes below it. *Chapter 11*.

**ensemble forecast** A set of forecasts of the same future, each from slightly different starting conditions, whose spread shows the uncertainty. *Chapter 7*.

**flood pool** The part of a reservoir kept empty in the wet season so that it can catch a flood. *Chapter 7*.

**free cooling** Cooling with cold outdoor air or water instead of a chiller whenever the weather allows. *Chapter 8*.

**friction stir welding** Joining metal plates with a spinning tool that heats and stirs them along the seam without melting them. *Chapter 3*.

**gate stroking** Moving a canal's gates on a precomputed schedule so that the water arrives where and when it is wanted. *Chapter 5*.

**Hamiltonian** The quantity that optimal control maximizes or minimizes at each instant, adding the running cost to the priced rates of change. *Chapter 1*.

**heating curve** A rule that sets the supply water temperature of a heating network from the outdoor temperature. *Chapter 6*.

**integer scheduling** Planning decisions that come in whole units, such as pumps on or off, usually by mixed-integer programming. *Chapter 1*.

**integrating process** A process whose output keeps moving as long as its input is off balance, like a level or an altitude. *Chapter 1*.

**Kalman filter** An estimator that blends a model's prediction with each new measurement, weighting each by how much it is trusted. *Chapter 1*.

**level band** A rule that starts pumps when a tank falls to a low level and stops them when it reaches a high level. *Chapter 11*.

**linear programming relaxation** The same problem with its whole-number decisions allowed to take any value, whose cost bounds every whole-number plan. *Chapter 1*.

**local upstream control** Canal control in which each gate holds the water level just upstream of itself. *Chapter 5*.

**MILP** Mixed-integer linear programming: an optimization with linear costs and limits in which some decisions must be whole numbers. *Chapter 11*.

**model mismatch** The difference between the controller's model and the plant it controls. *Chapter 1*.

**move blocking** Letting the plan change the input only at a few chosen steps of the horizon, to make the problem smaller. *Chapter 2*.

**move suppression** A price on each change of the input that makes a predictive controller act more gently. *Chapter 1*.

**MPC** Model predictive control: planning the next moves on a model over a horizon, applying the first and planning again. *Chapter 1*.

**night flow** The flow into a water network in the small hours, when use is lowest, used to estimate how much leaks away. *Chapter 11*.

**nitrification** The step of treatment in which bacteria turn ammonia into nitrate, using oxygen. *Chapter 10*.

**non-anticipativity** The rule that a decision made now cannot depend on which future later turns out true. *Chapter 1*.

**offset-free tracking** Reaching the setpoint exactly in steady state even when the model is wrong or an upset persists. *Chapter 1*.

**output bias** A correction that adds the present gap between measured and predicted output to every future prediction. *Chapter 1*.

**PID** Proportional-integral-derivative: the common feedback controller that acts on the error, its sum over time and its rate of change. *Chapter 1*.

**PLC** Programmable logic controller: a rugged industrial computer that reads inputs and drives outputs by a stored program. *Chapter 1*.

**Pontryagin's principle** The rule that an optimal control makes the Hamiltonian best at every instant, used to find the shape of an optimal run. *Chapter 12*.

**power usage effectiveness** The total energy of a data center divided by the energy its computers use, and a value of one would mean no overhead. *Chapter 8*.

**precooling** Cooling a building harder before a costly period so that its structure can absorb heat later. *Chapter 9*.

**prediction horizon** How far ahead a predictive controller looks when it plans its moves. *Chapter 1*.

**pressure-reducing valve** A valve that holds the pressure downstream of it at a set value whatever the pressure upstream. *Chapter 11*.

**price of time** In a train plan, the energy one second of journey time is worth, which decides where to coast. *Chapter 12*.

**QP** Quadratic program: minimizing a quadratic cost under linear limits, the problem a predictive controller solves each period. *Chapter 1*.

**receding horizon** Planning over a fixed horizon, applying only the first move, and planning again from the new measurement. *Chapter 1*.

**regenerative braking** Braking a train with its motors working as generators, so that the energy can be fed back to the line. *Chapter 12*.

**RTU** Remote terminal unit: a field computer at a distant site that reports measurements and moves equipment on a control center's orders. *Chapter 1*.

**rule curve** A chart that sets the target storage of a reservoir for each day of the year. *Chapter 7*.

**Saint-Venant equations** The equations of mass and momentum for shallow water flowing in an open channel. *Chapter 5*.

**scenario tree** A branching set of possible futures over which a plan is made, with one decision shared at each branch point. *Chapter 1*.

**self-optimizing control** Choosing which variable to hold constant, and at what value, so that the plant stays near its best without reoptimizing. *Chapter 10*.

**slack variable** An extra decision that measures by how much a soft limit is broken, so that the break can be priced. *Chapter 1*.

**sliding pressure** Running a power unit with its turbine valves wide open and letting the steam pressure fall and rise with load. *Chapter 4*.

**sludge age** The average time the bacteria stay in a treatment plant before they are removed as waste sludge. *Chapter 10*.

**soft constraint** A limit that a plan may break if it must, at a price, so that the problem always has a solution. *Chapter 1*.

**split-range control** One controller output shared by two actuators, such as a burner over half its range and a vent over the other. *Chapter 2*.

**step response model** A model given as the output change at each step after a unit step of the input. *Chapter 1*.

**step test** Moving one input by a fixed step and recording how the output answers, to build a model. *Chapter 3*.

**stochastic dynamic programming** Dynamic programming in which the future inflow or load is random and the cost to come is an expected value. *Chapter 7*.

**swell** The brief rise of drum level when steam demand rises, because falling pressure makes the bubbles in the water grow. *Chapter 4*.

**terminal cost** A value placed on the state left at the end of the horizon, so that a plan does not empty its stores. *Chapter 1*.

**thermal mass** The heat a building's structure can store, which slows how fast its rooms warm or cool. *Chapter 9*.

**three-element level control** Drum level control that adds the measured steam and feedwater flows to the level itself. *Chapter 4*.

**time-of-use tariff** An electricity price that changes with the hour of the day by a published table. *Chapter 9*.

**transport delay** A delay that comes from material having to travel, so that it is longer when the flow is slower. *Chapter 6*.

**tube MPC** A predictive controller that plans a nominal path with tightened limits, so that the real path stays in a tube around it. *Chapter 4*.

**turbine-following** A way of running a power unit in which fuel sets the output and the turbine valve holds the steam pressure. *Chapter 4*.

**value function** The best cost still to come from each state, the quantity that dynamic programming computes. *Chapter 1*.

**valve position control** Adjusting a shared pressure so that the most open valve served by it stays almost fully open. *Chapter 10*.

**VAV terminal** A box in a duct that varies the air flow into a room by a damper, to hold the room temperature. *Chapter 9*.

**water value** The worth of one more unit of water left in storage, which decides whether releasing it now pays. *Chapter 7*.

**wet bulb** The lowest temperature that evaporating water can reach in the outdoor air, which limits a cooling tower. *Chapter 8*.

**wheel slip** The difference between a wheel's rolling speed and the train's speed when the rail cannot carry the force asked. *Chapter 12*.

# Index of Control Topics


active constraint, 8
  moving with the conditions, 172
AI agents
  engineering practice, 318
  open problems, 322
  reproduction, 320
  truth audit, 320
  verification record, 319
  while studying, 17
ammonia-based aeration control, 245

back-calculation, 61
back-off, 8, 246
boiler-following, 79
branch and bound, 10, 273, 281
bumpless transfer, 15, 61, 85, 103

cascade control, 32, 183
  time-scale separation, 240, 257
chance constraint, 159, 167
constraint tightening, 86
constraints
  floor row, 273
  hard and soft, 7
coordinated control, 76, 92
costate, 294

data export, 16
data-enabled predictive control, 135
dead time, 22
  in the measurement, 56
  transport delay, 27
distant downstream control, 98, 119
disturbance model, 6, 34, 85
droop, 82, 92
dual decomposition, 10, 126
duality gap, 135
dynamic matrix control, 5, 22, 32, 49
dynamic programming, 2, 9, 287
  stochastic, 157

economic model predictive control, 8
economic MPC, 316
  margins, 314
economic objective, 179
equal-compliance comparison, 249

fault-tolerant control, 273
feedforward, 103, 297
  released ahead, 99
feedforward control, 71
  measured load, 183
  model-based, 85
first-order plus dead time model, 59
force control, 51, 69
forecast correction, 276

gain scheduling, 66, 188
grid search, 184

Hamiltonian, 294
Hankel matrix, 135
heating curve, 133, 143

infeasibility, 161, 273, 300
integer scheduling, 10, 265
  optimality gap, 275
integrating process, 22, 30, 104
  output bias, 39
inverse response
  swell and shrink, 81
ITAE tuning, 61

Kalman filter, 16
  input disturbance, 85, 106, 217, 247
  steady-state, 134

leak detection, 274
level control
  three-element, 85
linear programming, 158
linear programming relaxation, 273
local upstream control, 99, 100
lumped-parameter model, 207

measurement filter, 110, 118
mixed-integer linear programming, 272
model mismatch, 314
model predictive control, 1, 42
  against PID, 315
  anti-windup, 34, 62
  benefits, 11, 311, 318
  design procedure, 314
  distributed, 10, 98, 119, 125, 144, 317
  economic, 184, 202, 234

in the apps, 15
move blocking, 33, 62, 86, 104, 134, 216, 245
non-cooperative, 107
offset-free, 76
open problems, 71, 228, 258, 320
prediction model, 4, 104
stochastic, 11, 149
successive linearization, 86
tube, 76
tuning, 5, 228
move suppression, 5, 34, 62
moving horizon estimation, 16, 71

Nash equilibrium, 107
non-anticipativity, 158
nonlinear gain, 49

observability, 85
offset, 88
offset-free tracking, 6, 34, 43, 106, 217, 247
output bias, 63
optimal control, 9, 287
override control, 31, 297

perfect-foresight bound, 159
phase lag, 30
PID control
anti-windup, 15
velocity form, 15, 31, 60, 103, 214, 243, 273, 297
plug flow, 128
Pontryagin's principle, 9, 288
prediction horizon, 5, 38, 62, 220, 319
pressure management
flow modulation, 273
price of time, 294
process identification, 322
operating point, 49
projected gradient, 35, 41
pump scheduling, 265
minimum run and rest, 272

quadratic program, 3, 5, 134
solvers, 316

rate limit, 93, 216
re-planning, 297, 304
receding horizon, 2, 5, 104, 216, 245, 273
reference trajectory, 34, 63
reservoir operation, 149
resonance, 99
rule-based control, 157, 272

scenario tree, 11, 158
scenarios, 16, 63
completion rules, 16
scorecard, 16
self-optimizing control, 236
self-regulating process, 185
sensor bias, 191, 194, 225
setpoint as manipulated variable, 215
shared actuator, 214
simulation engine, 14
simulator framework, 13
singular arc, 294
slack variable, 7, 105, 273
soft constraint, 105, 134, 185, 217, 246
split-range control, 27
steady-state optimization, 8, 314
step response model, 4, 32
step test, 32, 59
stochastic MPC, 316
supervisory control, 12, 69, 202, 234
integral time, 250
supervisory optimization, 172

terminal constraint, 27, 273
terminal cost, 134
time-scale separation, 213
transport delay, 98
flow-dependent, 125
tuning
servo against regulator, 63
turbine-following, 79

value function, 157
valve characteristic
equal percentage, 195, 240
valve position control, 244

water value, 158

zone control, 203

# Index of Applications


activated sludge, 234
added mass, 26
adhesion, 292, 305
aeration
    share of site electricity, 235
affinity laws, 269
air handling unit, 206
air header, 240
aluminum alloy 6061, 52
ammonium analyzer, 245
anoxic zone, 237
atmospheric river, 150
automatic generation control, 84
automatic train operation, 287

backing anvil, 52
balise, 290
ballast, 39
balloon ascent model, 25
balloon envelope
    energy balance, 27
blower
    turndown, 237, 251
boiler-turbine unit, 76, 78
braking curve, 295
building HVAC, 202
building thermal mass, 130
buoyancy, 25

centrifugal blower, 237, 256
chiller
    air-cooled, 204
    centrifugal, 176
clearwell, 268
coasting, 288
coefficient of performance, 178
combined heat and power, 131
comfort band, 203, 227
computer room air handler, 176
cooling coil, 206
cooling tower
    hybrid, 176
creep curve, 292
critical node, 268, 279
cross-regulator, 101, 117
cycles of concentration, 179

data center cooling, 172, 174
Davis resistance, 292
demand charge, 210
demand response, 197, 205, 316
denitrification, 239
density altitude, 39
descent envelope, 27
dewatering liquors, 252
diffuser fouling, 251
discharge permit
    rolling mean, 241
dissolved oxygen probe, 256
district heating, 125, 127
drum boiler, 76, 90
drum level, 81
duct static pressure, 208

economizer, 208
electric train, 287
elevated tank, 266, 267
energy-efficient driving, 288, 289
ensemble forecast, 155, 167, 316
environmental flow, 154, 166
escape weir, 101

fine-bubble diffuser, 237
fixturing, 67
float temperature, 28
flood control, 149
flood pool, 151
flow stress, 54
flue-gas condenser, 131
free cooling, 175
friction stir welding, 49, 51, 69
    quality window, 50, 69, 70
fuel heating value, 79

gate stroking, 99
governor, 82, 91
gravity wave, 99

Hazen–Williams equation, 269
heat exchanger
    effectiveness, 130, 177
heat generation
    frictional and plastic, 54
heat network, 125
heat rate, 84

heated storage tank, 3
hot-air balloon, 22, 24
hot-aisle containment, 175
hydropower, 154

irrigation canal, 98, 118

landing, 36, 42
leakage, 269
    FAVAD exponent, 267
level band, 272
line receptivity, 293

main break, 269
makeup water, 179
moving heat source, 56
Muskingum routing, 154

night flow, 274
nitrification
    cold weather, 251
    Monod kinetics, 239

office building, 202
open-channel flow, 102, 117
oxygen saturation
    submergence, 240
oxygen transfer
    alpha factor, 240

parachute vent, 24
peak boiler, 131
pipe heat loss, 128, 144
platform stop, 297
plunge, force-limited, 55
power usage effectiveness, 175
precooling, 205, 206
pressure floor, 268
pressure management, 267
pressure-independent valve, 130
pressure-reducing valve, 268
propane burner, 24, 41
    fuel burn, 27
pump
    variable-speed drive, 142, 268
pump curve, 269, 280

quadratic drag, 25

rainfall-runoff model, 153
recirculation over containment, 176
regenerative braking, 292, 303
reservoir, 149
return temperature limiter, 130
rotating-mass factor, 292
rule curve, 151
run-out, 56

Saint-Venant equations, 102
server fans
    cube law, 177
server inlet temperature
    recommended limit, 175, 196
sliding pressure, 79
sludge age, 239
solids retention time, *see* sludge age
speed restriction, 290, 303
spillway gate, 165
standard atmosphere, 28
stored energy, 77
stream gauge, 165
substation, 130, 142
surge
    blower, 241

tachometer, 293, 303
temperature inversion, 28
thermal (rising air), 39
thermal mass, 205
thermal storage
    building fabric, 210
    stratified tank, 131
thermocouple
    rotating tool, 52, 69
time-of-use tariff, 210
timetable, 293
tool wear, 56
turnout, 101

unit hydrograph, 153

variable air volume, 204
variometer, 25, 41
VAV terminal, 208, 226
ventilation minimum, 208

wastewater treatment, 234
water distribution, 265, 281
    pumping cost, 266
water use effectiveness, 175
water-side economizer, 176
weld defect
    anvil strike, 57
    flash, 57
    wormhole, 57
wheel slip, 292

## About the Editor

John D. Hedengren is a professor in the Department of Chemical and Biological Engineering at Brigham Young University, where he leads the Process Research and Intelligent Systems Modeling (PRISM) group. He received the B.S. and M.S. degrees in chemical engineering from Brigham Young University and the Ph.D. degree from the University of Texas at Austin in 2005. He developed advanced process control applications for polymer production at ExxonMobil Chemical Company and commissioned nonlinear model predictive controllers with PAS, Inc. before joining the faculty at Brigham Young University in 2011.

His teaching covers process dynamics and control, engineering computing, data science, machine learning, and dynamic optimization, with course material published online for use inside and outside the classroom. He developed the Temperature Control Lab, a portable heat transfer device now used by many universities for process control and data science instruction, and he maintains the APMonitor and GEKKO optimization software used in those courses. He serves as Chair of Control Education for the IEEE Control Systems Society and as a trustee of the CACHE Corporation. He received the AIChE David Himmelblau Award in 2014, the AIChE Computing Practice Award in 2018, and the John R. Ragazzini Award in 2024, and he was inducted into the Process Automation Hall of Fame in 2025.

## Volumes in This Series

*Interactive Control Engineering: Principles and Applications* has 20 volumes. Nine are organized by control technology and eleven by industry or application area. Every chapter in every volume has a simulator at https://apmonitor.com/apps/.

1 **Process Dynamics, Modeling, and Identification.** How plants respond, and how to find the model.
2 **Practical PID Control.** Tuning, anti-windup and implementation that works on real loops.
3 **Feedforward, Cascade, and Ratio Control.** Choosing a structure from where the disturbance enters.
4 **Override, Split-Range, and Constraint Control.** When limits decide what the controller may do.
5 **Delays, Slow Measurements, and State Estimation.** Controlling what arrives late or cannot be measured.
6 **Multivariable and Plantwide Control.** Pairing loops, managing inventories, closing the recycle.
7 **Nonlinear Dynamics, Oscillations, and Stabilization.** Plants that will not stay put without control.
8 **Batch, Run-to-Run, and Hybrid Control.** Recipes, endpoints and learning from the last run.
9 **Model Predictive Control and Economic Optimization.** Planning moves ahead, and choosing where to operate.
10 **Control in Upstream and Midstream Oil and Gas.** From the wellbore to the terminal.
11 **Control in Petroleum Refining.** Throughput, quality and limits from crude unit to blending.
12 **Control in Petrochemicals and Polymers.** Severity, selectivity and grade from cracker to pellet.
13 **Control in Chemical Plants.** Recycle, equilibrium and energy in continuous chemical processes.
14 **Control in Semiconductor Manufacturing.** Run-to-run control where little can be measured in the chamber.
15 **Control in Mining and Mineral Processing.** Variable ore, late assays and equipment at its limit.
16 **Control in Iron, Steel, and Metals Production.** Slow furnaces, fast strip, and measurements that arrive once.
17 **Control of Energy Conversion and Storage.** Fast regulation and slow dispatch on a changing grid.
18 **Control of Nuclear and Fusion Systems.** Reactor physics, protection limits and coordinated control.
19 **Control of Biological Systems and Human Performance.** Variable subjects, sparse samples and limits that cannot be crossed.
20 **Motion Control, Robotics, and Autonomous Systems.** Tracking, navigation, contact and allocation.