The Pennsylvania State University
The Graduate School

ECONOMIC ISSUES IN NUTRIENT POLLUTION CONTROL

A Dissertation in
Agricultural, Environmental, and Regional Economics
by
Aaron Cook

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

May 2019

The dissertation of Aaron Cook was reviewed and approved* by the following:

James S. Shortle
Distinguished Professor of Agricultural and Environmental Economics
Dissertation Advisor, Chair of Committee

Douglas H. Wrenn
Assistant Professor of Environmental and Resource Economics

Karen Fisher-Vanden
Professor of Environmental and Resource Economics

Mort D. Webster
Associate Professor of Energy Engineering

Edward C. Jaenicke
Professor of Agricultural Economics
Chair of Graduate Program

*Signatures are on file in the Graduate School.

Abstract

Nutrient pollution represents one of the most significant threats to water quality in the United States and worldwide due to its physical complexity and the magnitude of its attendant environmental costs. Nutrient pollution problems involve elements of hydrology, biology, and engineering that complicate the economic analysis of optimal management and the design of efficient policy. These elements include 1) the persistent nature nutrient pollution, 2) the capital intensity of nutrient abatement processes, 3) lags times between nutrient discharge and delivery, and 4) the need to manage multiple pollutants jointly. Each essay in this dissertation treats some combination of these four elements.

The first essay examines the combined implications of elements 1, 2, and 3, developing a model to capture these aspects of the nutrient pollution problem and solving for the optimal time path of nutrient reductions across two polluting sectors—wastewater and agriculture. The model is calibrated to conditions in the Chesapeake Bay watershed and the optimal solution is compared to the reductions specified by the Chesapeake Bay's current Total Maximum Daily Load (TMDL) policy. The optimal plan calls for much more aggressive nutrient reductions in early periods relative to the TMDL, and the TMDL's total social cost exceeds the least-cost dynamic solution by 5-9% (depending on the lag length in the agricultural sector). An alternative policy—a time-invariant plan that jumps immediately to and maintains the optimal steady state loads for all time—exceeds the cost of the dynamically optimal plan by only 0.05%, suggesting the gains to a time-varying policy to be small despite the inherently dynamic character of the problem.

The second essay examines the implications of lag times for the design of markets for nutrient reductions. I characterize the first-best solution to the problem of managing discharges among sources with varying lag lengths, noting that optimality requires separate "regimes" of control corresponding to sets of sources that delivery their pollution at the same time. While this first-best solution would be prohibitively complex with either a forward market or a trading ratio system, the essay proposes a second-best trade ratio system that incorporates an adjustment to the trading rules based on the lag length disparity between the sources

involved in the trade. This second-best system will implement the optimal steady state loads in the long run, representing a practical approach to governing trades between the point and nonpoint sectors given differences in lag lengths.

The third essay examines the implications of complementarity in the costs of nitrogen and phosphorus removal at wastewater treatment facilities for the timing of policy implementation. When policies for two or more interdependent pollutants are implemented sequentially, potential cost savings may be overlooked. I develop a conceptual framework for evaluating the efficiency loss associated with managing two pollutants through a sequential policy. Analysis shows that the sequential policy is inefficient only for a subset of possible joint discharge targets (even when cost interdependencies exist). This framework is useful not only for evaluating and designing markets for nutrient reductions where municipal wastewater dischargers feature prominently, but also for other areas of environmental policy such as land conservation, habitat protection, and carbon sequestration where multiple environmental goods are produced jointly.

Overall, the essays represent three novel approaches for modeling several complex elements of the nutrient pollution problem. The findings therein offer conceptual guidance for the design of policies to help control it.

Table of Contents

List of Figures

List of Tables

Acknowledgments

For noteworthy contributions personal and professional, my gratitude goes out to:

Jim—for your support and guidance throughout the full duration of my dissertation-writing process. Your depth of experience and common sense have routinely cut through my confusion and set me on new, productive lines of thinking.

Doug—for the multiple opportunities to serve alongside you in the classroom. Teaching undergraduate environmental economics with you has been perhaps the most energizing and rewarding part of this PhD.

Mort—for your clear thinking and outstanding teaching.

Karen—for stepping in on short notice to round out "team dissertation." Thanks also for your work organizing and promoting our department seminar series. It has enriched my PhD experience greatly.

Lauren Chenarides—for your friendship and conversation during a dry and difficult stretch.

Christian Scott—for picking up where Lauren left off.

Kaleb Cook—for helping me see God more clearly and obey him more completely.

Jesus—for your kindness and forbearance. In reading about you and getting to know you more, I'm consistently amazed at how little effort you spent trying to convince others of the "rightness" of your arguments. I want to be more like that.

Chapter 1

Introduction

In the summer of 2014, the Collins Park Water Treatment Plant in Toledo, Ohio detected harmful levels of microcystin[1] in the city's drinking water, prompting Mayor Michael Collins to issue a do-not-drink advisory for area residents [28]. Ohio Governor John Kasich soon thereafter declared a state of emergency as over a half-million people were left without clean drinking water for parts of three days [17]. The elevated microcystin levels were attributed to an algal bloom that formed over the western part of Lake Erie near the Toledo's water-intake pipe. Algal blooms have become commonplace in Lake Erie, especially after spring rains flush excess fertilizer into the lake from surrounding farmland. Indeed, this was not the first occasion an Ohio community's drinking water had been tainted by harmful algae—less than a year prior, nearby Carroll Township had also been forced to issue a similar do-not-drink advisory during the second largest algal bloom in the lake's history [80].

Problems of this kind are not restricted to Toledo. In the past several decades, the ecosystems of countless streams, lakes, and coastal waters around the world have seen increases in toxic algae prevalence, declines in dissolved oxygen levels (hypoxia), and a reduction of economically and ecologically important stocks of fish and other aquatic animals [16, 76]. Governments worldwide have mobilized resources to assess the nature and causes of these problems. In 1974, seven European nations agreed to the terms of the Helsinki Convention to determine the extent to which the declining ecological health of the Baltic Sea was attributable to human influence [11]. Japan took similar action in 1978 to study the onset of similar problems in the Seto Inland Sea [6]. As a result of these and similar initiatives, scientists have identified excessive nutrient enrichment, or "nutrient pollution," as the main

[1] a toxic by-product of cyanobacteria (blue-green algae) known to impair liver function

source of the water quality impairments previously described. While nutrient concentrations can fluctuate through natural hydrological processes, long run data suggest that anthropogenic infusions of limiting nutrients such as nitrogen (N) and phosphorus (P) are driving these changes [7, 15]. The problem has become so widespread in the United States that an EPA task group recently identified nutrient pollution among the costliest and most challenging environmental problems of the 21st century [44], with the Chesapeake Bay and the Gulf of Mexico representing the highest profile examples [61].

Chesapeake Bay has been at the center of national nutrient pollution policy discussions in the United States. During the last several decades, the country's largest estuary has seen increases in the average sizes of hypoxic and anoxic zones, a higher prevalence of toxic algae, declining in water clarity, and a lower prevalence of submerged vegetation [48, 34, 29]. In the interest of protecting the Bay's ecological resources, the U.S. EPA, along with the Commonwealths of Virginia and Pennsylvania, the State of Maryland, and the District of Columbia agreed to coordinate efforts to improve the estuary's ecological health, signing the Chesapeake Bay Agreement of 1983 [63]. Failure to produce material improvements in the Bay's ecological health over the ensuing years eventually led the EPA in 2010 to establish a total maximum daily load (TMDL) for the Chesapeake Bay's 64,000 square mile watershed [41]. This, the largest TMDL in history, defined limits on the quantities of N, P, and sediment delivered annually to the Bay and requires that states develop watershed implementation plans (WIPs) to specify how these limits would be attained [63]. While the economic analysis in this dissertation applies to nutrient pollution problems in general, the ecological and political significance of the Chesapeake Bay make it a natural application for the economic models developed herein.

Nutrient pollution problems involve physical, biological, and engineering features that complicate the economic analysis of optimal management and the design of efficient policy. First, rates of algae growth and their attendant damages to aquatic environments depend importantly on the accumulated stocks of N and P [71]. This biological feature creates a need for dynamic analysis since the decision about how much nutrient delivery to allow over time must take into account current and future effects. Second, the processes required to reduce N and P discharges are more capital intensive in some sectors than others. In particular, the procedures for removing nutrients from municipal wastewater often involve large, irreversible investments in pipes, basins, and aeration tanks [68]. These stocks of capital are costly to adjust in the short run, which introduces an additional dynamic element to the problem.

Third, sources of nutrients to receiving waters differ widely with respect to the delay

between the time the nutrients leave the source until their eventual delivery. Many of the most prominent sources of N and P are agricultural operations whose nutrient runoff may not arrive at the receiving waters until years or decades later [55]. The substantial heterogeneity in the speed of delivery from various sources must be accounted for in decisions about where to allocate nutrient reduction effort across a watershed and when considering how sources should be allowed to offset one another.

Fourth, nutrient pollution involves multiple pollutants that are interdependent with respect to both environmental effects and reduction costs. Research in the natural sciences have found that reductions of *both* N and P are critical for maintaining healthy estuaries and coastal waters [12, 64] and attention must therefore be paid to their joint control. Likewise, engineering principles dictate that the capital costs associated with reducing N and P concentration in wastewater are interdependent. Cost-effective targeting of joint reductions must therfore keep this in mind.

The essays in this dissertation each treat some combination of these four complicating features. Chapter 2 examines the combined implications of the persistence of nutrients in water systems (element 1), the differences in the capital intensity of reduction measures among various nutrient sources (element 2), and differences in lag times among those sources (element 3) for the timing of nutrient reduction effort. The management question becomes how to jointly manage harmful nutrient stocks and stocks of nutrient abatement capital over time.

The Chesapeake Bay TMDL of 2010 set a goal to reduce nitrogen discharges by 68 million lbs per year by 2025, relative to the 260 million lbs. discharged in 2009. The policy required 60% of these annual reductions to be performed by 2017 (the midpoint of the time frame) with the remaining 40% completed by 2025. The 2017 and 2025 targets represent a simple "time path" for nitrogen load reductions where loads are drawn down gradually from a high baseline. The policy question to be explored is whether the gradual reductions under this plan bear any relation to an optimal reduction time path that accounted for pollution stock dynamics, capital stock dynamics, and the lags that characterize load reductions in the agricultural sector.

Chapter 2 develops a model to capture the relevant hydrological and economic aspects of this problem. The model is calibrated to conditions in the Chesapeake Bay watershed and is used to compare the optimal solution to the reductions specified by the Chesapeake Bay's current Total Maximum Daily Load (TMDL) policy. The optimal plan calls for much more aggressive nutrient reductions in early periods relative to the TMDL, and the TMDL's total

social cost exceeds the least-cost dynamic solution by 5-9% (depending on the lag length in the agricultural sector). An alternative policy—a time-invariant plan that jumps immediately to and maintains the optimal steady state load—exceeds the cost of the dynamically optimal plan by only 0.05%.

Chapter 3 examines the implications of lag times (element 3) for the design of markets for nutrient reductions. Past research has examined how spatial heterogeneity [58, 43] and risk [54, 37, 36, 39] influence the optimal design of these markets. The fact that sources delivering pollution to the body of water can have such different lag lengths implies that these discharges are not perfect substitutes. To guarantee efficiency of a market involving sources with dissimilar lag lengths, a set of trading rules must be established to account for the lag length disparity.

The essay first characterizes the first-best solution to the problem of managing discharges among sources with varying lag lengths, noting that optimality requires separate "regimes" of control corresponding to sets of sources that deliver their pollution at the same time. While this first-best solution would be prohibitively complex to achieve with either a forward market or a trading ratio system, the essay then proposes a second-best trade ratio system that incorporates an adjustment to the trading rules based the lag length disparity between the sources involved in the trade. If established immediately, this second-best system will implement the optimal steady state loads in the long run. The simple design proposed here represents a practical approach for governing trades between the agricultural and wastewater treatment sectors given differences in lag lengths.

Chapter 4 examines the implications of interdependencies in the costs of N and P removal at wastewater treatment facilities (element 4) and the costly adjustment of these processes (element 2) for the timing of policy implementation. While the chemical and biological processes that remove N and P from wastewater are distinct, the fact that these systems share certain components implies that upgrades for N (P) removal are less costly where P (N) removal capacity already exists [68]. Lence et al.'s [52] analysis of optimal nutrient removal upgrades among municipal wastewater facilities on the Willamette River show that the sequence of policy implementation matters for the cost-effectiveness of nutrient removal investments when there are complementarities in the pollution reduction process. When policies for two or more pollutants are implemented sequentially, potential cost savings can be overlooked. Policies implemented with the *joint* reduction target in mind from the start will be more cost-effective.

While Lence et al. [52] identify this issue, their empirical approach (optimizing over a

set of discrete pollution reduction options) precludes them from drawing more general conclusions about how the degree of cost complementarity affects the magnitude of the losses associated with a sequential policy relative to a simultaneous one. In this essay, I develop a conceptual framework for evaluating the efficiency loss associated with managing two pollutants through a sequential policy. The analysis shows that the sequential policy is inefficient only for a subset of possible joint discharge targets, even when these cost interdependencies exist. This framework is useful not only for evaluating and designing markets for nutrient reductions where municipal wastewater dischargers feature prominently [65, 66], but also for other areas of environmental policy such as land conservation, habitat protection, and carbon sequestration where multiple environmental goods are produced jointly [82, 53].

Overall, these essays represent three novel approaches for modeling several complex elements of the nutrient pollution problem. The findings therein offer conceptual guidance for the design of policies to help control it.

Chapter 2

Dynamically Efficient Nutrient Load Allocations: Should Managers Pay Attention to Time Path?

2.1 Introduction

The physical characteristics of nutrient pollution outlined in Chapter 1 make it an immensely difficult management problem. In particular, the problem is complicated by hetergeneity among nutrient sources with respect to the capital intensity of their nutrient reduction processes and the elapsed time between a source's nutrient discharge and the eventual delivery to the receiving waters. In general, point source discharges of nutrients are characterized by quick delivery times, but reductions often involve expensive, long-lived investments in nutrient removal capital. Nonpoint sources almost invariably have lower nutrient reduction costs per pound, but the actual effects of these reductions are often felt only after considerable delay. The question for management is how to allocate nutrient reduction effort across point and nonpoint sectors over time.

Because of the way nutrients accumulate in aquatic environments, it is important to consider not only the size of the nutrient loads but also their timing. Load timing matters for efficiency because the ecological damages incurred from nutrient pollution tend to be a nonlinear function of the accumulated nutrient stock [71], with damages becoming more and more severe as the pollution stock grows. Loads of equivalent size may thereby inflict different costs depending on the size of the the existing pollution stock when the new loads

are delivered.

The timing is also important insofar as the various polluting sectors differ with respect to their ability to influence load deliveries on short notice. In the case of agricultural runoff, nutrients often move slowly from topsoil to groundwater to streams, sometimes taking a decade or more to finally reach their destination [55]. Meanwhile, past runoff (perhaps percolating slowly through groundwater systems) will be delivered to the receiving waters regardless of current nutrient management that farmers undertake. Current efforts to curb nutrient runoff will therefore affect delivered nutrient loads only after considerable delay. On the other hand, municipal wastewater facilities often discharge nutrient-laden wastewater directly to surface water and therefore deliver their nutrient loads much faster on average. The fact that reductions from the wastewater sector can prevent immediate nutrient build-up makes their reductions more ecologically valuable than equivalent reductions from agriculture. Decisions over where to allocate nutrient reduction effort must therefore take into account differences in how quickly benefits will accrue.

An additional complicating factor is that the wastewater and agricultural sectors also differ in terms of the capital intensity of their nutrient abatement process. Load reductions from the wastewater sector often entail lumpy and irreversible investments in nutrient removal capacity that could "commit" the wastewater sector to certain reduction levels for many years to come [68]. Agricultural abatement, on the other hand, is based on much more flexible decisions about output, fertilizer usage, and fertilizer application techniques [10]. In some cases, the farm might make landscape alterations (to prevent the speed or the extent of runoff), which may constitute a medium-term investment but not on the scale of a nutrient removal upgrade in the wastewater sector. How much to invest in load reductions in the wastewater sector depends on the immediate severity of the nutrient stock level, combined with the reduction costs in the wastewater sector relative to the cheaper (but potentially delayed reduction measures) that could be undertaken in the agricultural sector. Ideally, the wastewater sector would assume a heavy burden of load reductions early and then scale back as the cheaper agricultural reductions eventually become effective. As suggested by the irreversible characteristics of wastewater nutrient upgrades, this "scaling back" may not be possible, and instead, managers must strike a balance between mitigating immediate pollution damages and preventing large fixed and irreversible costs.

The economic question becomes how to optimally manage the growth of both the nutrient pollution stock and the stock of nutrient removal capital. Since TMDLs frequently specify timelines for how load limits should be phased in, it is natural to ask whether these timelines

make economic sense given these stock dynamics. This essay develops a model to solve for the dynamically efficient nutrient load allocations across two polluting sectors (municipal wastewater and agriculture) accounting for the damage costs associated with the pollution stock, the abatement costs required to make load reductions in each sector, and the processes of pollution accumulation and decay.

I calibrate the model to conditions in the Chesapeake Bay watershed and compare the timing of these optimal load reductions to the timing of those outlined by the Chesapeake Bay TMDL, one of the largest and most comprehensive nutrient reduction policies in the world. Because I'm primarily interested in these time paths, I assume the long run steady state loads implied by the TMDL are optimal and ask whether the load reduction time path recommended by the TMDL is efficient. I compute excess costs of the TMDL under a set of lag lengths (0, 2, 4, ..., 20 years) and find the costs[1] of the Chesapeake Bay TMDL to be 5-9% greater than the dynamically optimal plan (depending on the lag length scenario). These cost excesses occur because the TMDL phases in load reductions more gradually than the optimal plans. This gradual path to the steady state, while perhaps easier to implement politically, increases overall costs by allowing ecological damages to mount in the policy's early years.

I also compare the optimal plan to a simple, time-invariant plan where loads jump immediately to and maintain the steady state optimal load allocations (that correspond to the correct lag length). Such a plan, by ignoring the path to the steady state, would be suboptimal but be easier to implement (a policy maker would merely announce nutrient load limits on each sector that would hold for all time). As it turns out, the cost of disregarding the optimal path and making this steady state "shortcut" would consist of a 0.05% increase in total costs relative to the dynamically optimal plan.

Finally, because lags introduce an additional complexity to the management problem (especially considering how they vary across space and management practice), what would be the cost of assuming them away? I evaluate a third plan similar to the second except for each lag scenario (0, 2, 4, ..., 20 years) loads jump to and maintain the steady state load allocations for the *zero-lag* scenario in all cases. As expected, the plan performs very well for short lags and less well as lag length increases. At its worst, i.e. under the 20-year lag scenarios, this plan costs 3.8% more than the optimal plan. In this case, even a time-invariant plan targeting an inefficient steady state allocation performs better than gradually phasing in load reductions as recommended by the TMDL.

[1] costs represent the sum of abatement costs (across the two sectors) and pollution damage costs

In the remaining sections of the paper I review the literature on dynamic pollution control, develop a conceptual model of nutrient pollution management, translate this to an empirical model, discuss time paths and costs associated with plans outlined above, and finally conclude.

2.2 Literature Review

Because nutrient pollution damages are a function of a nutrient stock, the problem requires dynamic management. The basic theory of pollution stock management was first formally presented by Keeler, Spence, and Zeckhauser [47] as an optimal growth problem where output could be allocated to consumption, capital accumulation or reduction of a pollution stock. Subsequent work includes Falk and Mendelsohn [21] in the context of global warming policy, Griffin [32] in relation to a persistent pollutant that varies across space, and Van Der Ploeg and Withagen [77] with respect to the use of taxes and subsidies to manage a pollution stock in a Ramsey-type model.

The most common way to model the evolution of stock nutrients is to have emissions at any point in time contribute directly and instantly to the stock. This is limiting in the nutrient pollution case because of lags between emissions and their contributions to damaging stocks. Winkler [81] and Augeraud-Veron and Leandri [3] provide continuous time frameworks for modeling lags, the former assuming a single, discrete delay and the latter allowing for distributed delays. (also [8]) Hart [35] solves for the optimal discrete time dynamic allocation between two nutrient reduction methods—agricultural reduction measures upstream becoming effective with a delay, and nutrient sequestration measures (mussel cultivation) downstream becoming effective immediately. While these studies deal with the lagged nature of certain pollution sources, they do not yet address the implications of abatement measures that entail long-lived, irreversible capital investments.

Since advanced nutrient removal processes depend heavily on specialized facilities and equipment, the size and scope of which can be costly to adjust in the short run, the appropriate model must address these capital constraints. Methods for modeling capital adjustment are developed in several papers. Singh and Weninger [70] investigate capital adjustment costs for the optimal management of a stock (in this case a biomass stock), though their model does not involve delays between control variables and effects on the state variables.

More directly relevant, Laukkanen and Huhtala [51] combine stock pollution with a lumpy, fixed capital investment. Instead of treating capital as a continuous variable, their

model more closely resembles and optimal stopping problem in which managers must decide the moment at which to invest in nutrient removal capacity in the wastewater sector, given mounting damages from a stock pollutant. In their model, the capital need not be maintained; in fact, the one-time decision to invest affords the wastewater sector the ability to reduce nutrient loads as low as desired, given enough variable expenditure. This latter assumption is problematic because to ratchet wastewater loads lower and lower requires additional expenditures on lumpy capital (in addition to the potentially larger operating cost expenses). The feature of requiring both capital investment and operating cost expense for any load reduction, which are described in sections 2.3 and 2.4 below, is absent from existing literature on wastewater pollution control.

An optimal management strategy for nutrient pollution control will necessarily define a time path for nutrient removal capital investment—lagged deliveries in the agricultural sector, capital adjustment costs in the wastewater sector, and pollution stock dynamics will influence these paths. While particular studies have examined pieces of this problem, none have yet dealt with them in combination. In what follows, a pollution control model that includes these dynamic elements is developed to study the characteristics of the optimal solution and consider their implications for policy.

2.3 Conceptual Framework

Prominent sources of nutrient pollution to the Chesapeake Bay include wastewater treatment facilities and agricultural operations, which consist of about 40% and 20% of total deliveries nitrogen [63]. Given the differences between these sources in terms of the timing of their deliveries and their nutrient reduction cost structure, optimal load allocations for the two sectors at a point in time will depend on the urgency of the reductions (i.e. the severity of near term damage costs associated with the existing pollution stock), disparities in abatement cost between sectors, and disparities in delivery time between sources.

A theoretical model that incorporates these features of the nutrient pollution problem is developed below. The model is implemented subsequently for the Chesapeake Bay. To enable a focus on the intertemporal tradeoffs between nutrient reductions in the two sectors, the model is simplified by assuming one WWTP and one farm contribute nutrients to a stock, S, which produces environmental damages according to the function $D(S)$. Each source may reduce their nutrient discharges at a cost, though the cost structures for the two sectors differ in economically important ways.

The WWTP receives wastewater from a municipal sewer system, having some baseline nutrient concentration $\bar{n}$ (mass/volume), at an exogenous rate ω (volume/time). The WWTP controls the nutrient concentration of its effluent by increasing its stock of nutrient removal capital K. It is useful to express the plant's capital in terms of its nutrient reduction capacity. Accordingly, let

$$K \;=\; \bar{n} - n \tag{2.1}$$

where n is the post-treatment nutrient concentration in the WWTP's effluent. K translates directly to the plant's capacity to reduce its nutrient concentration so that $K = 0$ implies no nutrient removal ($n = \bar{n}$) and $K = \bar{n}$ represents complete nutrient removal ($n = 0$). After installing capital at level K, the plant discharges nutrients to S at rate $(\bar{n} - K)\omega$ (mass/time). Since nutrient removal capital is durable, K is preserved wholly or partially from one period to the next depending on the depreciation rate.

The incremental costs of reducing nutrients at the WWTP are twofold. First, by investing in nutrient removal capital at rate I, the plant incurs costs $g(I)$. It is assumed that investment in period t affects the plant's nutrient removal capacity in period $t+1$. Second, in the process of performing the removal, the plant incurs variable costs $v(K)$. If capital is able to substitute for variable inputs in nutrient removal, then v would be a decreasing function of K. If on the other hand an upgraded nutrient removal process requires more intense use of variable inputs, v would be an increasing function of K.

The farm has some baseline nutrient discharge rate $\bar{e}$ from which it can abate some portion $A \in (0, \bar{e})$ through the adoption of a nutrient BMP, incurring costs $c(A)$. Some farm-level BMPs involve creating structures that provide nutrient pollution control over multiple time periods. (e.g. riparian buffers, manure storage units, barnyard runoff filters), while others involve alterations to planting or nutrient application patterns that do not have effects beyond the practice period. In either case, these measures represent less of a fixed capital investment relative to the abatement process at WWTPs. To simplify the model and highlight this contrast in capital intensity between abatement at the farm and the WWTP, it is assumed that the farm's BMPs are effective only during the implementation period and the cost of a given level of A is independent of previous period's choices of A.

The two sources also differ with respect to the timing of nutrient deliveries relative to abatement actions. Since wastewater treatment facilities often discharge directly to surface water, delivered nutrient loads respond immediately to abatement actions in the wastewater sector. On the other hand, nutrients that discharge from agriculture often take slower paths over the land surface or through the groundwater system before reaching the zone where

the nutrient pollution problems occur. To capture this difference, I assume the WWTP's discharges in period t, $(\bar{n} - K_t)\omega$, arrive in period $t+1$, whereas discharges in period t from the agricultural source, $(\bar{e} - A_t)$, arrive in period $t+l+1$.

The growth of S and K are given by

$$S_{t+1} \;=\; (1-\delta)S_t \;+\; (\bar{n} - K_t)\omega \;+\; (\bar{e} - A_{t-l}) \tag{2.2}$$

and

$$K_{t+1} \;=\; (1-\gamma)K_t \;+\; I_t \tag{2.3}$$

respectively, where $\delta \in [0,1]$ is the per-period decay rate of the pollution stock and $\gamma \in [0,1]$ is the per-period depreciation rate of the WWTP's capital stock. Assuming $\gamma > 0$, the WWTP's concentration will eventually revert to the baseline $\bar{n}$ in the absence of investment. The terms on the right-hand side of (2.2) represent (in order) the nutrients that carry over from the previous period, the WWTP's loads discharged in the current period, and discharges from agriculture l periods ago.

2.3.1 The Model

The regulator's objective is to allocate nutrient reductions between the WWTP and agriculture given their distinct cost and delivery features. Let Z denote the present value of the stream of abatement expenditure and damage costs incurred in each period. Formally, the goal is to

$$\min_{I_t,\; A_t} \; Z \;=\; \sum_{t=0}^{\infty} \left[g(I_t) \;+\; v(K_t) \;+\; c(A_t) \;+\; D(S_t) \right] (1+r)^{-t} \tag{2.4}$$

subject to (2.2) and (2.3) and initial states S_0 and K_0. The Lagrangian for this problem is

$$\begin{aligned} \mathcal{L} \;=\; & \sum_t \left[g(I_t) \;+\; v(K_t) \;+\; c(A_t) \;+\; D(S_t) \right] (1+r)^{-t} \;+\; \sum_t \mu_{t+1} \left[K_{t+1} \;-\; (1-\gamma)K_t \;-\; I_t \right] \\ & +\; \sum_t \lambda_{t+1} \left[S_{t+1} \;-\; (1-\delta)S_t \;-\; (\bar{n} - K_t)\omega \;-\; (\bar{e} - A_{t-l}) \right] \end{aligned}$$

with first order conditions

$$\frac{\partial \mathcal{L}}{\partial I_t} \;=\; \frac{g'(I_t)}{(1+r)^t} \;-\; \mu_{t+1} \;=\; 0 \tag{2.5a}$$

$$\frac{\partial \mathcal{L}}{\partial A_t} \;=\; \frac{c'(A_t)}{(1+r)^t} \;-\; \lambda_{t+l+1} \;=\; 0 \tag{2.5b}$$

$$\frac{\partial \mathcal{L}}{\partial S_t} \;=\; \frac{D'(S_t)}{(1+r)^t} \;+\; \lambda_t \;-\; (1-\delta)\lambda_{t+1} \;=\; 0 \tag{2.5c}$$

$$\frac{\partial \mathcal{L}}{\partial K_t} \;=\; \frac{v'(K_t)}{(1+r)^t} \;+\; \mu_t \;-\; (1-\gamma)\mu_{t+1} \;+\; \omega\lambda_{t+1} \;=\; 0 \tag{2.5d}$$

that must hold at the optimum for all t. The recursive structures of (2.5c) and (2.5d) imply

$$\lambda_t \;=\; -\frac{1}{(1+r)^t}\left[\sum_{i=0}^{\infty}\left(\frac{1-\delta}{1+r}\right)^i D'(S_{t+i})\right] \tag{2.6}$$

and

$$\mu_t \;=\; -\frac{1}{(1+r)^t}\left[\sum_{i=0}^{\infty}\left(\frac{1-\gamma}{1+r}\right)^i v'(K_{t+i})\right] \;-\; \omega\left[\sum_{i=0}^{\infty}\lambda_{t+1}(1-\gamma)^i\right] \tag{2.7}$$

Derivations of (2.6) and (2.7) are found in appendix A.

Intuitively, λ_t represents the damage cost savings (in present value terms) associated with a marginal change in S_t. A nutrient "pulse" that reaches the stock in period t will generate marginal damage costs equal to $D'(S_t)/(1+r)^t$, followed by a cascade of damage costs through time as period $t+i$ retains $(1-\delta)^i$ of this pulse. Each subsequent period's marginal damage costs, $D'(S_{t+i})$, are therefore discounted both for the extent of decay, $(1+\delta)^i$, and for the timing of the effects, $(1+r)^{-(t+i)}$.

Similarly, μ_t represents the abatement cost savings (in present value terms) associated with a marginal change to K_t. The two bracketed terms in (2.7) signify two channels through which the size of the nutrient removal capital stock may influence total costs. The first channel represents capital's effect on operating costs, while the second represents effects on future damage costs via changes in pollution discharge. Each is discussed in turn.

The first term in (2.7) resembles (2.6), except the marginal operating cost function $v'(K)$ replaces $D'(S)$ and the capital depreciation rate γ replaces δ. Analogous to (2.6), this first term in (2.7) represents the operating cost savings (in present value terms) associated with adding a small increment to the capital stock in period t. As the capital stock augments, current period operating costs will increase or decrease by $v'(K)/(1+r)^t$, followed by a cascade of marginal effects on operating costs through time as period $t+i$ retains $(1-\gamma)^i$

of this additional capital. The sign of the effect on operating costs could be positive or negative in principle. If capital were a substitute for variable inputs in the nutrient removal process, additional capital would reduce present and future operating costs. Alternatively, if a more capital intense process required greater use of variable inputs (energy, chemicals, etc.) additional capital would increase present and future operating costs.

The second term in (2.7) accounts for the damage cost savings associated with augmenting the capital stock. Recalling equation (2.1), additional capital in period t will reduce the nutrient concentration in the WWTP's effluent by $(\bar{n} - \Delta K_t)$, which in turn will reduce the mass of nutrients discharged in period t by $(\bar{n} - \Delta K_t)\omega$ (hence ω multiplies each λ_{t+i}). This nutrient reduction in period t will have persistent effects across all future periods and these effects are fully captured by λ_t, as illustrated by (2.6). The first term, λ_t, fully accounts for the cascade of damage costs arising from the "pulse" sent out in period t, but the subsequent terms $\lambda_{t+1}, \lambda_{t+2}$, etc., also appear in the bracketed expression because an adjustment to K_t will generate changes in nutrient discharges in all future periods by virtue of the capital stock's durability. In this way, capital augmentation that occurs in period t modifies all future discharges, but the effects diminish over time as capital depreciation sets in. Depreciation is reflected in the factors $(1 + \gamma)^i$ applied to each λ_{t+i}. Altogether, μ_t captures the present value of operating cost savings and damage cost savings that arise from a marginal adjustment to period t's nutrient removal capital stock.

With λ_t and μ_t expressed as functions of state variables, S and K, I examine conditions (2.5a) and (2.5b) which dictate the optimal loads from wastewater and agriculture in light of the pollution and capital stock dynamics. Condition (2.5a) states that in every period, the present value of marginal investment costs must equal the present value of cost savings (including both operating costs and pollution damage costs) associated with marginally increasing the capital stock in the following period [2]. Were marginal investment costs greater than (less than) the marginal value of future cost savings, investment would decrease (increase) until equality is restored. In turn, condition (2.5b) states that, in every period, the present value of marginal cost savings of increasing agricultural emissions must equal the present value of pollution damage costs associated with marginally increasing the pollution stock $l + 1$ periods later[3]. Whenever the marginal pollution damages become greater than (less than) the marginal cost of agricultural nutrient abatement, abatement should increase (decrease) until equality is restored.

[2]This accounts for the one-period delay between investment and capital augmentation

[3]This accounts for the $l + 1$ period delay between discharges from agriculture and downstream effects

2.3.2 Steady State Conditions

I next consider optimal behavior in the long-run steady state, where $S_t = S^s$, and $K_t = K^s$ for all t. Combining these steady state conditions with (2.6) and (2.7),

$$\lambda_t \;=\; -\frac{D'(S^s)}{(r+\delta)}(1+r)^{-t+1} \tag{2.8}$$

and

$$\mu_t \;=\; \left[-v'(K^s) \;+\; \omega\frac{D'(S^s)}{(r+\delta)}\right]\frac{(1+r)^{-t+1}}{(r+\gamma)} \tag{2.9}$$

Equations (2.8) and (2.9) together with (2.5a) and (2.5b) imply that steady state investment and agricultural abatement, I^s and A^s, must satisfy

$$\frac{g'(I^s)}{\omega} \;=\; -\frac{v'(K^s)}{\omega(r+\gamma)} \;+\; \frac{D'(S^s)}{(r+\delta)(r+\gamma)} \tag{2.10}$$

$$c'(A^s) \;=\; \frac{D'(S^s)}{(1+r)^l(r+\delta)} \tag{2.11}$$

where marginal investment costs and marginal operating costs are expressed per unit of wastewater flow. Since g and v are expressed as costs per unit of nutrient concentration ($\frac{\$}{\text{lb./gal.}}$), dividing by ω (gal.) converts each to cost per unit of nutrient mass ($\frac{\$}{\text{lb./gal.}} \cdot \frac{1}{\text{gal.}} = \frac{\$}{\text{lb.}}$). In this way, $\frac{g'}{\omega}$, $\frac{v'}{\omega}$, c', and D' are denominated in the same units.

Regarding the steady state optimality conditions (2.10) and (2.11), note first that g' and c' are both increasing functions of abatement effort, any "change" in the right-hand side of either expression will require steady state reductions in either sector to "respond" in the same direction. For example, a shift in D' that makes the pollution stock environmentally more costly will increase the right-hand sides of both (2.10) and (2.11). To restore equality, the left-hand sides must also increase, which is achieved by increasing abatement effort in both sectors. Following this logic, compare $v' < 0$ (where operating costs fall as K rises) versus $v' \geq 0$ (where operating costs increase with or are independent of K). If $v' < 0$, the first term on the right-hand side of (2.10) will be positive, implying positive cost savings associated with maintaining a higher capital stock and higher investment in the steady state (relative to $v' \geq 0$).

The parameters l and γ will also inform the relative load allocations of the two polluting sectors. Agricultural lag length l appears only in (2.11) and has an unambiguously negative effect on (2.11)'s right-hand side. As such, abatement effort would shift unambiguously

from the farm to the WWTP with any increase in l. Similarly, the depreciation rate γ will have a negative effect on (2.10)'s right-hand side (provided $\frac{D'(S^s)}{r+\delta} > \frac{v'(K^s)}{\omega}$) where higher depreciation rates will be associated with less abatement in the PS sector relative to the NPS sector.

2.4 Empirical Model

Wastewater first entering a treatment facility may contain nitrogen and phosphorus concentrations ranging 20-85 mg/L and 3-7 mg/L, respectively [68]. Standard wastewater treatment processes typically will reduce these concentrations down to 10-20 mg/L for nitrogen and to 1-2 mg/L for phosphorus, though many municipalities have mandated more stringent nutrient standards in response to the harmful effects of nutrient over-enrichment. Nutrient load reductions at WWTPs therefore represent extensions of their existing nutrient reduction capacity. I will consider the treatment levels from these standard processes as the baseline ($\bar{n}$) from which investment in more advanced nutrient removal (K) will occur.

Let $f(K)$ represent the cumulative costs of acquiring and installing K's worth of nutrient removal capital. I define the cost associated with augmenting the capital stock from K' to K'' as

$$g(K'', K') \;=\; f(K'') \;-\; f(K')$$

the difference in cumulative capital costs at each of the two levels of K. Letting I_t denote the incremental capital that is added in period t, g becomes the basis for the investment cost function in (2.4). We would expect investment costs at time t to depend on the size of the upgrade, I_t, and the existing level of K at the time of the investment, $(1-\gamma)K_{t-1}$ (i.e. the *undepreciated* portion of the previous period's capacity). To illustrate why existing K matters for investment costs, note that reducing the nitrogen content of wastewater from 11 mg/L to 10 will be easier than reducing from 3 mg/L to 2 even though these upgrades represent equivalent reductions. Because of these diminishing returns, we would expect a given change in K to be costlier the larger K becomes. Investment costs at time t can therefore be expressed

$$g(I_t, K_{t-1}) \;=\; f\Big[(1-\gamma)K_{t-1} + I_t\Big] \;-\; f\Big[(1-\gamma)K_{t-1}\Big] \tag{2.12}$$

Provided that f is characterized by diminishing returns to K, g will inherit these properties.

Constructing investment costs as the simple difference between total capital costs at

two different levels of treatment inevitably abstracts away from adjustment costs and, more fundamentally, whether nutrient removal capacity can truly be "built on top of" existing capacity. Nevertheless, this specification serves as a useful simplification and a point of departure for modeling these other, more complex elements.

To construct a model of point source nutrient reduction costs consistent with existing studies, I use the model put forward in Horan and Shortle [39]

$$\phi(PS) \;=\; u(\bar{PS} - PS)^3$$

as a basis for $g(I)$ and $v(K)$, where PS denotes annual point source nitrogen emissions (million lbs.), $\bar{PS}$ denotes the annual emissions baseline (million lbs.) and u is a scalar. While $\phi(PS)$ captures the essential features of the relationship between the total mass of nitrogen reduced and its associated cost (costs increase as emissions fall below baseline at an increasing rate), it obscures important distinctions between operating costs and capital costs which become relevant in a dynamic setting. It also ignores the engineering fact that WWTP's do not adjust nutrient *mass*, per se, but rather adjust the nutrient *concentration* of the treated wastewater they discharge. As public utilities, WWTPs are obliged to accept and treat everything that enters the sewer system, so unlike, say an industrial polluter, they do not have the option to "reduce output" as a way of keeping emissions low. It is therefore more natural to consider the volume of incoming wastewater as exogenous and define the WWTP's control variable as the nutrient concentration in that volume as it exits the plant.

Modifications to ϕ occur in three steps, each documented in appendix B. The first involves a basic unit conversion where I re-express ϕ as a function of nutrient concentration rather than total mass (B.1). The second decomposes ϕ into a linear combination of capital costs and operating and maintenance (O&M) costs (B.2). The third translates annualized capital costs into lump sum costs, scaling up the capital cost component to reflect the present value of real resources acquired and installed at the treatment facility (B.3). This last modification converts point source treatment capacity to a stock of capital to be managed over time rather than a set of independent annual decisions. At each step, the structure of $\phi(PS)$ is preserved such that the final result is a more generalized point source cost function of which $\phi(PS)$ is a special case.

Function	Description	Structure
$g(I_t)$	investment cost (given K_{t-1})	$c^{\text{CAP}}\left[I_t^3 \ + \ 3(I_t)^2 K_{t-1}(1-\gamma) \ + \ 3I_t(K_{t-1})^2(1-\gamma)^2\right]\Omega$
$v(K_t)$	operating and maintenance costs K_t	$c^{\text{OM}}(K_t)^3\omega$
$c(A_t)$	reduction costs in NPS sector	$c^{\text{NPS}}(A_t)^3$
$D(S_t)$	environmental damage costs	$d(S_t)^2$

Table 2.1. Structure of Cost Functions

Parameter	Description	Value	Units
Ω	WWTP design flow	1935	million gal. per day
ω	wastewater flow rate	1451	million gal. per day
r	discount rate	0.05	
δ	pollution stock decay rate	0.6	
γ	abatement capital stock depreciation rate	0.04	
$\bar{n}$	baseline nutrient concentration in PS sector	11.8	mg per liter
$\bar{e}$	baseline emissions from agriculture	113.8	million lbs. per year
S_0	initial nutrient stock	433.7	million lbs.
K_0	initial capital stock	0	mg per liter
c^{CAP}	capital cost coefficient	0.0205	
c^{OM}	O&M cost coefficient	0.0006	
c^{NPS}	NPS sector reduction cost coefficient	0.0095	
d	environmental damage cost coefficient	0.0485	

Table 2.2. Model Parameters

Following these modifications, point source reduction costs at t become

$$\begin{aligned} \theta(K_t, K_{t-1}) \ &= \ f(K_t) \ - \ f(K_{t-1}) \ + \ v(K_t) \\ &= \ g(I_t) \ + \ v(K_t) \end{aligned} \tag{2.13}$$

Table 2.1 summarizes the structure of each function in expression (2.4) and Table 2.2 summarizes the parameter values used in the numerical simulation.

The 2010 Chesapeake Bay TMDL specified allowable nutrient loads for various polluting sectors, including agriculture, wastewater, stormwater, septic, and forest. The plan used 2009 loads as a baselines and required a draw-down each year before meeting final sector-level load targets in 2025. Draw-downs were planned with the intermediate goal that 60%

	2009	2017	2025
Agriculture	113,798,000	88,666,000	71,912,000
Wastewater	52,179,000	43,585,000	37,855,000
Other	94,269,000	87,282,000	82,624,000
Total	260,246,000	219,533,000	192,391,000

Table 2.3. Chesapeake Bay TMDL, nitrogen targets by sector (lbs.)

of the 2025 load targets would be met by 2017. I show these targets in Table 2.3 for the agricultural and wastewater sectors, for all other sectors ("Other"), and for the total. In calibrating my empirical model to the Chesapeake Bay I used the values in Table 2.3 to inform both the baseline loads as well as the costs. Because my question deals specifically with the time path of the TMDL policy, I assumed loads given by the 2025 targets represent the steady state optimal solution, balancing abatement and damage costs. Supposing this to be the case, I measure the efficiency of the TMDL time path relative to the optimal time path.

2.5 Results

The optimal load allocations in each sector and in each time period depend importantly on the initial pollution stock and its decay rate, the wastewater sector's initial capital stock and its depreciation rate, and the lag length associated with the agricultural sector's nutrient deliveries. A higher initial pollution stock will increase abatement from both sectors in early periods since damages are more severe and immediate measures to reduce the pollution stock carry increased benefits. All else equal, the abatement would skew relatively more toward the wastewater sector for a longer lag in the agricultural sector. Once the efficient steady state has been achieved, the relative load allocations between sectors will be dictated by a combination of differences in abatement costs and differences in lag lengths. The optimal load at a point in time balances the magnitude of abatement costs against the magnitude and timing of the corresponding ecological benefits. When lags are present the benefits of agricultural nutrient reductions are deflated because they accrue later in time and the efficient steady state agricultural loads will be higher for longer lag lengths. This will be clear in Figure 2.1 below.

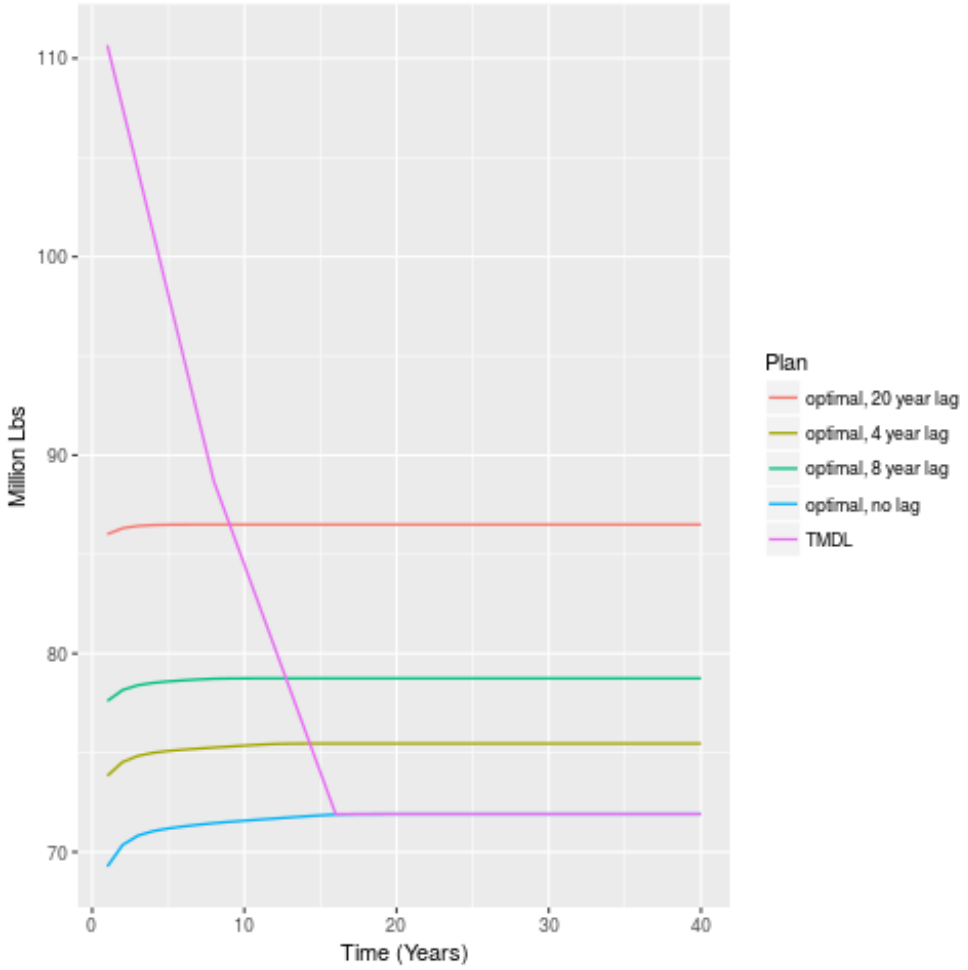

Figure 2.1. Time path of agricultural loads under TMDL and optimal plans

2.5.1 Optimal Time Paths

Figure 2.1 plots the time paths of the optimal agricultural loads under various lag length scenarios (no lag, 4-year lag, 8-year lag, and 20 year lag) coupled with the time path implied by the TMDL's 2017 and 2025 targets. Though the TMDL doesn't mandate specific targets in each individual year, I assumed the time path would consist of a straight-line decrease from the 2009 baseline to the 2017 targets (where 60% of the long run reductions would be achieved) and from the 2017 targets to the 2025 targets (where the remaining 40% of the long run reductions would be achieved).

The optimal paths differ from the TMDL paths in two ways. First, the optimal plans call for the most dramatic reductions in the very first period. Recalling that baseline agricultural loads are 113.8 million lbs. at $t = 0$, reductions in the first year of the optimal plans range

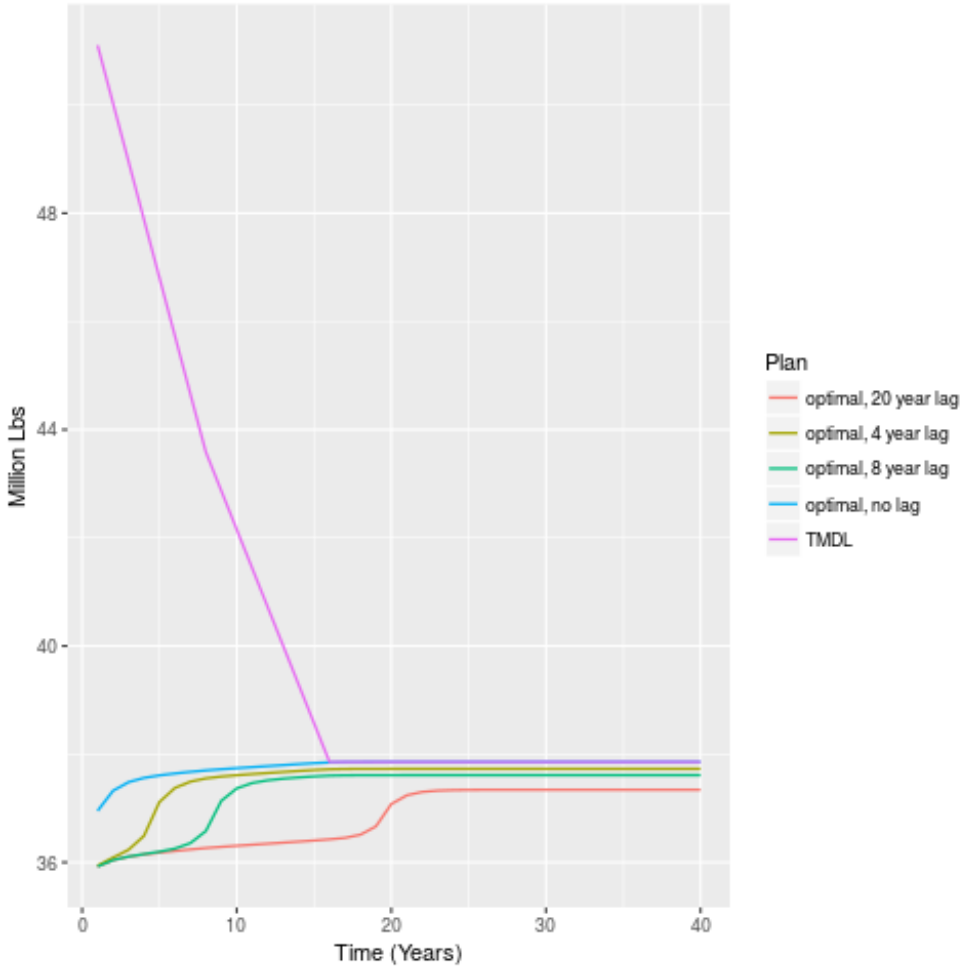

Figure 2.2. Time path of wastewater loads under TMDL and optimal plans

from about 28 to 45 million lbs. (depending on lag length) compared to just over 3 million lbs. under the TMDL. These results suggest that, even assuming its long run targets are optimal, the TMDL is too relaxed in the pace at which it approaches these targets. Second, because the TMDL does not account for agricultural sector lags, the optimal plan only lines up with the TMDL in the no-lag case. In all other cases, the optimal steady state load remains above the TMDL's 2025 target by 3.6 million lbs. in the case of a 4-year lag and by 14.6 million lbs. in the case of a 20-year lag.

Like its counterpart, Figure 2.2 plots the time paths of the optimal wastewater loads under various lag length scenarios. Again, though the TMDL does not mandate wasteload targets in each individual year, I interpolated the path based on the 2009 baseline loads and the 2017 and 2025 targets.

The optimal paths for wastewater loads resemble those in agriculture insofar as the most dramatic reductions occur in the initial period then follow a gradual increasing path toward the steady state optimal loads. The shape of the paths change as lags are introduced where the longer the lag length, the more slowly the load rises to meet the eventual steady state. Recall that, in the presence of lags, even if farms undertake immediate measures to curb their emissions, deliveries from the agricultural sector will remain at the baseline levels for l years. Loads in the wastewater sector must be held down longer in order to make up for the fact that agricultural abatement actions will not yet come into effect. Note in Figure 2.2 that the time at which wasteloads in the optimal paths begin rapidly approaching the steady state levels roughly corresponds with each scenario's lag length (the inflection points of the curves occur around 4, 8, and 20 years). Finally, as with the agricultural loads, the lag length also affects steady state wastewater loads. In this case, because lags exist in the other sector rather than in its own sector, a higher lag length implies smaller steady state loads. Assuming that the TMDL is efficient in the long run, with some positive lag length in agriculture, the TMDL overallocates loads to the wastewater sector.

2.5.2 Cost Comparisons

Having highlighted divergences between time paths of load allocations under the optimal plans and the TMDL plan, I turn next to the issue of how costly the TMDL and some alternative static plans are relative to the optimal plan. Figure 2.3 plots the total 40-year present value costs of four plans at various lag lengths, with all costs expressed as percentages of the optimal plan (where the optimal plan is normalized to 100). These costs include all those represented in the model—investment costs for nutrient removal capital, operating costs for the wastewater treatment facilities, costs of reducing loads in the agricultural sector, and the damage costs associated with the pollution levels each year. The three plans I compare against optimality are 1) the current Chesapeake Bay TMDL, 2) a static plan where loads in both the wastewater and agricultural sectors jump immediately to the steady state levels that prevail in the optimal solution, and 3) a static plan where loads in both sectors jump immediately to the steady state levels that prevail in the optimal solution for the *zero*-lag scenario. I plot these static plans for the agricultural and wastewater sectors in Figures 2.4 and 2.5, respectively. Plans 2) and 3) will have identical costs for lag length of zero since plan 2) always picks the steady state level corresponding to the correct lag length, and plan 3)'s "zero-lag assumption" happens to be correct in this particular case.

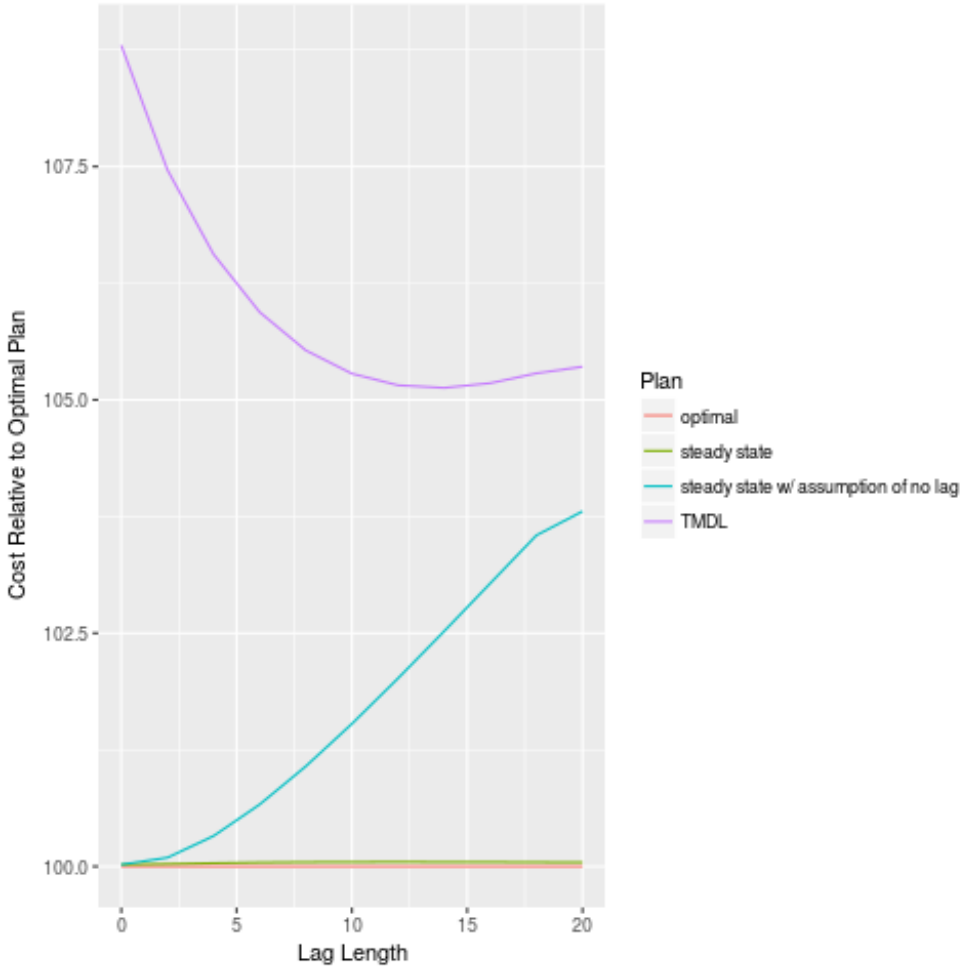

Figure 2.3. Cost of each plan under various lag lengths

Among the three non-optimal plans, the TMDL has the highest cost under any lag length scenario. The TMDL entails 8.8% higher costs than the optimal plan under a lag length of zero and a 5.1% higher cost than the optimal plan under a lag length of 14 years, with results at other lag lengths falling between these values according to the curve in Figure 2.3. The TMDL becomes relatively less costly for lag lengths greater than zero because, as seen in Figure 2.1, the optimal agricultural loads get higher and higher as lag length increases, and the gap between the TMDL path and the optimal paths therefore becomes less severe in earlier periods when lags are present. Meanwhile, as the steady state agricultural loads rise optimally with lag length they also make the TMDL steady state inefficiently low in later periods. The change is a net efficiency gain because the excess costs incurred early in the time horizon are worth more than those incurred later (from discounting). These gains

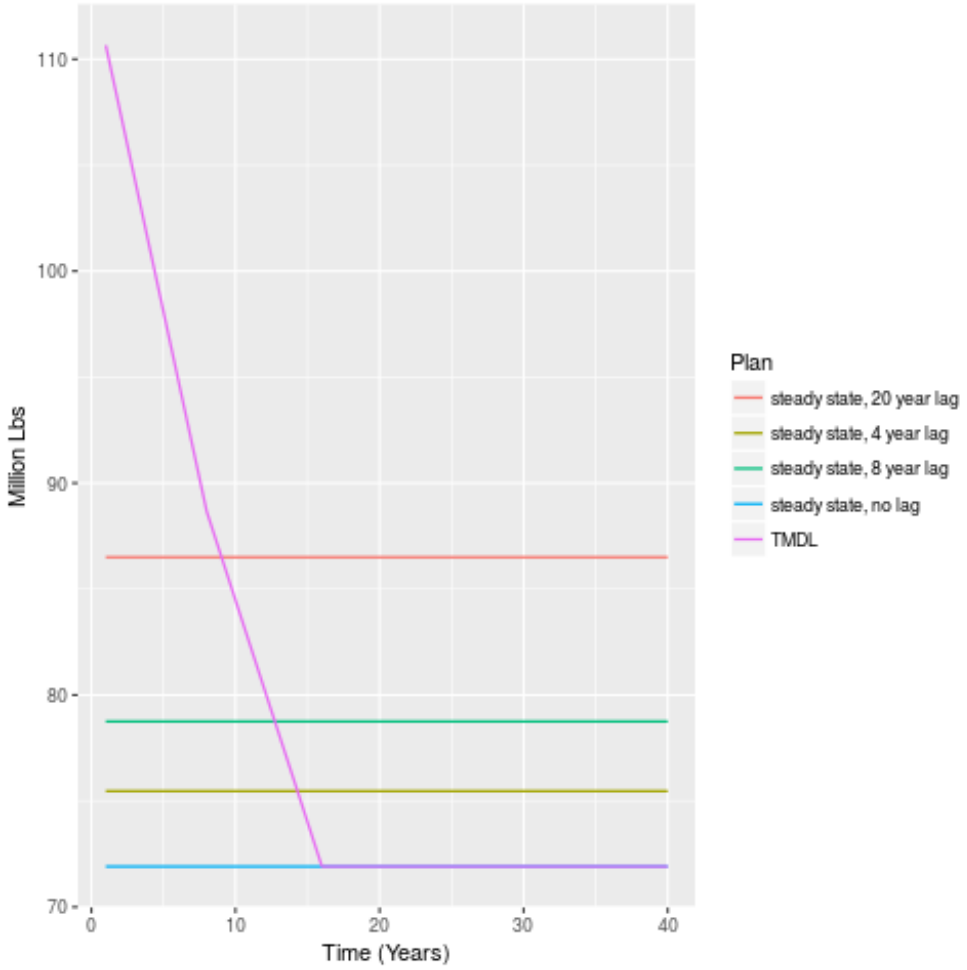

Figure 2.4. Time path of agricultural loads under TMDL and steady state plans

eventually run out around a lag length of 14 years and the excess costs of a dramatically high long run gap outweigh the benefits of being closer to the TMDL path in early periods. It is for these reasons that relative costs initially fall with lag length, reach a trough, and then increase.

At the other end of the cost spectrum, the steady state plan, while only imperfectly capturing the optimal path, produces costs that are within 0.05% of the optimal plan. Such plans would be highly efficient and relatively simple to implement—no time-varying policy would be required (although one would need a dynamic model to estimate the efficient steady state levels). The steady state plan is so efficient relative to the TMDL primarily because it wastes no time making early and aggressive nutrient reductions, even if they don't precisely track the optimal load time path. This result seems to indicate that a crude, time-invariant

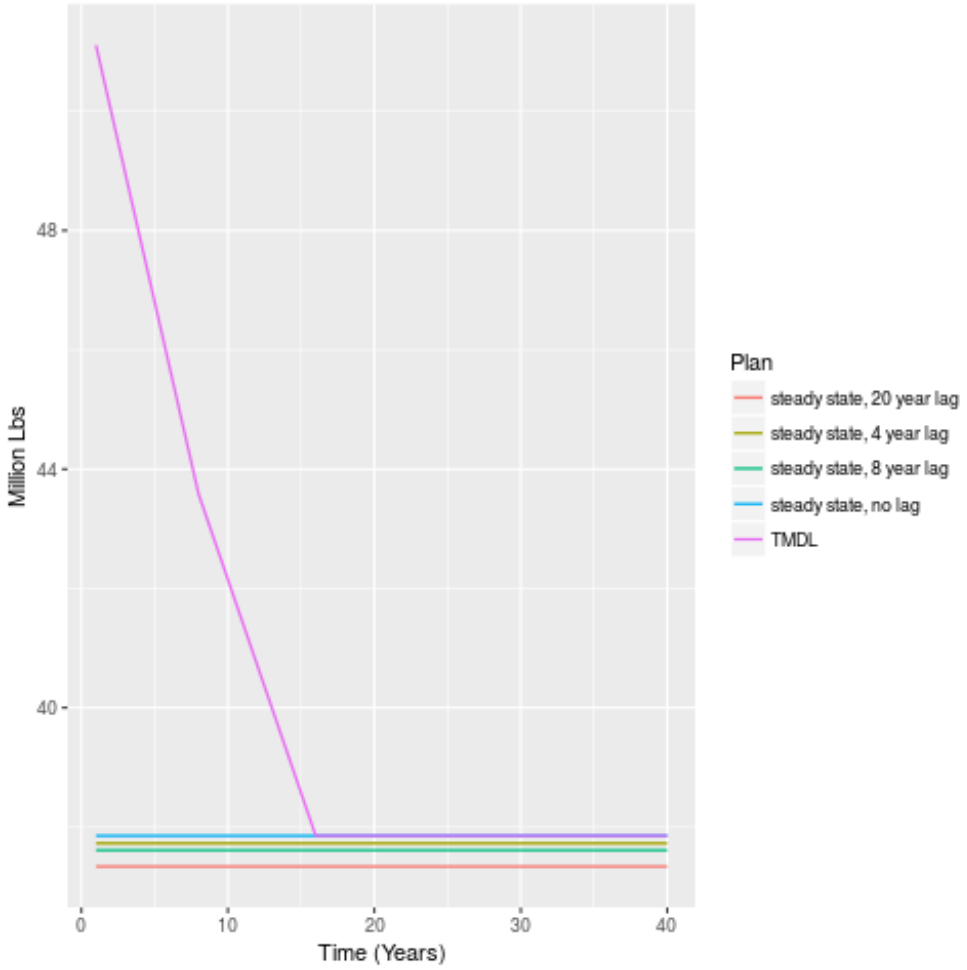

Figure 2.5. Time path of wastewater loads under TMDL and steady state plans

policy could provide nearly all the efficiency of the least-cost solution, without requiring the nearly the same complications associated with administering a time-varying policy.

Between these two extremes is a plan that acts like the optimal steady state plan, but is "blind" to the actual lag length, always following the load allocation given by the steady state solution to the no-lag scenario. The performance of this plan is excellent for short lag lengths, but for a lag length of 20 years, the costs become 3.8% higher than the optimal. The cost excess of this plan is essentially the cost of assuming the equivalency of wastewater and agricultural reductions when in fact the timing of their effects differ. This plan is costlier for long lag lengths because it is too lenient for wastewater loads and too strict for agricultural loads (since it assumes that reductions in agriculture will provide immediate ecological benefit). These results suggest that even a time-invariant plan that ignores lags

altogether might improve upon the recommendations of the TMDL because reductions under this plan are sufficiently aggressive in early periods.

2.6 Concluding Remarks

The essay began by asking whether the time path matters for crafting nutrient reduction plans. In principle, it does, though nailing the precise time path seems less quantitatively important than landing in the general vicinity of the path. Specifically, my analysis shows that implementing a time-invariant steady state policy, even if not exactly following the first-best path, can achieve within 0.05% of the least-cost solution under the prevailing conditions in the Chesapeake Bay.

Approaching the optimal steady state loads "from below" (as in Figures 2.1 and 2.2) is optimal whenever the pollution stock is critically high because it puts immediate downward pressure on the pollution stock level. In contrast, the Chesapeake Bay TMDL approaches the steady state loads "from above," perhaps with the aim of phasing in load reductions as smoothly as possible. There are at least two explanations for this design. First, calling for sharp and immediate nutrient load reductions likely would be met with fierce opposition from stakeholders. A gradual path perhaps represents a compromise between the polluters (on whom the burden of cleanup would fall) and those concerned with Chesapeake Bay health. Second, the existence of adjustment costs could make rapid load reductions economically undesirable. In the event that the cost of load reductions increase as the year-on-year *difference* in loads increases, a more gradual reduction path could be cost-effective. Examination of adjustment costs in the agricultural and wastewater sectors could help determine whether the optimal paths in Figures 2.1 and 2.2 are in fact optimal.

Even without taking account of these factors, the strategy of phasing in load reductions gradually each year until reaching the long run steady state did not produce a grossly inefficient cost outcome—my analysis found TMDL to be at most only 9% costlier than the first-best solution. By comparison, Kaufman, et al. [46] found that better targeting of agricultural best management practices (relative to the implementation plans recommended by the TMDL) could reduce agricultural sector abatement costs by 60%, and RTI International [78] found that opening up nutrient trading between significant point sources (wastewater) and agricultural nonpoint sources could reduce TMDL compliance costs by 36%. In this light, the allocation of abatement effort within and between sectors is perhaps a more pressing issue for managers than getting the exact time path correct. Even so, my results suggest

that further efficiency gains could be on the table if load timing is accounted for, though it remains to be seen whether adjustment costs and implementation costs would eliminate these gains. These questions remain for future research.

Chapter 3

Trade Ratio Design for Lagged Pollution

3.1 Introduction

There has been significant interest in cost-effective mechanisms to implement the Chesapeake Bay TMDL policy described in Chapter 1 [45]. Economic evaluations of the recommended sector-level nutrient reductions under this policy indicate the potential to meet the same load targets at much lower cost by reallocating nutrient reductions from high cost to low costs sources [78, 46]. The mechanism of "water quality trading" (WQT) represents one means through which these reallocations can occur. While the specifics vary by program, WQT generally involves the establishment of a market for pollution reduction "credits" where sources that reduce their pollution below their legal requirements generate credits which they can sell to other sources who purchase the credits in leiu of making their own (presumably costlier) pollution reductions. Despite concerns about whether such systems are consistent with the Clean Water Act [26, 13], EPA has officially endorsed WQT provided that water quality standards are maintained [75].

An important design choice for these trading programs involves setting the "trade ratio"—the rate at which excess reductions at one source are allowed to offset forgone reductions at another [40]. Allowing offsets to occur one-for-one may be undesirable given that discharges from various pollution sources often differ with respect to their impacts across space and time, as well as with respect to risk [38]. In general, trade ratios should be chosen to correct for the imperfect substitution of discharges at various sources. The problem of choosing an optimal trade ratio has been studied extensively as it pertains to managing trades between sources with heterogeneous spatial effects. Indeed, Montgomery's [58] seminal

piece on tradable pollution licenses described a system whereby emissions originating from different sources could trade according to their impacts at various receptors. Others have tried to improve upon this design [49, 25, 20] including Hung and Shaw [43] who adapted this system for trading water quality offsets between upstream and downstream sources along a river. Another carefully examined area has been the issue of choosing optimal ratios for trades between sources whose discharges are subject to different degrees of uncertainty [54, 37, 36, 39]. In chapters 1 and 2, I noted how point sources and nonpoint sources differ significantly with respect to how quickly their discharges arrive at the receiving waters. This imperfect temporal substitution of nutrient loads implies the need for an additional correction on trades between sources with unequal delays (on top of any corrections for spatial or risk factors). While differences in the risk profiles of pollution from agriculture versus pollution from the wastewater treatment sector remains an important concern, this essay focuses on the problem of designing trade ratios to account for differences across sources in the timing of pollution delivery relative to discharge.

I begin by developing a conceptual framework of pollution control over sources with varying lag lengths, noting that optimality requires separate "regimes" of control corresponding to sets of sources that deliver their pollution at the same time. I analyze the time paths of aggregate loads and individual discharges that are implied by the transitions between these regimes. Next, I examine two alternative market designs: 1) a forward market where the participants trade directly on pollution *deliveries* and 2) a trading ratio system where they trade on pollution *discharges*, which are then delivered at different points in time depending on each source's lag length. The forward market can, in principle, implement the first-best optimum, though we would expect the complexity of such a market to suppress trade considerably. I derive the time structure of permit prices that would result under a perfect forward market.

In the case of the trading ratio system, implementing the first best would be at least as administratively complex as doing so with a forward market, though I propose a second-best trade ratio design that balances the competing needs of correct for lags and maintaining simplicity. The mechanism involves market participants trading contemporaneous discharges according to the aggregate cap and trade ratios that prevail in the final regime of the first-best solution described above. Finally, I use a two-period, two-polluter model to examine the problem of choosing an optimal trade ratio prior to the arrival of the steady state. While first-best trade ratios will be greater than one when there are lag disparities between trading partners, second-best trade ratios under the same lag disparities may be less than one when

the overall cap on discharges is set sufficiently small. The optimal second best trade ratios must strike a balance between the abatement costs saved and the damage costs allowed.

3.2 Conceptual Framework

Suppose M point sources and N nonpoint sources contribute to the instantaneous pollution levels at time t. Let $WL(t)$ and $L(t)$ denote aggregate point source wasteloads and nonpoint source pollution loads, respectively. For the purposes of this analysis, the difference between point and nonpoint sources lie solely in the timing of their pollution delivery relative to discharge—discharges from each point source $j \in 1, 2, \ldots, M$ are delivered immediately, while those from each nonpoint source $i \in 1, 2, \ldots, N$ are delivered after a source-specific delay, l_i.[1] Henceforth, I refer to individual point and nonpoint sources as PS_j and NPS_i, respectively. Without loss of generality, let nonpoint sources be indexed such that NPS_1 has the shortest lag length, NPS_2 has the next-shortest and so on, with NPS_N having the longest lag.

Let $x_i(t)$ denote the quantity of pollution *discharged* from NPS_i at time t and delivered to the receiving waters at $t + \ell_i$. Let $w_j(t)$ denote the quantity of pollution discharged from PS_j at time t and delivered to the receiving waters instantaneously. Since pollution control measures in the nonpoint sector are incapable of affecting delivered pollution immediately, exogenous "legacy loads" associated with past discharges will be delivered in early periods. Let $\bar{x}_i(t)$ denote the legacy loads delivered from NPS_i at time t (discharged at $t-\ell_i$). Finally, let $b(t)$ denote natural background loads delivered at time t, which are included in the total nonpoint loads $L(t)$. The time structure of aggregate pollution delivery for $WL(t)$ and $L(t)$ is as follows:

$$WL(t) \quad = \quad \sum_{j=1}^{M} w_j(t) \quad \text{for } t \in [0, \infty) \tag{3.1}$$

[1] Undoubtedly, point and nonpoint source pollution differ in other ways (ease of monitoring, dependence on stochastic weather outcomes), but this paper focuses exclusively on their differences with respect to the timing of delivered pollution relative to the implementation of pollution control measures.

$$L(t) \;=\; \begin{cases} b(t) \;+\; \sum_{i=1}^{N} \bar{x}_i(t) & \text{for } t \in [0, \ell_1) \\ b(t) \;+\; \sum_{i=k}^{N} \bar{x}_i(t) \;+\; \sum_{i=1}^{k-1} x_i(t-\ell_i) & \text{for } k \in \{2,\ldots,N\}, t \in [\ell_{k-1}, \ell_k) \\ b(t) \;+\; \sum_{i=1}^{N} x_i(t-\ell_i) & \text{for } t \in [\ell_N, \infty) \end{cases} \tag{3.2}$$

Note in (3.2) that the delivery of nonpoint source loads between $t = 0$ and $t = \ell_1$ is entirely exogenous, consisting of natural background loads and the legacy loads from all N nonpoint sources. Starting at $t = \ell_1$, discharge (abatement) at time zero from the shortest lagged nonpoint source begins affecting delivered loads while the remaining $N - 1$ nonpoint sources continue delivering legacies from past discharges. With the arrival of each subsequent ℓ_i, the index k increases incrementally, and another nonpoint source exits the "legacy loads" term $\sum_{i=k}^{N} \bar{x}_i(t)$ to join the set of "managed loads" in the term $\sum_{i=1}^{k-1} \bar{x}_i(t-\ell_i)$. This proceeds until $t = \ell_N$ where discharges at time zero from the longest lagged source finally arrive and the discharge (abatement) choices of all nonpoint sources influence the size of load deliveries. Point source wasteloads and nonpoint source loads combine to produce the total loads, $TL(t)$. Formally,

$$TL(t) \;=\; WL(t) \;+\; L(t) \tag{3.3}$$

Thus far, expressions (3.1)–(3.3) only characterize the physical process of pollution delivery process. I next introduce the model's economic features. Let $g_j(w)$ represent PS_j's cost associated with any discharge level w, where $g_j' < 0$ (because costs increase as discharges fall) and $g_j'' > 0$ (because discharges become costlier to reduce at an increasing rate as discharges fall). Similarly, let $c_i(x)$ represent NPS_i's cost associated with any nonpoint discharge x, where $c_i' < 0$ and $c_i'' > 0$ for the same reasons as in g_j. Finally, let $D(TL_t)$ represent the pollution damage costs associated with delivered loads TL, where $D' > 0$ (because damages increase with total loads) and $D'' > 0$ (because damages increase with total loads at an increasing rate). Let Z represent the present value of the total costs of pollution and pollution cleanup over the continuous time horizon $t \in [0, \infty)$. Formally,

$$Z\big[w_j(t), x_i(t), TL(t)\big] \;=\; \int_0^{\infty} \Big\{ \sum_j g_j\big[w_j(t)\big] \;+\; \sum_i c_i\big[x_i(t)\big] \;+\; D\big[TL(t)\big] \Big\} e^{-rt}\, dt \tag{3.4}$$

where r is the discount factor. The regulator's problem is to minimize the total present value of pollution damage costs and pollution cleanup costs over time given the time structure of pollution delivery. This can be expressed

$$\min_{w_j(t),\, x_i(t)} Z\big[w_j(t), x_i(t), TL(t)\big] \quad \text{subject to (3.1)-(3.3)} \tag{3.5}$$

3.2.1 The First-Best Optimum

Nitrogen and phosphorus both accumulate in water systems when incoming loads exceed the rate at which the nutrients are either absorbed or flushed away [67]. This persistence would naturally call for a dynamic management plan where a planner regulates loads to modify the growth of a pollution stock [32]. To focus specifically on the problem of *lags* for optimal management, I assume no accumulation; rather, I regard pollution damages as relating only to the instantaneous flow of incoming pollution. In practice, nitrogen residence times in surface water tend to be relatively short (on the order of months, [9, 27]), making the no-accumulation assumption relatively benign. Analyzing the problem this way, the pollutant is not subject to stock dynamics, and therefore minimizing Z (the cumulative costs over the entire time horizon) merely requires minimizing the integrand (the instantaneous costs) at each t. The set of cost-minimizing $w_j(t)$ and $x_i(t)$ therefore must satisfy

$$-g_j'\big[w_j(t)\big] \;=\; D'\big[TL(t)\big] \qquad \forall j,\ t \geq 0 \tag{3.6a}$$

$$-c_i'\big[x_i(t)\big] \;=\; D'\big[TL(t+\ell_i)\big]e^{-r\ell_i} \qquad \forall i,\ t \geq 0 \tag{3.6b}$$

along with the relationships between w, x, and TL expressed in (3.1)-(3.3). Conditions (3.6a) and (3.6b) indicate that at the optimum, the marginal cost of reducing *discharges* from any source at any time must equal the time-discounted marginal pollution damages associated with the load *deliveries* corresponding to those discharges.

Whereas condition (3.6b) as written holds for $t \geq 0$, an equivalent way to express it is

$$-c_i'\big[x_i(t-\ell_i)\big]e^{r\ell_i} \;=\; D'\big[TL(t)\big] \qquad \forall i,\ t \geq \ell_i \tag{3.7}$$

The fact that the right hand sides of (3.6a) and (3.7) are the same allows for a direct comparison of their left hand sides, which reveals how the existence of lags affects the relative timing of optimal abatement effort for point versus nonpoint source discharges. Observe that

the time interval corresponding to the i^{th} condition in (3.7) depends explicitly on i,making it clear that the solution to (3.5) depends only on subsets of nonpoint polluters during early time intervals. The time structure of (3.7) implies that the optimal solution consists of $N+1$ pollution control "regimes" each of which corresponds to a unique set of sources whose discharges arrive at the same date. To see why these regimes exist, consider the fact that nonpoint sources cannot affect delivered loads prior to $t = \ell_1$, only point sources can. The first regime therefore consists solely of point sources choosing discharges to optimally manage $TL(t)$ during $t \in [0, \ell_1]$, given the levels of background loads and nonpoint legacy loads. Similarly, observe that only point sources and NPS_1 can affect delivered loads during $t \in [\ell_1, \ell_2)$. The second regime therefore involves choosing $w_j(t)$ for all j during $t \in [\ell_1, \ell_2)$ together with $x_1(t)$ during $t \in [0, \ell_2 - \ell_1)$ in order to optimize pollution deliveries during $t \in [\ell_1, \ell_2)$. Figure 1 illustrates how the optimal nonpoint discharges in the second regime (the bold lines) are offset in time by $-\ell_1$ relative to the point source discharge implying that NPS_1 belongs to the second regime from the outset. Following this reasoning, the third regime (the short dashed lines in Figure 3.1) involves point sources during $t \in [\ell_2, \ell_3)$, NPS_1 during $t \in [\ell_2 - \ell_1, \ell_3 - \ell_1)$ and NPS_2 during $t \in [0, \ell_3 - \ell_2)$. With each subsequent regime, the nonpoint source with the next-shortest lag length is added to the set of sources involved in the previous regime, and in general, the k^{th} regime involves all M point sources together with NPS_i for $i \in \{1, \ldots, k-1\}$. For $t = \ell_N$ and beyond, *managed* discharges from all nonpoint sources affect delivered loads at the receiving waters and therefore all $M+N$ sources are involved in choosing discharges to optimally manage deliveries in $t \geq \ell_N$. In this way, the $N^{th}+1$ discharge regime, in which discharges from all sources are jointly optimized, will eventually produce a steady state in load deliveries for $t \geq \ell_N$.

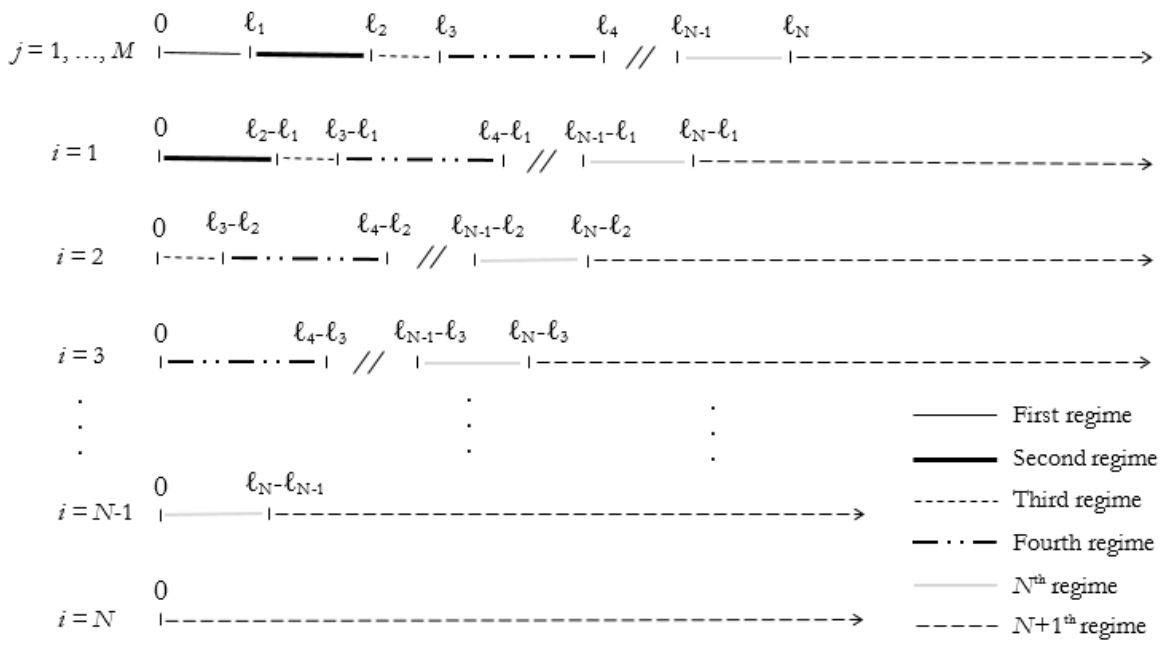

Figure 3.1. Regime timing for sources of different lag lengths

Letting $w_j^*(t)$ and $x_i^*(t)$ represent the point and nonpoint source discharges that solve (3.1)-(3.3) and (3.6), optimal aggregate load deliveries $TL^*(t)$ are, by definition

$$TL^*(t) \;=\; \begin{cases} \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=1}^{N} \bar{x}_i(t) & \text{for } t \in [0, \ell_1) \\ \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=k}^{N} \bar{x}_i(t) \;+\; \sum_{i=1}^{k-1} x_i^*(t-\ell_i) & \text{for } k \in \{2,\ldots,N\}, t \in [\ell_{k-1}, \ell_k) \\ \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=1}^{N} x_i^*(t-\ell_i) & \text{for } t \in [\ell_N, \infty) \end{cases} \tag{3.8}$$

At each t, aggregate load deliveries are a combination of managed discharges, legacy loads from nonpoint sources, and natural background loads.

The time paths of $TL^*(t)$, $w^*(t)$, and $x^*(t)$ undergo discrete jumps corresponding to transitions from one regime to the next. These jumps result from the fact that the minimum cost of aggregate load reductions is inherently a function of the number of sources participating in those reductions. Since regime $k+1$ adds a new source to those already participating in regime k, each regime is associated with a unique marginal abatement cost (MAC) curve.

Denote the aggregate MAC curve for regime k by MAC^k. Regime k's curve traces out the marginal cost of load reduction by load size, given that abatement is allocated among all sources in regime k at least cost. Aggregate MAC curves for select regimes are depicted in Figure 3.2's leftmost panel.

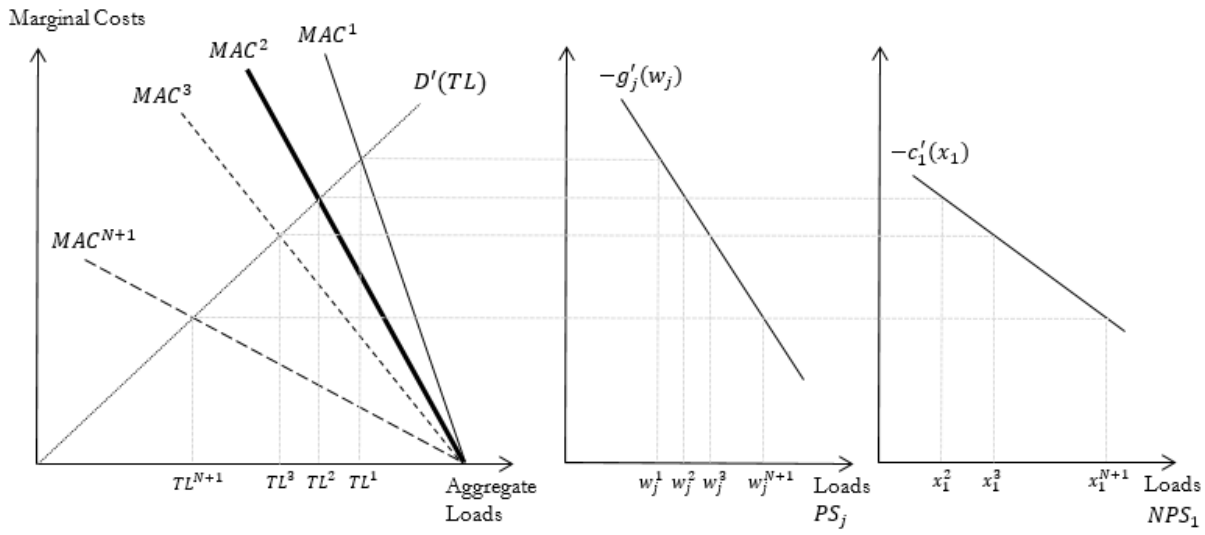

Figure 3.2. Time path of the aggregate marginal abatement cost curves and loads across regimes

To illustrate how optimal aggregate load deliveries and optimal source-level discharges evolve from one regime to the next, consider point source discharges in the first regime. Since nonpoint source deliveries consist solely of exogenous legacy loads during $t \in [0, \ell_1)$, the entire burden of load reduction during this interval falls on point sources, and the aggregate MAC curve during this interval, represented by MAC^1 in Figure 3.2, is as steep as it will ever be. Optimal aggregate loads in the first regime emerge where $MAC^1 = D'(TL)$ and these reductions will be shared optimally among point sources in accordance with condition (3.6a). Let TL^1 and w_j^1 denote the optimal regime 1 aggregate loads and the optimal regime 1 discharge level for PS_j, respectively. With the arrival of $t = \ell_1$, point sources transition to the second regime in which the burden of load reduction is shared among the M point sources and NPS_1, discharging at time zero. The fact that NPS_1 can contribute reductions to $TL(\ell_1)$ implies that the slope of the aggregate MAC curve suddenly becomes flatter than it was during $t < \ell_1$. This is depicted in Figure 3.2 by the movement from MAC^1 to MAC^2. Optimality requires loads be chosen such that $MAC^2 = D'(TL)$, which results in aggregate loads equal to TL^2, loads from PS_j equal to w_j^2, and loads from NPS_1 (discharged at $t = 0$) equal to x_1^2. At $t = \ell_1$, optimal aggregate load deliveries follow a discrete, downward jump

(as abatement suddenly becomes less costly), while optimal individual discharges from point sources follow a discontinuous, upward jump (as marginal pollution damages become less severe). These jumps are illustrated in Figures 3.3 and 3.4.

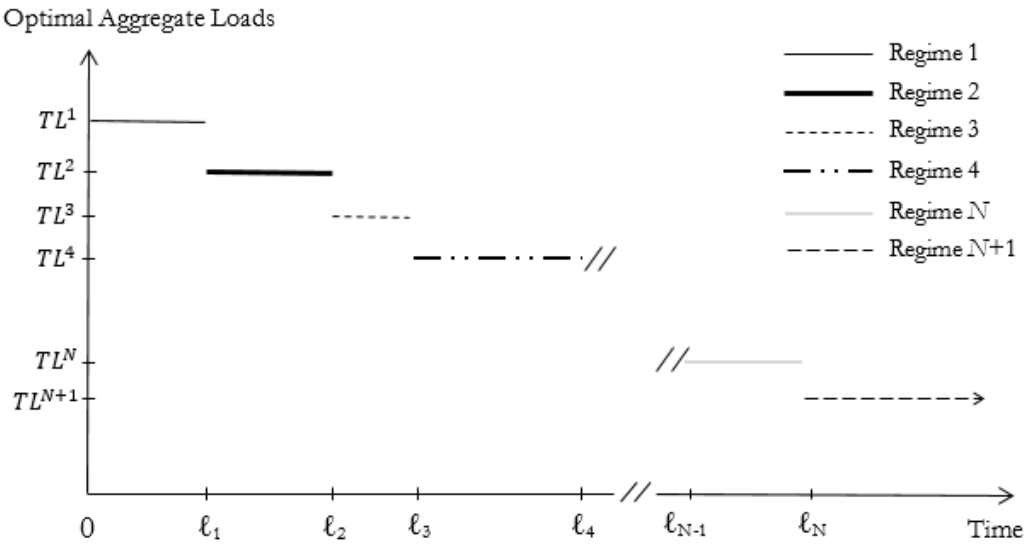

Figure 3.3. Time path of the optimal aggregate (delivered) loads

Similarly, when $t = \ell_2$ arrives, the reductions from NPS_2 begin affecting delivered loads, further lessening the burden of reductions on the sources involved in regime 2 and thereby reducing the cost of meeting any aggregate load target. In this transition to regime 3, the aggregate MAC curve shifts from MAC^2 to MAC^3, resulting in aggregate loads TL^3, loads from PS_j equal to w_j^3, and loads from NPS_1 (discharged at $t = \ell_2 - \ell_1$) equal to x_1^3. As happened at $t = \ell_1$, total load deliveries at $t = \ell_2$ experience a discontinuous, downward jump while individual discharges for PS_j and NPS_1 both jump upward discontinuously, the former at $t = \ell_2$ (Figure 3.4) and the latter at $t = \ell_2 - \ell_1$ (Figure 3.5).

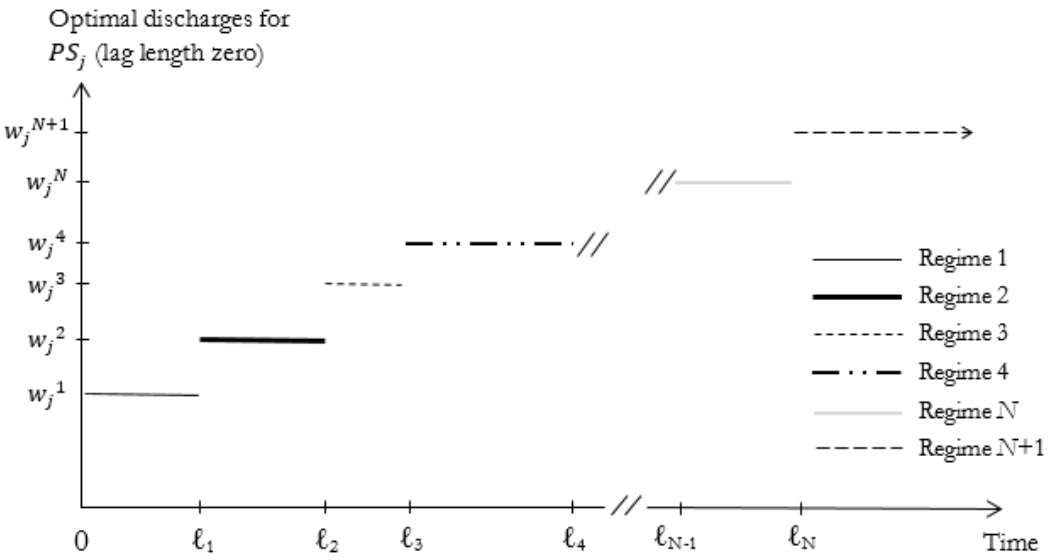

Figure 3.4. Time structure of optimal loads from each point source j

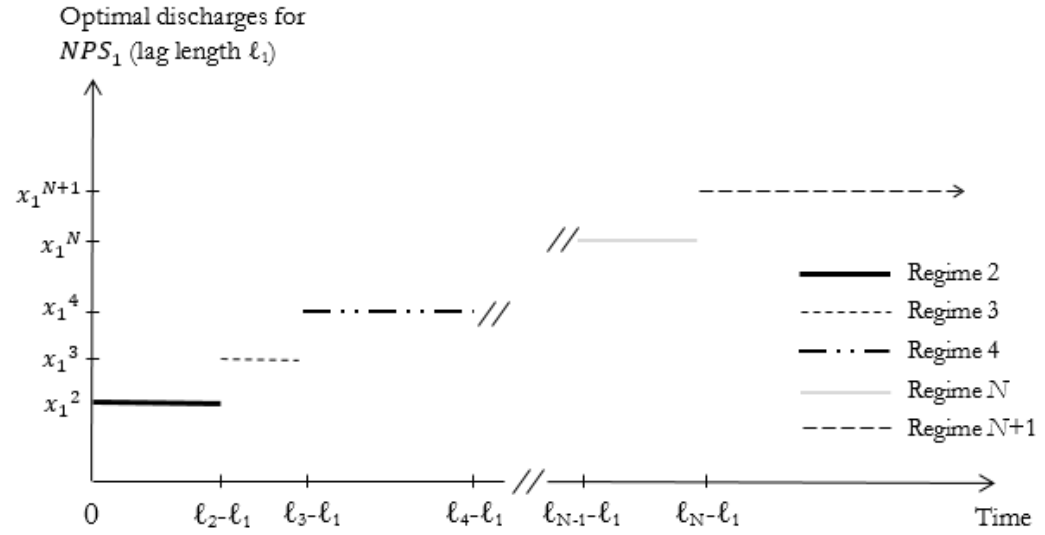

Figure 3.5. Time structure of optimal loads for nonpoint source $i = 1$

In general, jumps in the levels of optimal point source discharges will occur at $t = \ell_i$ for $i \in 1, \ldots, N$ (Figure 3.4) since the arrival of the each $t = \ell_i$ coincides with the transition to the next regime. Jumps in the levels of optimal discharges for nonpoint source $i = 1$ follow the same structure except the first regime is omitted and the remaining jumps are shifted back in time by ℓ_1 relative to point sources[2] (Figure 3.5). Figure 3.6 illustrates the same

[2]For example, NPS_1's transition from the second to the third regime, rather than occurring at ℓ_2 as it

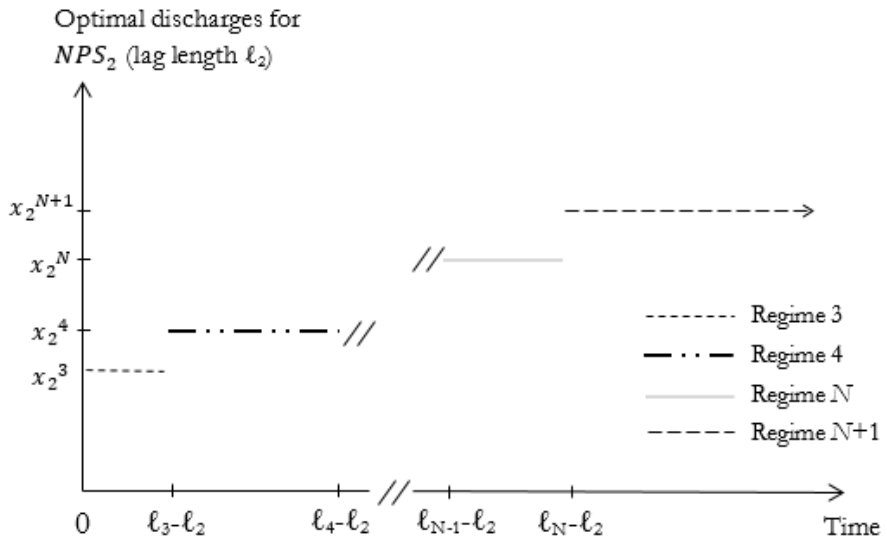

Figure 3.6. Time structure of optimal loads for nonpoint source $i = 2$

pattern for NPS_2, which begins the planning period as part of regime 3 and has the timing of its jumps shifted back by ℓ_2 relative to point sources. Broadly, any nonpoint source i will begin the planning period in regime $k = i + 1$, and will transition from the k^{th} to the $k^{th} + 1$ regime at $t = \ell_k - \ell_i$.

[Note somewhere that the load levels are only flat within a regime if background loads, nonpoint legacy loads, and cost functions are constant over the relevant t. Generally, optimal TL will be nonincreasing over time and, assuming some pollution control at each nonpoint source is preferred to none, will decrease in discontinuous fashion at $t = \ell_i$ for all i (Figure 3.3). Generally, optimal w_j and x_i will be nondecreasing over time and again assuming some pollution at each nonpoint source is preferred to none, w_j will increase in discontinuous fashion at $t = \ell_i$ for all i and x_i will increase in discontinuous fashion at $t = \ell_k - \ell_i$ for $k \in \{i + 1, \ldots, N\}$.]

Conditions (3.6a) and (3.7) imply that the present value of marginal abatement costs for loads delivered in period t must be equalized across all sources. Formally, the optimal discharge allocation must have

$$g_j'\big[w_j(t)\big] \;=\; g_{j'}'\big[w_{j'}(t)\big] \qquad \forall j,\ j',\ t \geq 0 \tag{3.9a}$$

$$g_j'\big[w_j(t)\big] \;=\; c_i'\big[x_i(t-\ell_i)\big]e^{r\ell_i} \qquad \forall j,\ i,\ \ t \geq \ell_i \tag{3.9b}$$

does for point sources, occurs instead at $\ell_2 - \ell_1$.

$$c_i'\big[x_i(t)\big] \;=\; c_{i'}'\Big\{x_{i'}\big[t-(\ell_{i'}-\ell_i)\big]\Big\}e^{r(\ell_{i'}-\ell_i)} \qquad \forall i,\ i'>i,\ t\geq \ell_{i'}-\ell_i \tag{3.9c}$$

Condition (3.9a) corresponds to the optimal point source allocation where, at the optimum, marginal costs of contemporaneous load reductions must be equal across all point sources. Condition (3.9b) corresponds to the optimal allocation between any point source j and any nonpoint source i. Optimality across the point and nonpoint source sector requires that the marginal cost of reducing discharges from any point source j at time t be equal to the marginal cost of reducing discharges from any nonpoint source i at time $t-\ell_i$, adjusted by the discount factor $e^{r\ell_i}$. Note that because point source lag times are zero, the ℓ_i implicitly represents the lag time *differential* between nonpoint source i and point source j. Condition (3.9c), which deals with the allocation between two nonpoint sources of different lag lengths, has the same form and interpretation as (3.9b), except ℓ_i is replaced by $\ell_{i'}-\ell_i$, the lag time differential between source i' and i.

Whereas conditions (3.9a)-(3.9c) describe the optimal allocation of discharges that are scheduled for delivery in the same period, an alternative way to think about the solution to (3.5) is in terms of the allocation of contemporaneous discharges between each source. First consider point sources. Since their lag times are identical, equating each source's marginal abatement costs with their corresponding marginal damage costs as in (3.6a) will imply the equivalence of marginal abatement costs among all M point sources. This condition has already been defined in (3.9a) and is the standard result for a cost-effective pollution control allocation. Next, consider the allocation between sources whose lag times differ. The allocation of contemporaneous discharges between any two sources must satisfy

$$g_j'\big[w_j(t)\big] \;=\; c_i'\big[x_i(t)\big]\frac{D'\big[TL(t)\big]}{D'\big[TL(t+\ell_i)\big]}e^{r\ell_i} \qquad \forall j,\ i,\ t\geq 0 \tag{3.10a}$$

$$c_i'\big[x_i(t)\big] \;=\; c_{i'}'\big[x_{i'}(t)\big]\frac{D'\big[TL(t+\ell_i)\big]}{D'\big[TL(t+\ell_{i'})\big]}e^{r(\ell_{i'}-\ell_i)} \qquad \forall i,\ i',\ t\geq 0 \tag{3.10b}$$

which follow directly from first order conditions (3.6a) and (3.6b). In (3.10), instead of simply equating marginal abatement costs among these sources, the equations are modified to account for the imperfect substitution of abatement from sources with different lag lengths. In (3.10a) the marginal abatement costs at point source j at t must equal the marginal abatement costs at nonpoint source i at t adjusted by a factor that embodies the differential impact each source has on pollution damage costs due to the lag length discrepancy. Specifi-

cally, this adjustment factor consists of the ratio of marginal damage costs from point source discharges (sustained at t) to the time-discounted marginal damage costs from nonpoint discharges (sustained at $t + \ell_i$). The adjustment factor on the right hand side of (3.10b) is conceptually the same—it is the ratio of the time-discounted marginal damage cost of source i's discharges (sustained at $t+\ell_i$) to the time-discounted marginal damage cost of source i''s discharges (sustained at $t + \ell_{i'}$).

A cost-effective allocation of discharges among a set of polluters implies equalization of marginal abatement costs across all polluters. Note that this is true among the point sources but not between point and nonpoint sources nor among nonpoint sources (since each is assumed to have a unique lag length). The modifications to the standard least-cost conditions described in (3.10) hint at the rationale for applying trade ratios between sources with different lag lengths, namely the nonequivalence of marginal damage costs for discharges that are subject to different degrees of delay.

3.2.2 Markets for Pollution Deliveries (A Forward Market Approach)

One way of achieving the optimal delivered loads given by (3.8) would be to set a cap on aggregate loads delivered at each t equal to $TL^*(t)$ and allow firms to reallocate *delivered* loads among themselves. Allowances would be issued to firms at each t such that the sum of pollution deliveries in any t is no greater than $TL^*(t)$. This market structure would imply that any firm could trade pollution reductions with another provided their emissions had the same delivery date. Two firms with identical lag lengths could trade contemporaneous discharges, or alternatively, two firms whose lag lengths differ by ℓ could swap reductions in period t for reductions in period $t + \ell$. Assuming that this tradable allowance system eliminates gains from trade, the equilibrium under this market design will result in

$$-g_j'\big[w_j(t)\big]e^{-rt} \;=\; p(t) \qquad \forall j,\ t \geq 0 \tag{3.11a}$$

$$-c_i'\big[x_i(t-\ell_i)\big]e^{-r(t-\ell_i)} \;=\; p(t) \qquad \forall i,\ t \geq \ell_i \tag{3.11b}$$

where $p(t)$ is the price of the right to deliver nutrients at time t. Note that the allocation of discharges among and across point and nonpoint sources will be cost-effective under this market design since (3.11a) and (3.11b) imply the conditions in (3.9). Provided the time-specific caps on aggregate delivered loads match those in (3.8), the discharge levels that

satisfy equilibrium under this market structure will match those in (3.6) and the permit market (in which pollution sources trade permits in *delivered* loads) implements the first-best solution.

Assuming gains from trade are exhausted after sources reallocate the initial allowance distribution, the price of an allowance will converge on the level of aggregate MAC at the total cap. As explained in the previous section and as depicted in Figure 3.2, aggregate MAC begins the management period steep (as only point sources can contribute abatement) and flattens out with time as the managed discharges from nonpoint sources start affecting delivered loads in later periods. Even though the optimal load caps decrease in each new regime, the cost savings more than make up for these stricter targets such that the aggregate MAC at the optimal load target in each new regime is smaller than in the previous regime. This implies that allowance prices will follow the same type of discontinuous time path as optimal loads, with jumps occurring at $t = \ell_i$ for all i. Prices will be high in the beginning of the planning horizon, due to the relative scarcity of sources with the ability to affect delivered loads, and will drop with the arrival of each new regime (Figure 3.7).

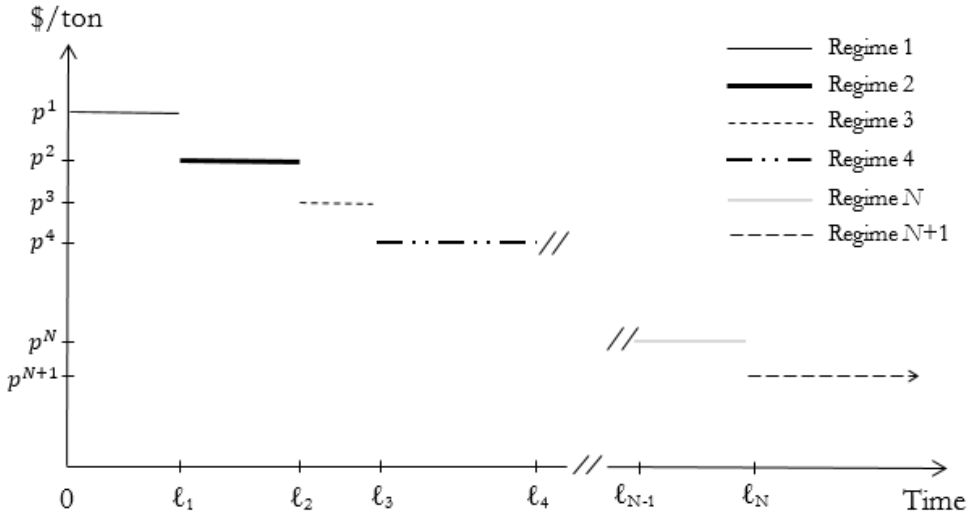

Figure 3.7. Time path of permit prices under the optimal cap

To implement the optimal market, the regulator must know the lag lengths of all N nonpoint sources and set $N + 1$ separate caps each applying to the intervals that correspond to each unique pollution control regime. Such a system would become administratively cumbersome. A regulator might approximate the vastly complicated lag structure by grouping nonpoint sources into a small number of lag length bins and set time-specific caps for this

simplified pollution delivery structure. In practice, transactions would resemble forward contracts with a seller agreeing to implement some BMP in period t, estimated to deliver x pounds of pollution reduction at some future date $t + \ell$ (dictated by the lag length of their pollution delivery process) and the buyer purchasing the right to increase pollution discharges above their permitted levels by x pounds at $t + \ell$.

This type of contract may be problematic in the context of nutrient pollution control for two reasons. First, the commodity that the seller is providing at time t (i.e. the amount of "delivered" pollution reduction) is not well-defined. The complex relationship between nutrient control measures performed on agricultural land and the ultimate timing and amount of pollution deliveries makes this so. Defining the commodity as "estimated nutrient reductions" as many existing trading programs do (e.g. [62]) is one way around this problem, although uncertainty remains as to whether future regulations could become more strict if water quality goals fail to be achieved on schedule. Regulators' affinity for this type of "adaptive management" may leave point sources uncertain as to whether the nonpoint reductions they purchase in the present will guarantee them the right to increase their future discharges. Pollution delivery uncertainty may spawn regulatory uncertainty.

Second, even if nonpoint pollution reductions can be delivered reliably, a TMDL may require point sources to make reductions sooner than reductions from nonpoint sources can be delivered. To satisfy these requirements, point sources may need to make long-lived investments in nutrient removal technologies that could render the future reductions in delivered pollution from nonpoint sources unnecessary. Allowing point and nonpoint sources to trade contemporaneous discharges according to some lag-specific trade ratio could open the door for point source abatement costs savings while accounting for the fact that point and nonpoint reductions are not ecologically equivalent (due to lag-length disparities). I discuss this system next.

3.2.3 Markets for Pollution Discharges (A Trading Ratio Approach)

Instead of prohibiting sources with different delivery dates from trading pollution reductions with one another, suppose these trades are allowed provided they are not one-for-one. In principle, the correct trade ratio should require the lagged source to reduce pollution in excess of the quantity that they are offsetting to account for the fact that reductions from lagged sources will provide environmental benefits later in the future (making them economically less valuable). This system would place a cap on the aggregate amount of pollution that can

be *discharged* at any point in time, but would be indifferent to how these discharges were allocated among the polluting firms. Firms with high reduction costs could pay low-cost firms to make reductions on their behalf, reducing overall control costs while maintaining aggregate discharges at a constant level. Using the socially optimal pollution discharges as a guide, let caps on discharges in period t, $TD(t)$, be based on the optimal discharges from section 3.2.1, where

$$TD^*(t) \;=\; \sum_j w_j^*(t) \;+\; \sum_i x_i^*(t)$$

The market equilibrium under this trading system is given by the solution to

$$\min_{x_i(t),\, w_j(t)} \int_0^\infty \Big\{ \sum_j g_j\big[w_j(t)\big] \;+\; \sum_i c_i\big[x_i(t)\big] \Big\} e^{-rt}\, dt$$

subject to

$$\sum_j w_j(t) \;+\; \sum_i x_i(t) \;\le\; TD^*(t) \quad \forall t \tag{3.12}$$

with corresponding Lagrangian expression

$$\mathcal{L} \;=\; \int_0^\infty \Big\{ \sum_j g_j\big[w_j(t)\big]\, e^{-rt} + \sum_i c_i\big[x_i(t)\big] e^{-rt} + \lambda(t)\Big[\sum_j w_j(t) + \sum_i x_i(t) - TD^*(t)\Big] \Big\}\, dt$$

Optimal discharges satisfy

$$-g_j'\big[w_j(t)\big] e^{-rt} \;=\; \lambda(t) \qquad \forall j,\; t \ge 0$$

$$-c_i'\big[x_i(t)\big] e^{-rt} \;=\; \lambda(t) \qquad \forall i,\; t \ge 0$$

implying that

$$g_j'\big[w_j(t)\big] \;=\; g_{j'}'\big[w_{j'}(t)\big] \quad \forall j,\; j',\; t \ge 0$$

$$g_j'\big[w_j(t)\big] \;=\; c_i'\big[x_i(t)\big] \qquad \forall j,\; i,\quad t \ge 0$$

$$c_i'\big[x_i(t)\big] \;=\; c_{i'}'\big[x_{i'}(t)\big] \qquad \forall i,\; i',\; t \ge 0$$

This outcome differs from (3.9) in two ways. First, marginal costs between the two sources are being evaluated on contemporaneous discharges, whereas in (3.9b) and (3.9c), the evaluation occurs on lag-adjusted discharges—marginal reduction costs for PS_j at t are being compared with marginal reduction costs for NPS_i at $t - \ell_i$ and marginal reduction costs for NPS_i

at t are being compared to marginal reduction costs for $NPS_{i'}$ at $t-(\ell_{i'}-\ell_i)$. Second, whereas nonpoint marginal reduction costs in (3.9b) and (3.9c) are inflated by the continuous time discount factor $e^{r\ell}$ (where ℓ denotes the lag discrepancy between the sources), no such adjustment is applied to nonpoint source costs under this discharge trading system. A regulator could fix this second issue by applying a trade ratio to nonpoint source i's discharges equal to $e^{r\ell_i}$, meaning that for every pound of pollution increased at a point source, its nonpoint trading partner would have to reduce its discharges $e^{r\ell_i}$ pounds. Formally, this design choice would require modifying the cap on discharges, turning (3.12) into

$$\sum_j w_j(t) \;+\; \sum_i x_i(t)e^{-r\ell_i} \;\leq\; TD^*(t) \quad \forall t \tag{3.13}$$

The first order conditions for the new Lagrangian expression are

$$\begin{aligned} -g_j'\big[w_j(t)\big]e^{-rt} &= \lambda(t) \qquad \forall j,\ t\geq 0 \\ -c_i'\big[x_i(t)\big]e^{-r(t-\ell_i)} &= \lambda(t) \qquad \forall i,\ t\geq 0 \end{aligned}$$

implying that the allocation of discharges between any point source j and nonpoint source i and any nonpoint source i and i' is given by

$$\begin{aligned} -g_j'\big[w_j(t)\big] &= -c_i'\big[x_i(t)\big]e^{r\ell_i} \qquad\qquad \forall j,\ i,\ t\geq 0 \\ -c_i'\big[x_i(t)\big] &= -c_{i'}'\big[x_{i'}(t)\big]e^{r(\ell_{i'}-\ell_i)} \quad \forall i,\ i',\ t\geq 0 \end{aligned}$$

These conditions will match (3.9b) and (3.9c) as long as $x_i(t)=x_i(t-\ell_i)$ for all i. Because legacy pollution from nonpoint source i is still being delivered until $t=\ell_i$, NPS_i's deliveries at t will exceed its discharges until $t=\ell_i$. Provided that NPS_i discharges at a constant rate, $x_i(t)$ will then equal $x_i(t-\ell_i)$ for all $t\geq\ell_i$. The earliest this can be true for *all* sources is $t=\ell_N$, again, provided each source discharges at a constant rate from the start of the management period. These facts imply that if a permit market is established at the outset in which simple adjustment factors are applied to nonpoint source discharges, as specified in the modified discharge cap (3.13), this system will implement the first-best solution for $t\geq\ell_N$. The discharges that result from this trading equilibrium will equal optimal discharges in the $N^{th}+1$ regime. These discharges will be suboptimal for $t<\ell_N$, but optimal thereafter. Getting the system of trade ratios to implement the first-best for all t would require (like the forward market system) establishing $N+1$ time-specific load caps and $N+1$ sets of time- and

source-specific adjustment factors to allocate discharges appropriately among the sources. Given the complexity of the actual nonpoint delivery process, these parameters would be enormously difficult to determine. In the next section, I analyze a simple two-polluter, two period problem to illustrate how an optimal pre-steady state trade ratio system would be designed in principle.

3.3 Two Polluter, Two Period Problem

Consider pollution originating from one point source whose discharges, w, deliver immediately and one nonpoint source whose discharges, x, deliver with an ℓ-period delay. As before, let the costs associated with these discharge levels be given by $g(w)$ and $c(x)$ and let TL_0 and TL_ℓ denote total loads delivered in period 0 and ℓ. Formally,

$$TL_0 \;=\; \bar{x} \;+\; w \tag{3.14a}$$

$$TL_\ell \;=\; x \;+\; \bar{w} \tag{3.14b}$$

where $\bar{x}$ represents exogenous deliveries of legacy loads from nonpoint and $\bar{w}$ represents exogenous point source discharges in period $t+\ell$. The regulator's problem is

$$\min_{w,\,x} \; g(w) \;+\; c(x) \;+\; D(TL_0) \;+\; D(TL_\ell)\delta^\ell \qquad \text{subject to (3.14)}$$

where $\delta = \frac{1}{(1+r)}$ is the discount factor. The optimal w and x must satisfy $-g'(w) = D'(TL_0)$ and $-c'(x) \;=\; D'(TL_\ell)\delta^\ell$, meaning that marginal pollution control costs at each source are balanced against the present value of the marginal damage costs associated with each discharge. Denote these first-best load allocations w^* and x^*. Note from the first order conditions that x^* depends on ℓ, whereas w^* is independent of ℓ. The trade ratio and discharge cap that implements this first-best solution under a particular ℓ must induce the lag-specific x^* while maintaining the w^* that would prevail under any lag length.

Consider a regulatory mechanism that establishes a trade ratio, dictating the rate of substitution between reductions at point and nonpoint sources, sets a cap on total pollution discharge denominated in terms of one or the other source, and gives permission to the sources to reallocate discharges among themselves subject to the trade ratio and the cap.

Under this market design, the polluters choose loads to

$$\min_{w,x} \; g(w) \; + \; c(x) \quad \text{subject to} \quad x \; + \; w\psi \; = \; TD \tag{3.15}$$

where ψ represents the trade ratio (denominated in pounds of nonpoint loads per pound of point loads) and TD represents the cap on total discharges (denominated in nonpoint loads). The Lagrangian expression that corresponds to (3.15) is

$$\mathcal{L} \; = \; g(w) \; + \; c(x) \; + \; \lambda\Big[x \; + \; w\psi \; - \; TD\Big]$$

Assuming the two sources exhaust gains from trade, they will reallocate discharge permits until

$$g'(w) = c'(x)\psi$$

Let $\tilde{w}(\psi, TD)$ and $\tilde{x}(\psi, TD)$ denote each polluter's equilibrium discharges for any ψ and TD. Using this trading equilibrium condition $g'(w) - c'(x)\psi \; = \; 0$, the credit balancing condition $x + w\psi - TD = 0$, and the implicit function theorem (Mas Colell, Whinston, and Green, 1995) the changes in the equilibrium loads with respect to ψ are

$$\frac{\partial \tilde{w}}{\partial \psi} \; = \; \frac{c'(\tilde{x}) \; - \; \tilde{w}\, c''(\tilde{x})\, \psi}{g''(\tilde{w}) \; + \; c''(\tilde{x})\, \psi^2} \qquad \text{and} \qquad \frac{\partial \tilde{x}}{\partial \psi} \; = \; \frac{-c'(\tilde{x})\, \psi \; - \; \tilde{w}\, g''(\tilde{w})}{g''(\tilde{w}) \; + \; c''(\tilde{x})\, \psi^2}$$

and the changes in equilibrium loads with respect to TD are

$$\frac{\partial \tilde{w}}{\partial TD} \; = \; \frac{c''(\tilde{x})\, \psi}{g''(\tilde{w}) \; + \; c''(\tilde{x})\psi^2} \qquad \text{and} \qquad \frac{\partial \tilde{x}}{\partial TD} \; = \; \frac{g''(\tilde{w})}{g''(\tilde{w}) \; + \; c''(\tilde{x})\, \psi^2}$$

The derivatives of $\tilde{w}$ and $\tilde{x}$ with respect to TD are unambiguously positive—increasing total allowable loads produces load increases at both sources under this trading system. The derivative of $\tilde{w}$ with respect to ψ is unambiguously negative—requiring greater nonpoint load reductions to offset a given point load increase will lead to smaller loads (larger reductions) from the point source. However, the sign of $\frac{\partial \tilde{x}}{\partial \psi}$ is less straightforward. Since $-c'(\tilde{x})\, \psi$ and $\tilde{w}\, g''(\tilde{w})$ are both positive, $\frac{\partial \tilde{x}}{\partial \psi}$ may be positive or negative depending on their relative magnitudes. Figure 3.8 illustrates the features of $\frac{\partial \tilde{w}}{\partial \psi}$ and $\frac{\partial \tilde{x}}{\partial \psi}$ for one particular case. In the figure, point source loads decrease monotonically as the trade ratio goes up, while nonpoint loads decrease before eventually turning back upward as the trade ratio climbs. This ambiguous effect of ψ on nonpoint loads is due to the implicit relationship between ψ and the TD.

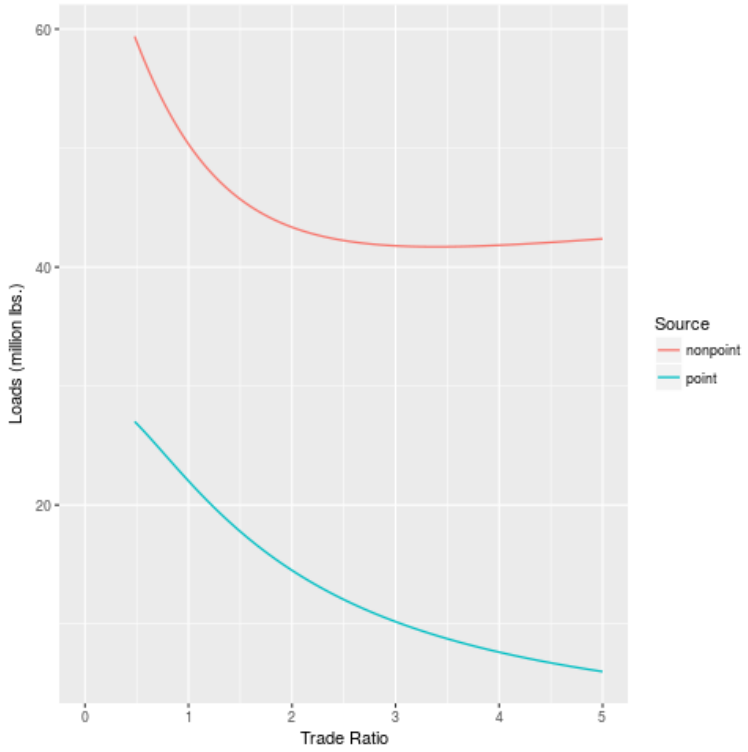

Figure 3.8. Loads under the optimal cap for various trade ratios

Recall the constraint in problem (3.15) (the discharge cap) and note how the point source sector's usage of the cap is given by the product of w and ψ. For large values of ψ, point source loads may shrink such that the overall size of $w\psi$ may decrease, leaving a larger share of cap left for nonpoint loads. In this way, changes in ψ produce both a *relative price effect* (shifting load reductions toward point sources) and an *endowment effect* (relaxing the cap and thereby increasing loads at both sources).

3.3.1 The First Best Regulation

Given the equilibrium outcome of a discharge trading system under any choice of trading ratio ψ and cap TD, consider next the optimal choice of ψ and TD. Formally, a regulator would select these parameters to

$$\min_{\psi, TD} \; g\big[\tilde{w}(\psi, TD)\big] \; + \; c\big[\tilde{x}(\psi, TD)\big] \; + \; D(TL_0) \; + \; D(TL_\ell)\delta^\ell \quad \text{subject to (3.14)}$$

The optimal values of ψ and TD must jointly satisfy

$$\left\{g'\left[\tilde{w}(\psi, TD)\right] \; + \; D'(TL_0)\right\}\frac{\partial \tilde{w}}{\partial \psi} \; + \; \left\{c'\left[\tilde{x}(\psi, TD)\right] \; + \; D'(TL_\ell)\delta^\ell\right\}\frac{\partial \tilde{x}}{\partial \psi} \; = \; 0 \quad (3.16a)$$

$$\left\{g'\left[\tilde{w}(\psi, TD)\right] \; + \; D'(TL_0)\right\}\frac{\partial \tilde{w}}{\partial TD} \; + \; \left\{c'\left[\tilde{x}(\psi, TD)\right] \; + \; D'(TL_\ell)\delta^\ell\right\}\frac{\partial \tilde{x}}{\partial TD} \; = \; 0 \quad (3.16b)$$

Since $\frac{\partial \tilde{w}}{\partial \psi}$ will not equal $\frac{\partial \tilde{w}}{\partial TD}$ in general and $\frac{\partial \tilde{x}}{\partial \psi}$ will not equal $\frac{\partial \tilde{x}}{\partial TD}$ in general, the terms in brackets must vanish to guarantee that both of these conditions are satisfied. This corresponds to the first order conditions of the social cost minimization problem above. Combining these first order conditions with the trading equilibrium condition indicates that at the optimum

$$\psi^* \; = \; \frac{D'(TL_0)}{D'(TL_\ell)\delta^\ell} \quad (3.17)$$

which states that the optimal rate of substitution between point and nonpoint discharges is exactly the ratio of the marginal damage costs of point source discharges to the time-discounted marginal damage costs of nonpoint discharges. This ensures that the relative allocation is correct, but TD must also be chosen to ensure that the levels of each discharge be correct. With both the trade ratio and the overall cap at their disposal, a regulator can, in theory, adjust both to implement the first-best discharges w^* and x^*. Because the discount factor δ is less than one, the optimal trade ratio, ψ^*, is positively related to lag length. Recall however that a) as ψ increases, point source loads in the trading equilibrium fall, and b) w^* is lag invariant. To accommodate a larger ψ while maintaining constant point source discharges, the optimal cap TD^* must increase in tandem with ψ^*. Figure 3.9 illustrates the relationship between nonpoint lag length and the optimal pair of trade ratio and cap for a particular set of abatement costs and damage cost functions. To make room for higher optimal nonpoint loads under a longer lag lengths, the regulator must adjust the cap upward (increasing loads at both sources) and then increase the trade ratio enough to bring point source loads back down to their previous level.

3.3.2 A Second Best Context

Since pollution damage costs are highly uncertain, regulators often choose a limit on total allowable pollution (perhaps based on biological criteria) and aim to meet this limit in the most cost effective way. Horan and Shortle [37] analyze this type of scenario in the context

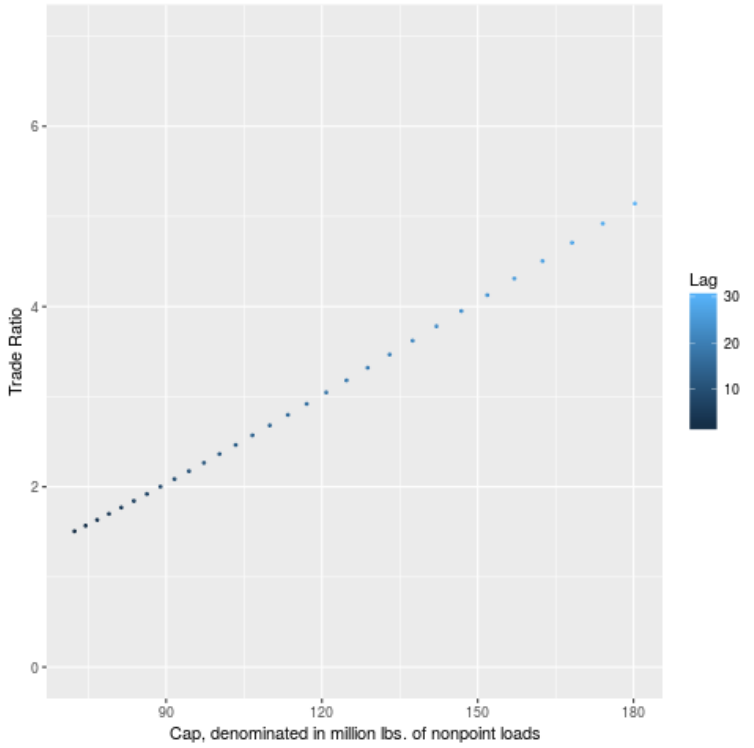

Figure 3.9. Optimal pair of trade ratio and cap for lag lengths of 1 to 30 years

of nutrient loads in the Susquehanna River basin. Along these lines, consider a discharge cap fixed exogenously at $\hat{TD}$ with the regulator seeking to minimize social costs solely through the choice of ψ. This time the optimal ψ must satisfy (3.16a) for $TD = \hat{TD}$. Here, the first-best outcome can be achieved only if the cap is set at TD^*. In the event that $\hat{TD} \neq TD^*$, the bracketed terms in (3.16a) no longer vanish and the second best optimal trade ratio, $\hat{\psi}$, obtains by substituting $c'(x)\psi$ for $g'(w)$ (from the trading equilibrium) in the left-hand side of (3.16a):

$$\hat{\psi} \;=\; \frac{-\Big[c'(x) \;+\; D'(TL_\ell)\delta^\ell\Big]\frac{\partial x}{\partial \psi} \;-\; D'(TL_0)\frac{\partial w}{\partial \psi}}{c'(x)\frac{\partial w}{\partial \psi}} \tag{3.18}$$

Rather than directly equating marginal abatement costs and marginal damage costs for each type of load separately[3] as would occur in a first-best context, the solution to the second-best problem strikes a balance between abatement cost savings (associated with shifting loads from nonpoint to point sources) and damage cost savings (associated with increasing

[3]Since $g'(w) < 0$, it essentially represents the benefits of discharging loads equal to w. The numerator and denominator on the right-hand side of (3.18) is therefore the damage costs associated with each type of load *net of the cost savings* associated with discharging loads of that size

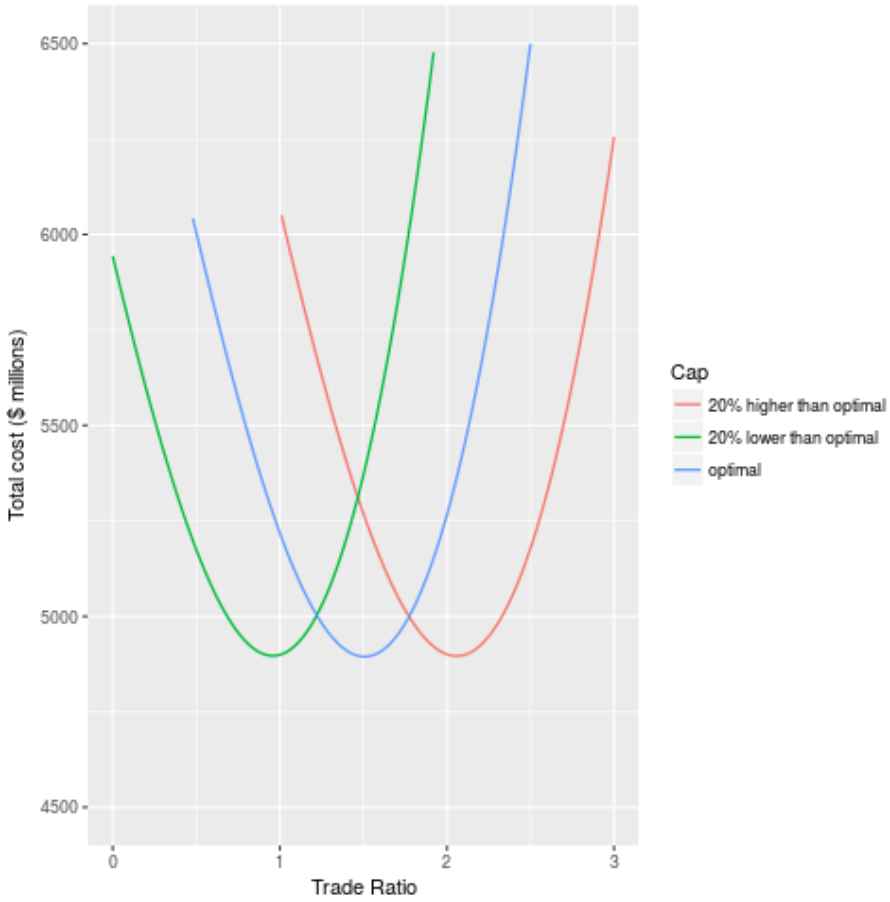

Figure 3.10. Total costs (abatement plus damage costs) for various trade ratios

point source reductions, thereby delivering more immediate ecological benefits).

Figure 3.10 plots the total costs (abatement plus damage) associated with various choices of trade ratios under three different discharge caps. The middle curve represents the total costs of various trade ratios under the optimal discharge cap, while the curves to the right and left illustrate the costs for discharge caps 20% larger and smaller, respectively, than the optimal cap. The minima of these curves represent the optimal trade ratio for the given cap. Note that in the case of the optimal cap and the "optimal plus 20%" cap, the cost-minimizing trade ratios are both greater than one. This would make sense based on the logic that an increase in point source loads must be compensated for by an extra bit of nonpoint reduction to make it worthwhile to wait for the delayed environmental benefits. However, $\psi > 1$ need not be true in general—case in point, the "optimal minus 20%" cap.

For the numerical example in Figure 3.10, under a cap that's 20% smaller than optimal, the second-best trade ratio is less than one. The reason this can persist even in a lagged

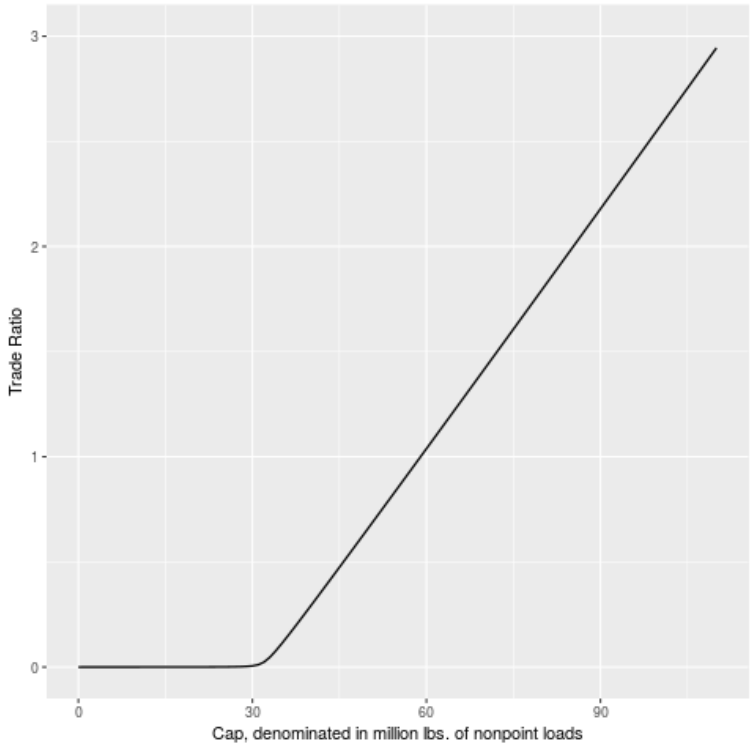

Figure 3.11. Optimal trade ratio given the cap (for 1-year lag length)

pollution context follows from (3.16a) (the condition governing the second best trade ratio) where the solution represents a tradeoff between suboptimal marginal abatement costs and suboptimal marginal damage costs. Under a shrinking cap, pollution sources face rising abatement costs, while pollution damages become less severe. Reducing the trade ratio below one in this scenario will shift loads toward point sources where abatement costs tend to be steepest. This inevitably will increase pollution damages but the overall tradeoff with be worthwhile. Figure 3.11 illustrates this relationship between the size of the cap and the optimal trade ratio for a simple numerical example. This result mirrors those found by Shortle [69] and Horan and Shortle [39] where the presence of risk in nonpoint pollution control does not theoretically preclude trade ratios less than one.

Consistent with the framework put forward in Horan and Shortle [37], this last result implies that trade ratios must be chosen keeping the overall load cap explicitly in mind. Under optimal caps, the presence of lags implies a nonpoint-point trade ratios greater than one, however, under suboptimal caps set especially far below the first-best level, trade ratios between lagged and nonlagged sources may be less than one.

3.4 Conclusion

This essay characterizes the solution to the general problem of managing pollution discharges from multiple sources with different lag lengths, noting that the optimal time-specific load allocations will consist of $N+1$ pollution control regimes where particular sets of source are optimized jointly during various intervals (Figure 3.1). The number of regimes corresponds to the number of distinct lag lengths that exist among the polluters under regulation. This first-best optimum would be achievable in theory if a set of $N+1$ regime-specific load caps were established and permits distributed to the firms belonging to each regime over the correct firm-specific time interval. Even if the lag structure across a watershed were greatly simplified (by perhaps placing sources into bins according to approximate lag length) this market design would require the use of forward contracts which would introduce new dimensions of complexity (time and uncertainty) for the market participants. Given the low participation rates in even simple water quality trading schemes [23], we could expect this design to suppress trading activity still further.

An alternative market design is proposed where participants trade contemporaneous discharges rather than time-dated load deliveries. Properly adjusting for lags using a trade ratio of $e^{\delta\ell}$ (where ℓ represents the difference in lag length between the trading partners and δ is the discount rate) would align the market outcome with the first-best solution during $t > \ell_N$ (i.e., after the period 0 discharges from the source with the longest lag length have been delivered). This interval corresponds to the final regime during which loads settle at steady state levels. While, this trading rule will not generally reproduce the first-best loads for $t < \ell_N$, the approach represents a simple policy design that will correct for at least some of the distortion in the market brought on by lag length discrepancies.

While designing a first-best trade ratio scheme prior to the steady state would entail the same type of regime-specific policy that makes forward markets prohibitively complex, I characterize first-best and second-best trade ratios for a simple two-period, two-polluter model. Adjusting both the discharge cap and the trade ratio, a regulator can, in principle, mimic the first-best solution for any lag length in the nonpoint sector (Figure 3.9). Modifying the cap allows discharges to increase while increasing the trade ratio shifts loads away from point sources and toward nonpoint sources. The optimal cap and the optimal trade ratio both increase with nonpoint lag length, and nonpoint-point trade ratio will exceed one whenever lags exist. In a second best context, the regulator takes a suboptimal cap as given and trades off the abatement cost savings associated with higher point source loads against the

damages prevented by allocated loads from point to nonpoint sources. Even in the presence of lags, optimal nonpoint-point trade ratios may be less than one when the cap is sufficiently small. Such cases result from the relative importance of abatement cost versus damage costs, the former tending to be large and the latter tending to be small under a stringent cap. As previous studies have shown in other contexts (see Horan and Shortle [37]; Horan and Shortle [39]), regulators must account for the size of the cap when designing trade ratios that account for lag length.

Chapter 4

Achieving Joint Emissions Targets for Multiple Pollutants: Sequential vs. Simultaneous Permit Trading

4.1 Introduction

The economic and ecological damages associated with waste disposal depend on both the amounts and types of waste that combine in environmental media (air and water) [14]. Many forms of harmful pollution result from the presence of multiple substances in combination. For example, smog (responsible for eye, throat, and lung irritation) results from the interaction of sunlight with both hydrocarbons and nitrogen oxide in the atmosphere [33]. Indeed, the problem of eutrophication in aquatic ecosystems discussed in chapter 1 results from the over-abundance of nitrogen (N) and phosphorus (P) in these waters [71, 42]. Depending on how two or more pollutants interact in the environment, the optimal policy may prescribe reductions of each pollutant in similar proportions or call for reductions exclusively of one pollutant or another. In the case of nutrient pollution, physical scientists have pointed out that reductions of *both* N and P are critical for maintaining healthy estuaries and coastal waters [12, 64]. The question then becomes how best to meet these dual objectives [79]. Theoretical arguments for the use of tradable pollution permit systems [14, 58, 74] and recent implementations of these systems [30] suggest potential for market mechanisms to allocate pollution reductions at multiple sources cost-effectively. This essay considers the design of markets for the joint control of multiple pollutants with an application to N and P reduction

in the wastewater treatment sector.

Determining the socially efficient level of pollution is an immensely difficult calculation [14, 4] and the need to consider the joint levels of N and P [12, 64] compounds this difficulty. Endres [19] and Beavis and Walker [5] were among the first to characterize the economic problem of choosing the correct levels of multiple pollutants with potentially nonlinear interactions with respect to environmental benefits and abatement costs. Following these studies, Ungern-Sternberg [79] and Kuosmanen and Laukkanen [50] show that the relative curvature of the benefit and abatement cost functions is decisive for whether abatement effort should focus on one pollutant or both. In dynamic settings, Michaelis [56], and Moslener and Requate [59, 60] examine the optimal time paths for the abatement of multiple stock pollutants given their jointness in both benefits and costs. These studies concern the choice of optimal pollution levels, but an alternative approach would be to target some "acceptable" pollution level and seek to achieve it at minimum cost [14, 4]. Gren et al. [31], and Elofsson [18] represent empirical analyses in this vein, computing least-cost nutrient load allocations among sources discharging to the Baltic Sea.

While the research above attempts to identify pollution targets for multiple pollutants based on either efficiency or cost-effectiveness criteria, the subsequent policy question concerns how best to implement these targets. Ambec and Coria [1] define the conditions under which taxes, tradable permits, or a mixed policy most efficiently regulate two pollutants under uncertainty, while Montero [57] derives rules for when markets for reductions of two pollutants should be integrated or separated in the presence of both uncertainty and incomplete enforcement. Another important design element when targeting multiple objectives is the choice of sequential versus simultaneous implementation—does implementing two policies at the same time lead to the same outcome as implementing them one after the other? Feng et al. [22] study this question in the context of land conservation policy where a land conservation program may have two fixed budgets: one to fund land retirement and the other to fund alterations to management practices on working land. Since a given land parcel can only be designated for *either* retirement *or* management alterations, the authors find that allocating these two budgets optimally in sequential fashion may lead to inefficient conservation choices relative to the case where the fixed budgets are allocated simultaneously.

The timing of implementation may also matter when the abatement of two or more pollutants exhibits economies of scope and investment in pollution control is irreversible [52][1].

[1]Referring to the wastewater treatment sector, Lence, Eheart, and Brill [52] point out that "if efficient capital investments are made in facilities to control one pollutant, those facilities may be inefficient when

Ermoliev, Michalevich, and Nentjes [20] show theoretically that a tradable permit market with bilateral, sequential trades will eventually converge on the least-cost emissions allocation, but they concede that these results depend critically on participating firms being able to costlessly adjust emissions levels up and down. This assumption of costless adjustment is inappropriate for nutrient reductions in the wastewater treatment sector where reduction measures predominantly involve discrete, irreversible capital investments [73].

Lence, Eheart, and Brill [52] study this issue empirically for a set of municipal wastewater dischargers along the Willamette River in Oregon, analyzing the relative performance of sequential versus simultaneous tradable discharge permit systems for three pollutants: N, P, and biological oxygen demand (BOD). They test multiple management scenarios (uniform discharge standards, sequential markets with trades occurring first in BOD then in P, sequential markets with trades occurring first in P then in BOD, etc.) and solve numerically for the cost-minimizing (market equilibrium) investments at the various facilities under each regulatory design. Due to interdependecies in treatment costs, they find that investing in reductions of each pollutant one by one is more costly than if investments were chosen with the *joint* standard in mind from the start. This finding is corroborated by engineering principles—in particular, Sedlak [68] notes that "biological processes for removal of [nitrogen and phosphorus] may be incorporated into the standard activated sludge secondary treatment process with relative ease. (p.170)"

While Lence et al. [52] show that a simultaneous market design outperforms a sequential design for the particular case of N, P, and BOD management along the Willamette River, a more general characterization of how implementation timing matters for multiple pollutant control has yet to be developed. These authors stress the importance of cost interdependence for determining the size of this performance gap[2], though their methodological approach (optimizing over a finite set of discrete abatement choices) precludes making a more precise statement about this relationship. This essay extends Lence et al.'s [52] analysis by modeling facility-level nutrient reduction costs as a continuous function of the facility's existing treatment levels (defined as the current N and P concentration of the facility's discharges) and considering the optimal pattern of investment required to meet any pair of N and P discharge standards. The model allows for an evaluation of the efficiency shortfall associated with a sequential policy (relative to a simultaneous one) for any joint discharge standard,

controls on other pollutants are subsequently imposed. (p.897)"

[2]Specifically, they state that"[w]hether any economic losses from poor timing of the individual permit markets are significant depends on the degree of interdependence of treatment costs for the various pollutants. (p.898)"

given the structure of joint abatement costs and the initial treatment capacities of facilities in the regulated sector. Analysis shows that a sequential market design falls short of a simultaneous one only for a subset of possible joint discharge targets, even in the presence of economies of scope. This framework is useful not only for evaluating and designing markets for nutrient reductions where municipal wastewater dischargers feature prominently [65, 66], but for many other cases in environmental policy where multiple environmental goods are produced jointly [82, 53].

I begin with a description of the joint control of N and P in wastewater treatment sector in section 4.2 before laying out a conceptual framework in section 4.3 for the cost-effective control of two pollutants discharged from a stylized wastewater treatment sector. Section 4.4 describes the empirical model of nutrient removal costs and estimates a joint N and P removal cost function based on cost data from N and P dischargers in the Chesapeake Bay watershed. Section 4.5 applies the modeling framework to analyze the efficiency gap between a sequential and simultaneous market design. The final section concludes.

4.2 The Case of Wastewater Treatment

Wastewater treatment plants (WWTPs) transform polluted influent through a combination of physical, chemical and biological processes, making it more suitable for discharge into the natural environment or a drinking water system. They are particularly prominent sources of nitrogen and phosphorus in small drainage basins containing dense population centers (e.g., the Passaic River Basin, the Long Island Sound Watershed, and the Narrangasett Bay Watershed in the northeastern United States), and even in larger watersheds, wastewater discharges may constitute as much as half of river flow under low-flow conditions [2]. For the purposes of nutrient pollution management, the important characteristics of a WWTP are the *concentrations* of N and P in its discharged effluent and its *rate* of discharge. The product of these two quantities gives the total mass of pollution discharged. Since WWTPs are public utilities (and therefore must accept whatever volume of wastewater that businesses and residents deliver) their discharge rates are not, for practical purposes, under their control. Rather, WWTPs only control the concentrations of the pollutants in these exogenous volumes. These concentrations can be reduced by augmenting standard wastewater treatment processes.

While the chemical and biological processes that remove N and P from wastewater are distinct, increasing the capacity to remove N may involve only modest adjustments to existing

processes for removing P, and vice versa. For example, the A/O process[3] for phosphorus removal requires wastewater to pass through an anaerobic zone to release P before moving through an aerobic zone where the P binds to the sludge to be later physically filtered out. This process can be readily converted to an A^2/O process[4] by adding an anoxic zone between the anaerobic and aerobic zones already part of the A/O process [68]. Similarly, the Bardenpho process for nitrogen removal involves wastewater passing through an anoxic zone where carbon in the wastewater facilitates denitrification, which is then released as nitrogen gas in the subsequent aerobic zone. Modifying this system for phosphorus removal involves adding an anaerobic zone at the front of this existing process [68]. In this way, the shared components (here, the anaerobic and anoxic zones) allow for less expensive upgrades for N (P) removal where P (N) removal capacity already exists. These facts of the engineering imply that the process of N and P treatment exhibit economies of scope where the incremental cost of upgrading a plant's N (P) treatment capacity depends on its current capacity to remove P (N).

Along the lines just described, a facility with a high existing N(P)-removal capacity may have a relatively low incremental cost of removing P (N) because the structures used in the removal of one nutrient may play a role in the removal of the other. The incremental cost of reducing N (P) at a particular plant will therefore be *inversely* related to how intensely the plant is already reducing P (N). At the same time, the principle of diminishing returns will cause the marginal cost of reducing N (P) at a particular plant to be *directly* related to how intensely the plant is already reducing N (P). Standard treatment processes will remove large amounts of N and P without any modification, but achieving concentrations below this baseline requires specialized basins and aeration tanks which become more sophisticated as the desired concentration falls [68]. Economies of scope and diminishing returns work in tandem to define a relationship between incremental nutrient removal costs and existing nutrient removal capacity. These effects matter for the cost-effective investments in N and P removal insofar as the existing treatment capacities of the WWTPs under consideration are heterogeneous.

As seen in Figure 4.1, hetergeneity in existing treatment capacity is prevalent among the municipal WWTPs that discharge to the Chesapeake Bay. Figure 4.1 plots the nutrient treatment levels at 416 municipal point sources in the Chesapeake Bay watershed, where a facility's position on the x-y plane corresponds to the concentrations of N and P in its

[3] A/O stands for "anaerobic/oxic"
[4] A^2/O stands for "anaerobic/anoxic/oxic"

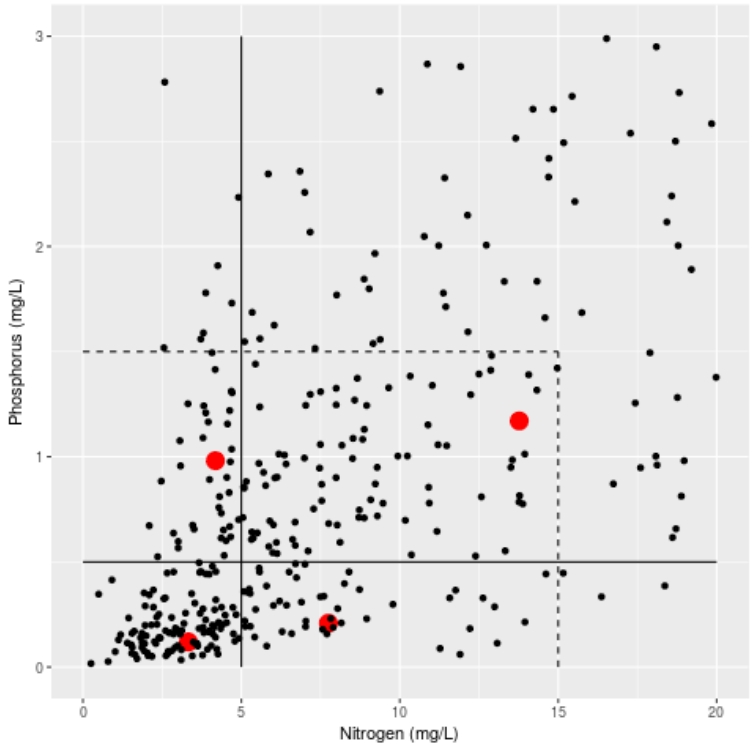

Figure 4.1. Significant Municipal Point Source Dischargers in Chesapeake Bay's Watershed

treated effluent. Facilities in the northwest quadrant have invested heavily in N removal but relatively little in P removal, and the reverse is true for facilities in the southeast quadrant. Facilities in the southwest quadrant have invested heavily in the removal of both N and P, while facilities in the northeast have not invested heavily in the removal of either. By the logic outlined above, the cheapest units of P removal will come from facilities in the northwest of Figure 4.1, while the cheapest units of N removal will come from southeast facilities. In this way, a facility's relative position in N-P concentration space matters influences how cost-effectively it can perform nutrient removal, and the cost-minimizing set of upgrades for any given joint discharge target will depend on the degree of complementarity in the joint abatement process and the distribution of existing treatment capacity in the set of polluting facilities.

Figures 4.2 and 4.3 further illustrate this principle, plotting the level curves (in concentration space) for the marginal cost of N and P reduction, respectively. The arrows in each figure indicate directions of increasing cost and each line represents combinations of N and P concentrations for which marginal reductions in N or P is equally costly. Any movement due west (smaller N concentration holding P constant) will decrease the marginal cost of

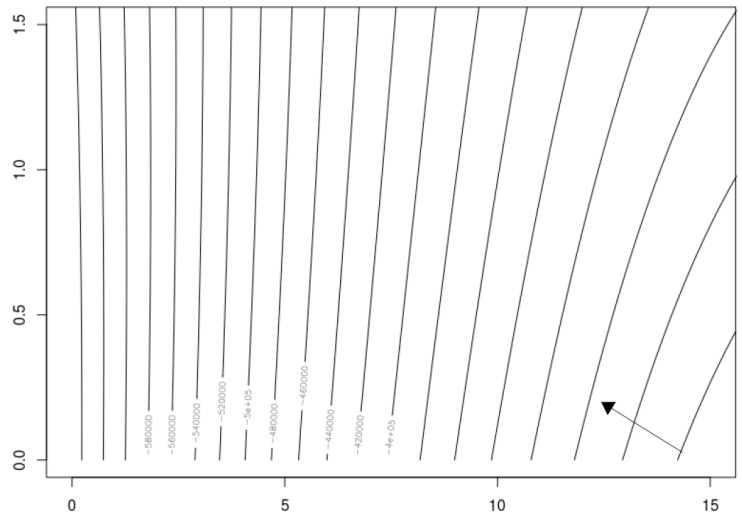

Figure 4.2. Iso-marginal cost lines for N removal (x-axis N mg/L; y-axis P mg/L)

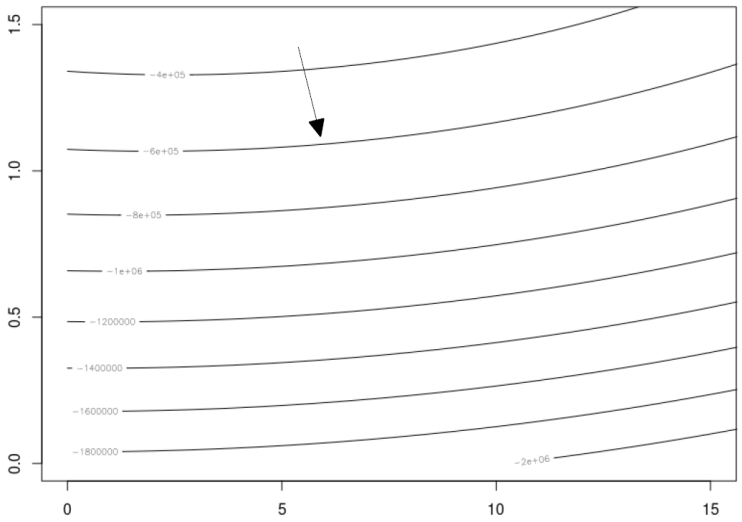

Figure 4.3. Iso-marginal cost lines for P removal (x-axis N mg/L; y-axis P mg/L)

removing P (Figure 4.3) and increase the marginal cost of reducing N (Figure 4.2). Similarly, any movement due south (smaller P concentration holding N constant) will increase the marginal cost of removing P (Figure 4.3) and decrease the marginal cost of reducing N (Figure 4.2). These principles form the basis of the continuous optimization framework that will be used to for evaluating alternative market designs.

4.3 Optimal Two-Pollutant Control with Economies of Scope

This section considers the optimal targeting of upgrades among a set of sources with heterogeneous existing treatment capacities and describes how two-pollutant discharge targets map onto the optimal set of facility-level upgrades. For conceptual clarity, the wastewater treatment sector will consist of three WWTPs that differ only with respect to the levels of nutrient reduction they've already achieved.

Let variables n_i and p_i denote the post-treatment concentrations (treatment capacities) of N and P in wastewater treatment facility i's discharges, where $i \in \{1, 2, 3\}$. Suppose that facilities have just two possible initial treatment capacities for each pollutant—facilities will have already reduced their N concentration down to either n^H or n^L (with $n^L < n^H$) and their P concentrations down to either p^H or p^L (with $p^L < p^H$). Let F_i denote facility i and let n_i^o and p_i^o refer to the initial N and P treatment capacities at F_i, where $\langle n_1^o, n_2^o, n_3^o \rangle = \langle n^L, n^H, n^H \rangle$ and $\langle p_1^o, p_2^o, p_3^o \rangle = \langle p^H, p^H, p^L \rangle$. Facilities accept and discharge wastewater at rates ω_i, and the means by which nutrient reduction occurs is solely through the lowering of the nutrient concentrations in these exogenous volumes (facilities do not have the option to "reduce output"). Finally, all facilities have an identical cost function $c(n, p)$ where costs increase in the reduction of either nutrient at increasing rates ($c_n < 0$, $c_{nn} < 0$, $c_p < 0$, $c_{pp} < 0$), and marginal costs of reducing one pollutant decreases the more the facility reduces the other pollutant ($c_{np} > 0$). Since cost structures at the facilities are the same, cost disparities stem only from differences in the initial treatment capacities.

Conceptually, the three modeling units introduced above represent facilities in the northwest, northeast, and southeast quadrants of Figure 4.1. These facilities are represented by the red circles in Figure 4.4, Panel B. Given the exogenous flow volumes and the initial treatment capacities of facilities in this stylized wastewater sector, baseline aggregate N and P discharges are given by

$$n^L\omega_1 \; + \; n^H\omega_2 \; + \; n^H\omega_3 \; \equiv \; N_0$$

and

$$p^H\omega_1 \; + \; p^H\omega_2 \; + \; p^L\omega_3 \; \equiv \; P_0$$

These baselines are represented by the red square in Figure 4.4, Panel A. Let the aggregate

cost of any set of nutrient removal capacities be given by

$$\pi(\boldsymbol{n}, \boldsymbol{p}) \;=\; \sum_{i=1}^{3} w_i \, c(n_i, p_i)$$

where $\boldsymbol{n}$ and $\boldsymbol{p}$ are vectors of treatment levels $\langle n_1, n_2, n_3 \rangle$ and $\langle p_1, p_2, p_3 \rangle$. Given any nitrogen discharge target $\bar{N}$ and any phosphorus discharge target $\bar{P}$, the problem is to

$$\min_{\boldsymbol{n},\boldsymbol{p}} \; \pi(\boldsymbol{n}, \boldsymbol{p}) \quad \text{subject to} \quad \sum_{i=1}^{3} n_i \omega_i \leq \bar{N}, \quad \sum_{i=1}^{3} p_i \omega_i \leq \bar{P}, \quad n_i \leq n_i^o, \quad p_i \leq p_i^o \qquad (4.1)$$

The first two inequality constraints represent the requirement to meet the aggregate discharge limits $\bar{N}$ and $\bar{P}$. The constraints on individual variables n_i and p_i represent irreversibility in treatment capacity, that is, sources may reduce their N or P concentration but not liquidate existing their capital stock.

Reducing the mass of nutrient discharges below the initial aggregate discharge levels (N_0, P_0) requires upgrading the treatment capacities of one or more of the polluting facilities. Doing so at minimum cost means upgrading the low cost facilities first and bringing along the costlier ones only if the emissions target requires it. Since cost disparities stem from differences in treatment capacities, the disparities that exist initially will diminish as upgrades at lower capacity sources occur and begin to catch up those that have higher treatment capacity.

Figures 4.4A and 4.4B depict the problem's initial conditions in terms of total discharges and facility-level concentrations. To achieve some emissions target $\bar{P} < P_0$, which facilities should be upgraded? As previously explained, P reductions are cheapest in the northwest corner of 4.4B and become more expensive with any movement south or east. For small reductions in P, upgrades need only occur at F_1, following the arrow in 4.4D. Along this path, diminishing returns set in, and eventually (at the dotted yellow line) F_1's marginal P removal cost will match F_2's. Let P_1 (Figure 4.4C) denote aggregate P discharges at this juncture. Once F_1 "catches up" to F_2 in terms of marginal P reduction costs, the burden of any further P reductions will be shared between them, and they will reduce their concentrations together so as to always maintain equal marginal costs. F_3 will not be required to upgrade until F_1 and F_2 both "catch up" to it (at the dotted yellow line in 4.4F). Let P_2 (Figure 4.4E) denote total P discharges at this juncture. For an emissions target $\bar{P} < P_2$, concentrations will be reduced at all three facilities (again, maintaining equal

marginal costs). The optimal investment in N reductions follows exactly the same logic, and Figure 4.5 depicts these analogous cases.

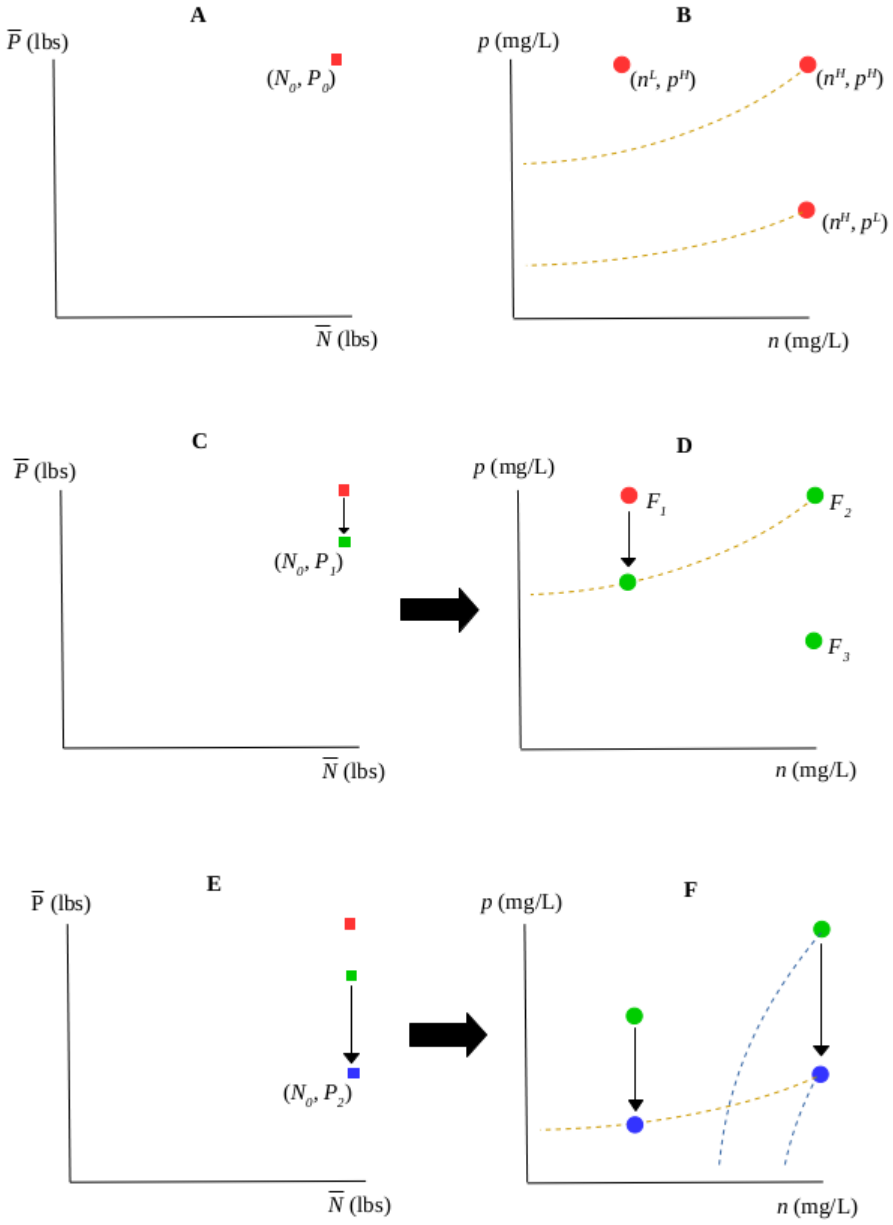

Figure 4.4. Mapping P reductions onto facility upgrades

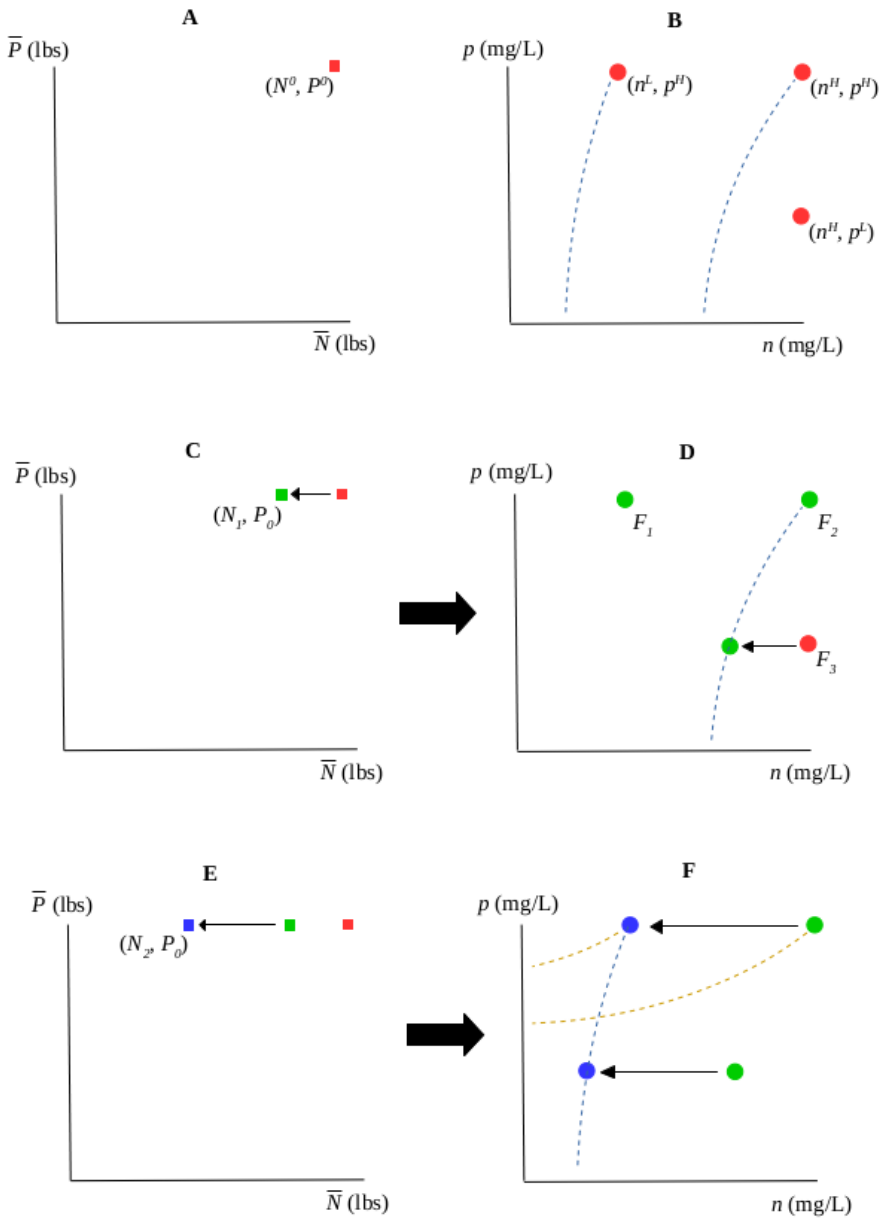

Figure 4.5. Mapping N reductions onto facility upgrades

The upgrade regimes in Figures 4.4 and 4.5 pertain to aggregate discharge standards that only target one pollutant or the other. More interesting situations arise when N and P are targeted in combination. Specifically, when economies of scope are present, the optimal upgrade regime for a given $\bar{N}$ may depend on $\bar{P}$ and vice versa. To see how $\bar{N}$ and $\bar{P}$ jointly influence the upgrade regime, consider making a small P reduction starting at the position in Figures 4.4A and 4.4B (designate this scenario "case 1") versus the position in Figure

4.5E and 4.5F (designate this scenario "case 2"). No upgrades have yet occurred in case 1, whereas N discharges have been reduced to N_2 in case 2. In case 1, F_1's marginal P reduction cost is strictly lower than F_2's, so the optimal plan for achieving a small P reduction with $\bar{N} = N_0$ would be to upgrade F_1 only. In case 2, F_2 has "caught up" to F_1 in terms of N reduction, making their marginal P reduction costs equal. The optimal plan for achieving a small P reduction with $\bar{N} = N_2$ would be to upgrade both facilities simultaneously. F_2 therefore begins upgrading P "sooner" (i.e., at a higher $\bar{P}$) when $\bar{N} = N_2$ than when $\bar{N} = N_0$.

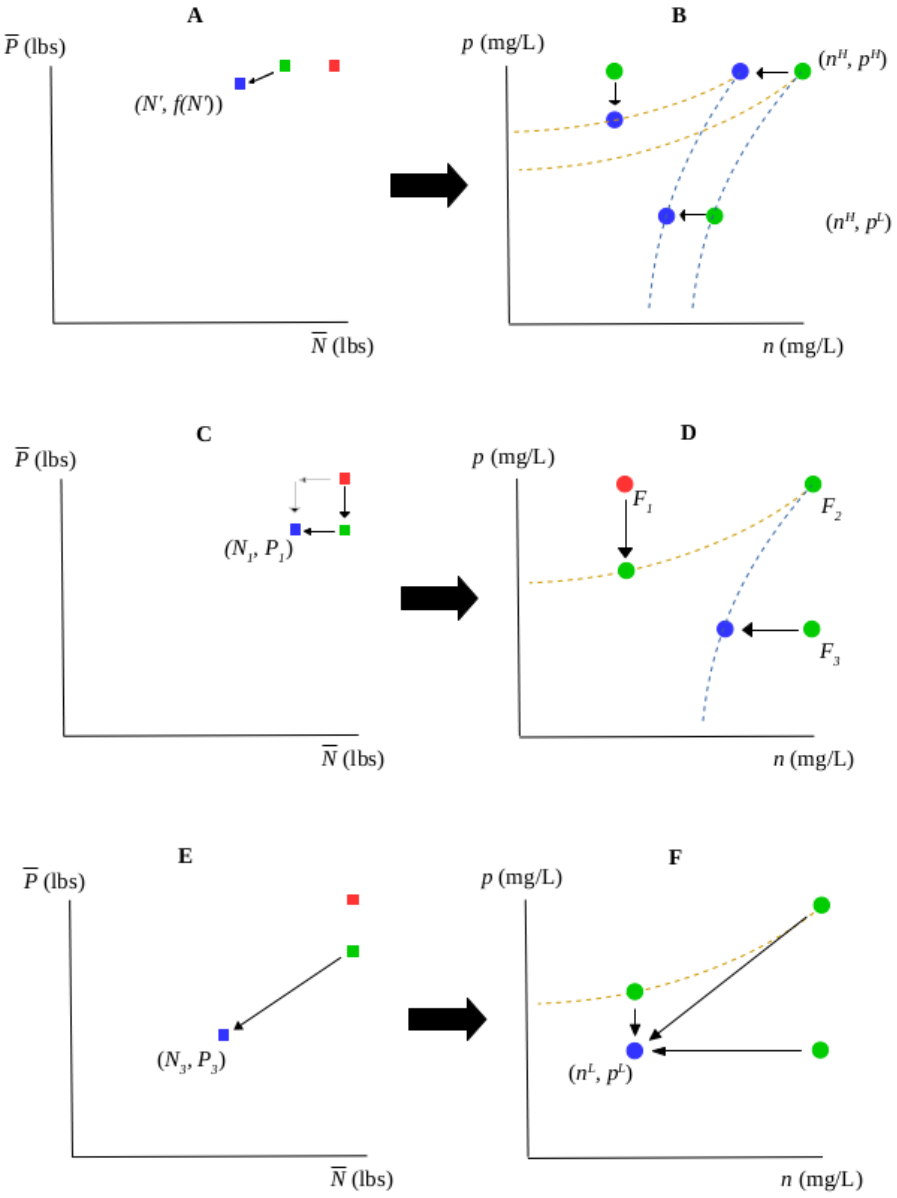

Figure 4.6. Mapping combinations of N and P reductions onto facility upgrades

While $\bar{N} = N_2$ represents one instance where this occurs, in fact there is an entire interval $\bar{N} \in (N_2, N_1)$ for which F_2 will begin contributing P reductions at some unique $\bar{P}$. One of these intermediate cases is shown in Figures 4.6A and 4.6B, where $\bar{N} = N'$. While F_2 doesn't technically upgrade P in Figure 4.6B as drawn, it becomes "active" in the sense that the very next bit of P reduction will require F_2 to join F_1 in the reductions. Note that if N concentrations were lower, F_2 would become active even earlier in the P removal

process. Also note that for any combination $\bar{N} \geq N_1$ and $\bar{P} \geq P_1$, F_2 is never called upon to reduce at all (see Figure 4.6D) and thus N and P reductions occur independently at F_3 and F_1, respectively. Outside of this region (i.e. anywhere outside the rectangle outlined by the arrows in Figure 4.6C), the optimal upgrade regimes will be determined by the joint selection of $\bar{N}$ and $\bar{P}$.

Let the function $f(\bar{N})$ define a set of regime-switching points, where f takes any $\bar{N}$ and maps it to the maximum $\bar{P}$ for which F_2 will become "active" with respect to both N and P upgrades. f will be a piecewise function whose domain is partitioned along $\bar{N} = N_1$. For $\bar{N} \leq N_1$, setting $\bar{P} < f(\bar{N})$ will induce F_2 to upgrade with respect to P, otherwise F_2 will remain at p^H. For $\bar{N} > N_1$, setting $\bar{P} < f(\bar{N})$ will induce F_2 to upgrade with respect to N, otherwise F_2 will remain at n^H. The right-hand panel of Figure 4.7 plots the general form of $f(\bar{N})$.

Another set of regime-switching points occurs where emissions targets become so strict as to induce N or P upgrades at the least efficient facilities (F_1 for N treatment and F_3 for P treatment). Again, because of economies of scope, these frontiers cannot be defined at fixed values of $\bar{N}$ or $\bar{P}$ but instead depend on the two values in combination. Let the function $g(\bar{N})$ define the boundary in discharge space that separates these regimes. g is a piecewise function whose domain can be partitioned along $\bar{N} = N_3$, where N_3 is the discharge level in Figure 4.6E at which all three facilities converge on (n^L, p^L) (Figure 4.6F). For any $\bar{N} \leq N_3$, setting $\bar{P} < g(\bar{N})$ will induce F_1 (initially the costliest source of N reduction) to upgrade with respect to N, otherwise F_1 will remain at n^L. Similarly, for any $\bar{N} > N_3$. setting $\bar{P} < g(\bar{N})$ will induce F_3 (initially the costliest source of P reduction) to upgrade with respect to P, otherwise F_3 will remain at p^L. The right-hand panel of Figure 4.7 plots the general form of $g(\bar{N})$.

The functions f and g, along with the points (N_1, P_1), (N_2, P_2) and (N_3,P_3) allow a complete characterization of the optimal upgrade regimes for any emissions target ($\bar{N}$, $\bar{P}$). Table 4.1 enumerates the 14 qualitatively unique upgrade regimes and identifies the sets of $\bar{N}$ and $\bar{P}$ that make them optimal. A "yes" under column n_i or p_i, in Table 4.1 indicates that the concentration of n or p at facility i is reduced below its baseline level for the joint discharge target specified in columns $\bar{N}$ and $\bar{P}$. The left-hand panel of Figure 4.7 depicts the complete set of boundaries between the upgrade regimes specified in Table 4.1, where the n's and p's in each zone indicate which nutrient levels must drop below their baseline to achieve the joint discharge target implied by that zone. The numbers in the upper left of each zone correspond to the regime numbers in Table 4.1.

#	Figure	$\bar{N}$	$\bar{P}$	n_1	p_1	n_2	p_2	n_3	p_3
1	4.4A, 4.4B, 4.5A, 4.5B	$\bar{N} = N_0$	$\bar{P} = P_0$	–	–	–	–	–	–
2	4.4C, 4.4D	$\bar{N} = N_0$	$P_1 < \bar{P} < P_0$	–	yes	–	–	–	–
3	4.4E, 4.4F	$\bar{N} = N_0$	$P_2 < \bar{P} < P_1$	–	yes	–	yes	–	–
4		$\bar{N} = N_0$	$0 < \bar{P} < P_2$	–	yes	–	yes	–	yes
5	4.5C, 4.5D	$N_1 < \bar{N} < N_0$	$\bar{P} = P_0$	–	–	–	–	yes	–
6	4.5E, 4.5F	$N_2 < \bar{N} < N_1$	$\bar{P} = P_0$	–	–	yes	–	yes	–
7		$0 < \bar{N} < N_2$	$\bar{P} = P_0$	yes	–	yes	–	yes	–
8	4.6C, 4.6D	$N_1 < \bar{N} < N_0$	$P_1 < \bar{P} < P_0$	–	yes	–	–	yes	–
9		$N_1 < \bar{N} < N_0$	$f(\bar{N}) < \bar{P} < P_1$	–	yes	–	yes	yes	–
10	4.6A, 4.6B	$N_2 < \bar{N} < N_1$	$f(\bar{N}) < \bar{P} < P_0$	–	yes	yes	–	yes	–
11	4.6E, 4.6F	$N_2 < \bar{N} < N_0$	$g(\bar{N}) < \bar{P} < f(\bar{N})$	–	yes	yes	yes	yes	–
12		$N_3 < \bar{N} < N_0$	$0 < \bar{P} < g(\bar{N})$	–	yes	yes	yes	yes	yes
13		$0 < \bar{N} < N_3$	$P_3 < \bar{P} < g(\bar{N})$	yes	yes	yes	yes	yes	–
14		$0 < \bar{N} < N_3$	$0 < \bar{P} < P_3$	yes	yes	yes	yes	yes	yes

a "yes" beneath n_i or p_i indicates the concentration is reduced below its baseline for the discharge requirement corresponding to that row

Table 4.1. Upgrade Regimes Across all ($\bar{N}$, $\bar{P}$) Combinations

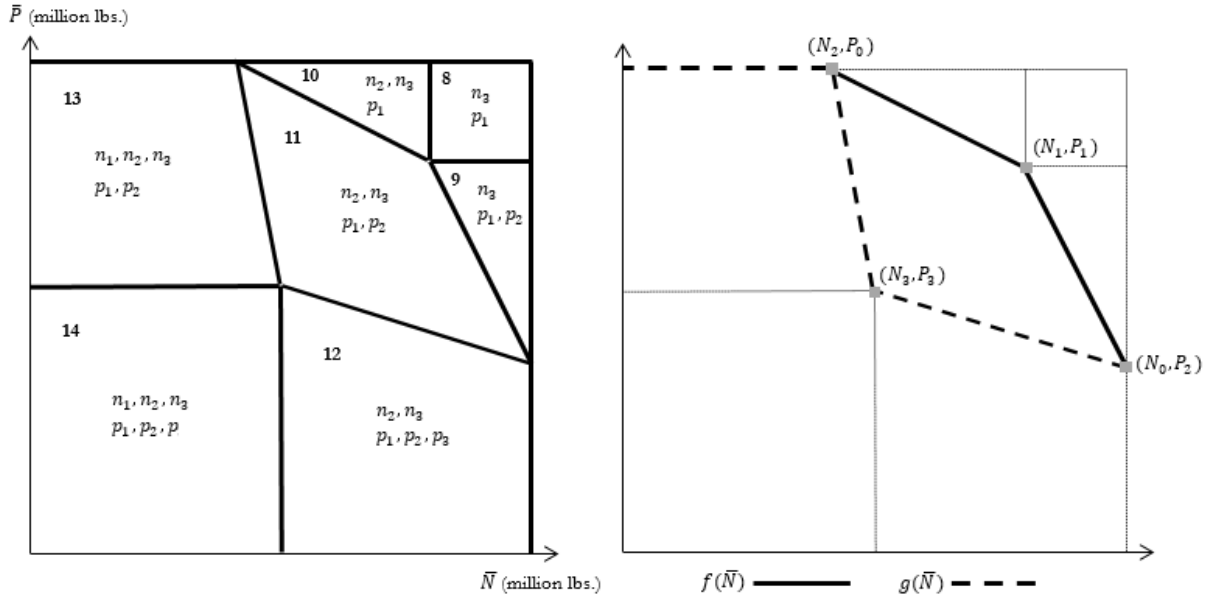

Figure 4.7. Zones of Qualitatively Similar Upgrade Regimes

4.4 Estimating the Nutrient Removal Cost Function

A recent Chesapeake Bay Program (CBP) publication [24] compiles a list of 304 wastewater treatment facilities in the Chesapeake Bay watershed and estimates their nutrient removal upgrade costs through a combination of direct contact with facility managers and engineering estimates based on the characteristics of the facilities. The report defines four tiers of N and P removal (Table 4.2) and lists the incremental capital costs associated with upgrading the facility from one tier to the next. While there are 304 facilities in the data set, the report includes the upgrade costs associated with moving from each tier to the next, so there can be multiple cost observations for a given facility. 31% of the facilities were already equipped to treat P above the Tier 1 level (< 1mg/L) and 11% were equipped to treat N above Tier 1 (< 8mg/L), in which cases upgrade costs were only reported for the tiers that represented an improvement in nutrient removal capacity. No facilities were already at Tier 4 capacity for both N and P.

	Nitrogen (mg/L)	Phosphorus (mg/L)
Tier 1	$N \geq 8$	$P > 1.0$
Tier 2	$5 < N < 8$	$0.5 < P < 1.0$
Tier 3	$3 < N < 5$	$0.1 < P < 0.5$
Tier 4	$N < 3$	$P < 0.1$

Table 4.2. Nutrient Treatment Tier Definitions from CBPO (2002)

In addition to these upgrade cost estimates, the report includes information on the initial N and P treatment levels at each plant, making it possible to specify upgrade costs as a function of existing treatment capacity and estimate the degree of interdependence in the capital costs of N and P removal. Because CBP [24] reports nutrient removal costs separately for N and P (holding the treatment level of the other nutrient constant), these values represent the "partial" incremental costs of upgrading treatment capacity. Since the analysis in section 4.3 is based on a *joint* N and P removal cost function, I use these two sets of "partial" incremental cost data to estimate a two-equation system from which to derive the parameters of the joint cost function. In what follows I specify the empirical model of nutrient removal costs and present results of the estimation.

4.4.1 Empirical Model

Let x denote a set of physical or chemical wastewater characteristics (nutrient content, prevalence of organic compounds or suspended solids, etc.). WWTPs receive water having a set of baseline characteristics, $\bar{x}$, and after applying various treatment processes, discharge water with a new set of characteristics x^o, thereby rendering it more desirable. While building the capacity to achieve a particular wastewater quality target requires large initial outlays, once these structures are in place, all incoming wastewater can be treated to those levels. To model this process, let $f(x;\bar{x})$ represent the capital costs associated with building the capacity to convert incoming wastewater from an untreated form $\bar{x}$ to a higher quality form x.

Taking x^o as the plant's initial treatment capacity, suppose a higher treatment level x' is desired. Let the capital costs associated with an upgrade from x^o to x' be the difference between the total capital costs of the new treatment capacity x' and those of the old capacity x^o. Using $g(x', x^o)$ to represent the *incremental* upgrade costs of adjusting treatment capacity from x^o to x' this specification implies

$$g(x', x^o) \;=\; f(x';\bar{x}) \;-\; f(x^o;\bar{x}) \tag{4.2}$$

Defining $g(\cdot)$ in this way pins costs down to the ultimate baseline $\bar{x}$, the "no-treatment" level where treatment costs are zero and a natural reference point against which all upgrades can be compared. Representing wastewater treatment capital in this way inevitably leaves out practical concerns such as adjustment costs, compatibility between the newly introduced structures and the old ones, and economies of scale in construction. For this analysis, I lay aside these concerns in order to focus on economies of scope in the treatment of multiple pollutants.

Next, consider a particular $f(\cdot)$ where the important wastewater characteristics are N and P concentrations. Let $f(n, p; \bar{n}, \bar{p})$ denote the capital cost associated with reducing N and P concentrations from their untreated baselines $(\bar{n}, \bar{p})$ down to a new set of concentrations (n, p). Let f take on a cubic structure with a vector of parameters $\boldsymbol{\theta} = [a, b, c, d, e, r, u, v, w]$

parameter	expected sign	interpretation (given expected sign)
a, b	> 0	N and P removal costs increase as concentration falls
c, e	> 0	marginal costs increase as concentration falls
r, w	> 0	marginal costs increase at an increasing rate as concentration falls
d	< 0	marginal costs of N (P) removal fall as concentration of P (N) falls
u, v	> 0	marginal costs of N (P) removal fall at a decreasing rate as concentration of P (N) falls

Table 4.3. Expected signs of the parameters of the nutrient removal cost function

such that

$$\begin{aligned} f(n,p;\bar{n},\bar{p},\boldsymbol{\theta}) \;=\; & a(\bar{n}-n) \;+\; b(\bar{p}-p) \\ & +\; c(\bar{n}-n)^2 \;+\; d(\bar{n}-n)(\bar{p}-p) \;+\; e(\bar{p}-p)^2 \\ & +\; r(\bar{n}-n)^3 \;+\; u(\bar{n}-n)^2(\bar{p}-p) \;+\; v(\bar{n}-n)(\bar{p}-p)^2 \;+\; w(\bar{p}-p)^3 \end{aligned} \tag{4.3}$$

The expected signs of the parameters of this joint cost function are listed in Table 4.3. There are three sets of parameters that relate to the independent effects on capital costs of reducing the concentration of either N *or* P in wastewater. It would be expected that total capital costs of nutrient removal will increase as nutrient concentration falls for all $n \in [0, \bar{n}]$ and $p \in [0, \bar{p}]$. Positive coefficients on the linear terms $(\bar{n} - n)$ and $(\bar{p} - p)$, that is $a > 0$ and $b > 0$, would ensure this. From diminishing returns, we'd also expect nutrient removal capital costs to increase as nutrient concentration falls for all $n \in [0, \bar{n}]$ and $p \in [0, \bar{p}]$. Positive coefficients on the quadratic terms $(\bar{n} - n)^2$ and $(\bar{p} - p)^2$, that is $c > 0$ and $e > 0$, would ensure this. Furthermore, since we'd expect the cost of nutrient removal to become prohibitively expensive at very low concentrations, marginal cost are likely to increase at an increasing rate as concentration falls. Positive coefficients on the cubic terms $(\bar{n} - n)^3$ and $(\bar{p} - p)^3$, that is $r > 0$ and $w > 0$ would make this happen will ensure that marginal capital costs increase at an increasing rate as nutrient concentration falls. There are also three parameters (d, u, and v) that relate to the effects of reducing the concentration of one pollutant on capital costs reducing the concentration of the other.

The costs of treating N and P will be interdependent, provided that d, u, and v are

non-zero. According to the structure of (4.2), let the "partial" incremental N removal costs be given by

$$g(n, n^o, p^o) \;=\; f(n, p^o) - f(n^o, p^o)$$

The first argument of g is the post-upgrade N concentrations while the second and third arguments are the pre-upgrade N and P concentrations. The incremental cost is a function of the magnitude of the treatment upgrade (the difference between n and n^o) and the existing treatment capacities of both pollutants (n^o and p^o). Expressed in terms of f and its parameters

$$\begin{aligned} g(n, n^o, p^o; \boldsymbol{\theta}) \;&=\; f(n, p^o; \boldsymbol{\theta}) - f(n^o, p^o; \boldsymbol{\theta}) \\ &=\; \Big[a + 2c\bar{n} + d\bar{p} + 3r(\bar{n})^2 + 2u\bar{n}\bar{p} + v(\bar{p})^2\Big]\,(n^o - n) \\ &+\; \Big[c + 3r\bar{n} + u\bar{p}\Big]\,\Big[(n^o)^2 - n^2\Big] \;+\; \Big[d + 2u\bar{n} + 2v\bar{p}\Big]\,(n^o - n)p^o \\ &+\; r\Big[(n^o)^3 - n^3\Big] \;+\; u\Big[(n^o)^2 - n^2\Big]\,p^o \;+\; v(n^o - n)(p^o)^2 \end{aligned} \tag{4.4}$$

Along the same lines, let the incremental P removal costs be given by

$$h(p, n^o, p^o) \;=\; f(n^o, p) - f(n^o, p^o)$$

where the first argument of h is the post-upgrade P concentration and the remaining two arguments are again the pre-upgrade concentrations. In terms of f and its parameters

$$\begin{aligned} h(p, n^o, p^o; \boldsymbol{\theta}) \;&=\; f(n^o, p; \boldsymbol{\theta}) - f(n^o, p^o; \boldsymbol{\theta}) \\ &=\; \Big[b + d\bar{n} + 2e\bar{p} + u(\bar{n})^2 + 2v\bar{n}\bar{p} + 3w(\bar{p})^2\Big]\,(p^o - p) \\ &+\; \Big[e + v\bar{n} + 3w\bar{p}\Big]\,\Big[(p^o)^2 - p^2\Big] \;+\; \Big[d + 2u\bar{n} + 2v\bar{p}\Big]\,(p^o - p)n^o \\ &+\; w\Big[(p^o)^3 - p^3\Big] \;+\; v\Big[(p^o)^2 - p^2\Big]\,n^o \;+\; u(p^o - p)(n^o)^2 \end{aligned} \tag{4.5}$$

To uncover $\boldsymbol{\theta}$, I estimate g and h as a linear two-equation system with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$

$$\begin{aligned} g(n, n^o, p^o; \boldsymbol{\alpha}) \;&=\; \alpha_1(n^o - n) \;+\; \alpha_2\Big[(n^o)^2 - n^2\Big] \;+\; \alpha_3(n^o - n)p^o \;+\; \alpha_4\Big[(n^o)^3 - n^3\Big] \\ &+\; \alpha_5\Big[(n^o)^2 - n^2\Big]p^o \;+\; \alpha_6(n^o - n)(p^o)^2 \\ h(p, n^o, p^o; \boldsymbol{\beta}) \;&=\; \beta_1(p^o - p) \;+\; \beta_2\Big[(p^o)^2 - p^2\Big] \;+\; \beta_3(p^o - p)n^o \;+\; \beta_4\Big[(p^o)^3 - p^3\Big] \\ &+\; \beta_5\Big[(p^o)^2 - p^2\Big]n^o \;+\; \beta_6(p^o - p)(n^o)^2 \end{aligned} \tag{4.6}$$

with the restrictions $\alpha_3 = \beta_3$, $\alpha_5 = \beta_6$, and $\beta_5 = \alpha_6$, which follow from the underlying structure of g and h (themselves both derived from f) shown in equations (4.4) and (4.5). Since a facility's size may affect its per gallon nutrient removal costs, each term in (4.6) is also interacted with size as measured in millions of gallons per day of design flow (MGD)[5]. Let $\boldsymbol{\alpha}^{size}$ and $\boldsymbol{\beta}^{size}$ denote the set of coefficients on the interactions of each of the terms in (4.6) with the facility's design flow.

The coefficient estimates from the joint estimation of (4.6) can then be used to compute the structural parameters of f according to the definitions of $g(n, n^o, p^o; \boldsymbol{\theta})$ and $h(p, n^o, p^o; \boldsymbol{\theta})$, specifically

$$\begin{aligned}
a &= \alpha_1 + 2\alpha_2\bar{n} + \alpha_3\bar{p} + 3\alpha_4(\bar{n})^2 + 2\alpha_5\bar{n}\bar{p} + \beta_5(\bar{p})^2 \\
b &= \beta_1 + \alpha_3\bar{n} + 2\beta_2\bar{p} + \alpha_5(\bar{n})^2 + 2\beta_5\bar{n}\bar{p} + 3\beta_4(\bar{p})^2 \\
c &= -\alpha_2 - 3\alpha_4\bar{n} - \alpha_5\bar{p} \\
d &= -\alpha_3 - 2\alpha_5\bar{n} - 2\beta_5\bar{p} \\
e &= -\beta_2 - \beta_5\bar{n} - 3\beta_4\bar{p} \\
r &= \alpha_4 \\
w &= \beta_4 \\
u &= \alpha_5 \\
v &= \beta_5
\end{aligned} \tag{4.7}$$

Given the structure of f in (4.3), these estimates of $\boldsymbol{\theta}$ characterize the joint N and P upgrade cost of a representative WWTP for any level of existing treatment capacity $[0, \bar{n}] \times [0, \bar{p}]$. Sedlak [68] reports that baseline N and P concentrations in raw wastewater typically range from 20-40mg/liter and 3-7mg/liter, respectively. For calculating θ in the next section, I assume $\bar{n} = 30$ and $\bar{p} = 5$.

4.4.2 Estimation Results

The results of the joint estimation of the two "partial" incremental nutrient removal cost functions are presented in Table 4.4. Because the independent variables in (4.6) are not necessarily intuitive, the estimates of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are difficult to interpret as shown. However, these regression coefficients can be used to back out structural parameters of f which are more readily interpretable. Using the relationships in (4.7) and assuming a facility size of 7.7MGD (the sample mean) I present the estimates of these parameters in Table 4.4.2. For

[5]Design flow refers to the maximum volume a facility can receive and treat at one time.

n = 632

	Variable					Coef.	Std. Err.	z	$P > \|z\|$
α_1	$(n^o - n)$					520,941.2***	72,233.6	7.21	0.000
α_2	$(n^o)^2 - n^2$					-1,687.4	5,246.4	-0.32	0.748
α_3	$(n^o - n)$	$\times$	(p^o)			19,629.0	18,197.3	1.08	0.281
α_4	$(n^o)^3 - n^3$					-274.7**	130.6	-2.10	0.035
α_5	$(n^o)^2 - n^2$	$\times$	(p^o)			-739.4	512.4	-1.44	0.149
α_6	$(n^o - n)$	$\times$	$(p^o)^2$			3,774.9**	1,920.2	1.97	0.049
α_1^{size}	$(n^o - n)$			$\times$	$size$	16,652.4**	6,553.4	2.54	0.011
α_2^{size}	$(n^o)^2 - n^2$			$\times$	$size$	-2,624.3***	839.1	-3.13	0.002
α_3^{size}	$(n^o - n)$	$\times$	(p^o)	$\times$	$size$	-4,494.1	2,925.3	-1.54	0.124
α_4^{size}	$(n^o)^3 - n^3$			$\times$	$size$	75.4***	26.0	2.90	0.004
α_5^{size}	$(n^o)^2 - n^2$	$\times$	(p^o)	$\times$	$size$	171.8*	90.8	-1.89	0.058
α_6^{size}	$(n^o - n)$	$\times$	$(p^o)^2$	$\times$	$size$	-814.7*	488.0	-1.67	0.095
β_1	$(p^o - p)$					1,859,467.0***	153,415.2	12.12	0.000
β_2	$(p^o)^2 - p^2$					-756,467.6***	50,396.3	-15.01	0.000
β_3	$(p^o - p)$	$\times$	(n^o)			19,629.0	18,197.3	1.08	0.281
β_4	$(p^o)^3 - p^3$					64,331.7***	6,466.1	9.95	0.000
β_5	$(p^o)^2 - p^2$	$\times$	(n^o)			3,774.9**	1,920.1	1.97	0.049
β_6	$(p^o - p)$	$\times$	$(n^o)^2$			-739.4	512.364	-1.44	0.149
β_1^{size}	$(p^o - p)$			$\times$	$size$	9,910.7	21,864.6	0.45	0.650
β_2^{size}	$(p^o)^2 - p^2$			$\times$	$size$	-5,090.7	11,628.8	-0.44	0.662
β_3^{size}	$(p^o - p)$	$\times$	(n^o)	$\times$	$size$	-4,494.1	2,925.3	-1.54	0.124
β_4^{size}	$(p^o)^3 - p^3$			$\times$	$size$	5,345.0**	2,361.4	2.26	0.024
β_5^{size}	$(p^o)^2 - p^2$	$\times$	(n^o)	$\times$	$size$	-814.7*	488.0	-1.67	0.095
β_6^{size}	$(p^o - p)$	$\times$	$(n^o)^2$	$\times$	$size$	171.8*	90.8	1.89	0.058

***significant at the 1% level, **significant at the 5% level, *significant at the 10% level

Table 4.4. Results of the Joint Estimation of System (4.6)

these calculations, I use only the coefficients that are significant at the 10% level.

The parameters on the linear and cubic terms (a and r for N, b and w for P) have the expected positive signs, whereas the parameters on the quadratic terms (c for N, e for P) have negative signs contrary to the expectations in Table 4.5. The negative signs of c and e cause the total cost function to have a negative second derivative (decreasing marginal cost) over some portions of its domain. Figure 4.8 plots the total capital costs (per million

gallons per day of design flow) of achieving a given level of N or P concentration given the parameter values in Table 4.4.2; capital costs are low for high concentrations (low reduction capacity) and high for low nutrient concentrations (high reduction capacity). The portions of decreasing marginal cost can be seen in this figure between 20 and 30 mg/L for N and between 2.5 and 5 mg/L for P (where the cost functions are concave). However, because 91% of facilities in the sample have existing N concentrations below 20 mg/L and 78% have existing P concentrations below 2.5 mg/L, the cost function does have the expected convex shape for the most relevant portions of the domain. In general, the magnitudes of the parameters on P concentration are larger than those on N due to the fact that 1mg/L represents a more significant reduction relative to the baseline concentration of P in wastewater (5 mg/L) than it does relative to the baseline concentration of N (30 mg/L).

The parameters that characterize the cost interdependence of N and P removal are d, u and v. The parameter on the quadratic interaction term, d, has a negative sign, implying that marginal capital costs for the reduction of one pollutant is inversely related to the level of reduction being performed with respect to the other. This statistical result corroborates the engineering principles previously described whereby the investment burden associated with reducing N (P) is lessened by having higher existing capacity to reduce P (N). Based on the positive sign of u, these cost savings diminish as N concentration falls (the benefit of having higher P removal capacity shrinks as N is pushed closer to zero). The opposite appears to be true for P, where based on the negative sign of v, the costs savings from having a higher N removal capacity are enhanced as P concentration falls.

4.5 Sequential vs. Simultaneous Markets

Having established in section 4.3 how combinations of N and P discharge targets map onto the optimal investments at the facility level, I illustrate how economies of scope influence the relative performance of sequential and simultaneous markets.

Consider a joint load target where $\bar{P} = P_1$ (from Table 4.1 and Figure 4.7) and $N_3 < \bar{N} < N_1$ (from Table 4.1 and Figure 4.7). Suppose the regulator proposes a sequential policy where the P target must be achieved first, followed by the N target. Achieving P_1 at least cost requires facility F_1 to reduce its P concentration down to the point at which its marginal P reduction cost just equals the marginal P reduction cost of F_2 at F_2's initial treatment capacity (given by the dashed line in Figure 4.9). Facility F_1's position in n-p space following this optimal reduction is labeled with a "1" in Figure 4.9 to indicate that this is the first of

parameter	value	term from $f(n,p)$
a	640,888.7	$(\bar{n}-n)$
b	2,732,052.5	$(\bar{p}-p)$
c	-15,312.85	$(\bar{n}-n)^2$
d	-58,357.28	$(\bar{n}-n)(\bar{p}-p)$
e	-747,159.3	$(\bar{p}-p)^2$
r	305.9	$(\bar{n}-n)^3$
u	1,322.9	$(\bar{n}-n)^2(\bar{p}-p)$
v	-2,498.3	$(\bar{n}-n)(\bar{p}-p)^2$
w	105,488.2	$(\bar{p}-p)^3$

Table 4.5. Estimated parameters of the joint N and P removal cost function

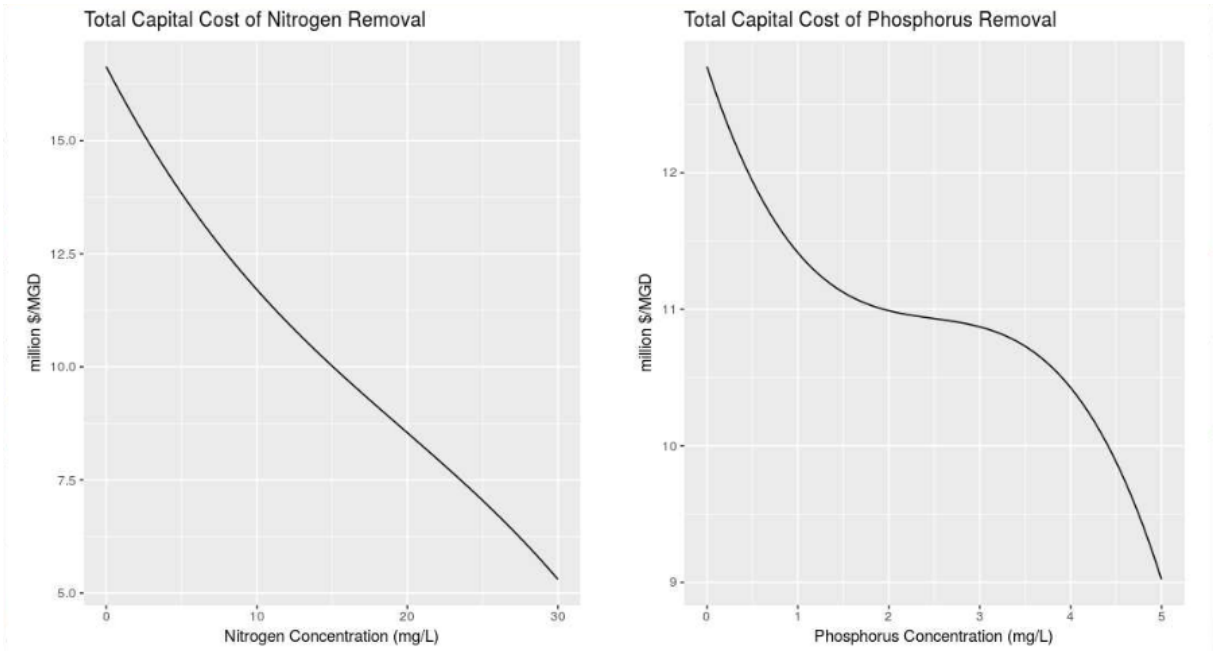

Figure 4.8. Total capital costs per MGD of design flow

two sequential stages. Given this irreversible investment in P reduction at F_1, the most cost-effective means of meeting the N target in the policy's second stage is to invest in upgrades at both facilities F_2 and F_3, bringing each to the positions labeled with a "2" in Figure 4.9.

The pattern of upgrades resulting from the sequential market design in the left-hand panel of Figure 4.9 satisfies the joint N and P load target. But is it cost-effective? The

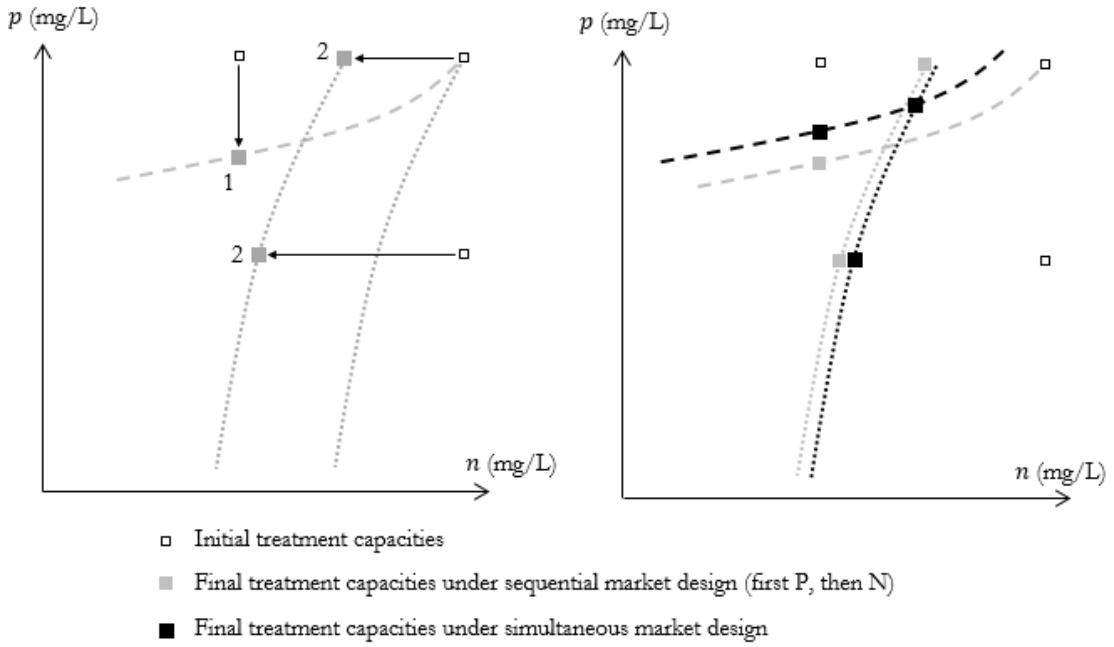

Figure 4.9. Reduction patterns under sequential (left) vs. simultaneous (right) market design

fact that the sequential policy leads to an arrangement where F_2's marginal P reduction cost is lower than F_1's [6] precludes this possibility—costs could be reduced by increasing P discharges at F_1 and decreasing them at F_2. The right-hand panel of Figure 4.9 shows the least-cost arrangement of upgrades that meets the same load targets in the left-hand panel. Due to the economies of scope in the abatement process, this least-cost solution involves F_2 increasing its capacity to treat both P and N relative to the upgrades it performed in the sequential market scenario. Not only are smaller P reductions are required from F_1, but as F_2 increases its capacity to treat P, its marginal costs of reducing N fall and it becomes efficient to reallocate N discharges between facilities F_2 and F_3. This leads to smaller N reductions at F_3 under the simultaneous design relative to the sequential design.

The key feature of this example that caused the divergence between the sequential and simultaneous policies is the the fact that the simultaneous policy requires F_2 to reduce both N and P. Even though F_2 was not initially the least-cost source of P reductions in the first stage, its N reductions in the second stage "retroactively" made it a cost effective source of P reductions along with F_1. Due to the sequential policy's lack of "foresight" with respect to the N reductions required in the second stage, F_1 and F_3 overinvested in reductions of P

[6] note that the marginal P reduction costs of F_1 and F_2 were equal after stage 1, but F_2's P reduction costs drop below this level after F_2 enhances its N removal capacity in stage 2

and N, while F_2 (the source able to benefit from economies of scope) underinvested. This misallocation will arise whenever the joint discharge targets fall within zones 11-14 in Figure 4.7, which includes all the cases in which at least one plant is called on to make reductions in both pollutants.

In cases where the least-cost solution does not require any facility to perform their upgrades with respect to both pollutants, the equilibrium outcome under a sequential design will exactly match that of a simultaneous design. The sets of N and P load targets for which sequential and simultaneous designs perform identically are highlighted in Figure 4.10. This region of the N-P discharge space corresponds to regime #'s 2-10 in Table 4.1, which includes the discharge targets for which the facilities that upgrade with respect to N do not upgrade with respect to P and vice versa. Economies of scope never enter the picture in these instances.

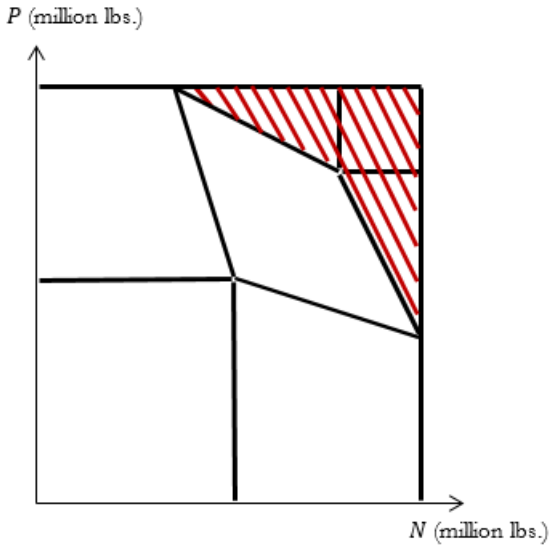

Figure 4.10. Sets of joint N and P discharge targets for which sequential markets perform identically to simultaneous markets

Having specified a joint cost function for N and P removal, it's possible to compute the excess cost of a sequential trading system relative to a simultaneous one, given the initial treatment capacities of the polluting sources. First, consider the optimal arrangement of upgrades for any joint discharge target under a sequential policy that regulates P in stage 1

and N in stage 2. The stage 1 problem is

$$\min_{p_1,p_2,p_3} \sum_{i=1}^{3} \omega_i \, c(n_i^o, p_i) \quad \text{subject to} \quad \sum_{i=1}^{3} \omega_i \, p_i \leq \bar{P}, \quad p_i \leq p_i^o$$

Let $\tilde{p}_i$ denote the least-cost values of P concentration at each facility, given that each facility's N reduction capacity is fixed at n_i^o. Due to the irreversibility of investments in nutrient removal capital, each facility's P reduction capacity is fixed at $\tilde{p}_i$ in stage 2 where the problem is

$$\min_{n_1,n_2,n_3} \sum_{i=1}^{3} \omega_i \, c(n_i, \tilde{p}_i) \quad \text{subject to} \quad \sum_{i=1}^{3} \omega_i \, n_i \leq \bar{N}, \quad n_i \leq n_i^o$$

Let $\tilde{n}_i$ denote the least-cost values of N concentration at each facility, given their prior investments in P reduction. The final nutrient concentrations that emerge from this sequential policy are $(\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{p}})$.

Next, consider the arrangement of upgrades that will result under a simultaneous policy for the same joint discharge target. The problem of choosing N and P reductions together to satisfy the discharge targets $\bar{N}$ and $\bar{P}$ if formulated in (4.1). Let $(\boldsymbol{n}^*, \boldsymbol{p}^*)$ represent the final nutrient concentrations that solve this problem. Letting Δ denote the excess cost associated with the sequential policy, the inefficiency from a sequential market design when pollution reductions are irreversible is given by

$$\Delta \;=\; \pi(\tilde{\boldsymbol{n}}, \tilde{\boldsymbol{p}}) \;-\; \pi(\boldsymbol{n}^*, \boldsymbol{p}^*) \tag{4.8}$$

As previously discussed, Δ will be greater than zero when the dual N and P targets fall within zones 11, 12, 13, or 14 (Figure 4.7, left). Because π is a function of the structural parameters of the joint cost function estimated in section 4.4, this framework provides a way to characterize how the degree of interdependence in the reduction of two pollutants affects the efficiency of a sequential policy implementation (relative to simultaneous implementation) for any joint N and P discharge standard.

4.6 Concluding Remarks

The need to regulate the joint levels of multiple interdependent pollutants has been established in well known air pollution cases [33, 56], and scientific evidence suggests the same is true for the case of N and P pollution in aquatic environments [12, 64]. Achieving pollution targets through market mechanisms has gained popularity in recent decades, though implementing markets for multiple pollutants introduces new design questions [1, 57]. One of these new questions relates to the relative timing of implementation for the control of two or more pollutants. Lence et al. [52] showed that in cases where the costs of reducing two or more pollutants are interdependent and the reduction process is defined by discrete, irreversible capital investments, the relative timing of policies for the control of two or more pollutants may affect the final costs of meeting a joint discharge standard. While these authors point out the existence of this issue, they do not formally characterize the relationship between joint pollution reduction costs and the importance of getting the timing of policies right.

To fill this gap, this essay develops a framework for modeling the joint costs of reducing two pollutants at the same time. The approach estimates the structural parameters of a joint cost function to test for the presence of cost interdependence in the wastewater sector, finding statistical evidence consistent with what is known about the engineering of joint N and P reduction. This cost function can then be used to estimate the incremental upgrade costs for N or P, given the facility's existing level of treatment for either pollutant. While this technique is applied to the wastewater sector, decisions about allocating resources to the joint production of environmental goods are prevalent in current environmental economics research [82, 53, 72]. This framework could be adapted to other contexts to determine the optimal policies for land conservation, habitat protection, and carbon sequestration.

Chapter 5

Conclusion

This dissertation tackles several elements of the physical complexity of the nutrient pollution problem, an issue that has been receiving enormous policy attention in past decades. Each essay studies the implications of one or more of these elements for the optimal allocation of resources toward the reduction of nutrient discharges and, in the case of chapters 3 and 4, the design of policy to implement these allocations.

Chapter 2 characterizes the nutrient pollution problem as one of managing the growth of two stocks over time—one, a stock of nutrients whose excess leads to the production of toxic algae and the depletion of oxygen supplies in aquatic environments, and the other a stock of capital used in the reduction of these nutrients from municipal wastewater. In theory, an optimal strategy would vary the levels of nutrient reduction effort based on the urgency of the environmental harm and the costliness of adjusting the size of the nutrient removal capital stock. In practice, these results determine that the advantages of a time-varying are relatively small (between 0.05 and 4%, depending on the delay associated with reductions in the agricultural sector), and that a policy of targeting a constant nutrient loads over time may do very well.

This result provides a sound basis for market design proposed in chapter 3, which shows that optimally managing the nutrient loads from an indefinite number of sources with dissimilar lag lengths would require policy to evolve over time with at least as many phases as the number of unique lag lengths. For the vast number of sources spanning a watershed the size of the Chesapeake Bay's, such a policy would be prohibitively complex. Instead, the essay proposes a design that balances the need to adjust for lag length with a competing need for simplicity. In this design, the steady state nutrient load limit is imposed from the

start and the discharges from one source may be allowed to offset those of another at a rate given by a simple function of the discount rate and their lag length disparity. Such a policy will be inefficient along the path to the steady state nutrient stock, but as chapter 2 shows, this inefficiency may not be very substantial.

Finally, while chapter 4 relates specifically to the cost-effective control of N and P in the point source sector, the modeling framework may have broader applications to other areas of environmental policy that deal with multiple environmental goods produced jointly. It is hoped that further refinement of this concept will help structure further analysis of these issues.

Appendix A

Lagrange Multipliers

A.1 Deriving λ_t from (2.5c)

Recall that λ_t is the shadow price of increasing the pollution stock in a given period. To see where its structure comes from, consider the effects on Z from an infinitesimally small "pulse" of nutrients added to the pollution stock in period t. This pulse will first increase current period damage costs by $D'(S_t)/(1+r)^t$ and then contribute to future damage costs as the fraction $(1-\delta)$ carries over into each subsequent period. Observe from (2.2) that each pollution stock subsequent to S_t can be expressed

$$\begin{aligned}
S_{t+1} &= S_t(1-\delta) &&+ x_t, \\
S_{t+2} &= S_t(1-\delta)^2 &&+ x_t(1-\delta) &&+ x_{t+1}, \\
S_{t+3} &= S_t(1-\delta)^3 &&+ x_t(1-\delta)^2 &&+ x_{t+1}(1-\delta) &&+ x_{t+2},
\end{aligned}$$

and so on, where the x's represent new loads contributing to the stock in subsequent periods. Following S_t through time in this way highlights the fact that differentiating Z with respect to S_t (using the chain rule) will yield

$$\frac{D'(S_t)}{(1+r)^t} + (1-\delta)\frac{D'(S_{t+1})}{(1+r)^{t+1}} + (1-\delta)^2\frac{D'(S_{t+2})}{(1+r)^{t+2}} + (1-\delta)^3\frac{D'(S_{t+3})}{(1+r)^3} + \ldots = -\lambda_t$$

A negative sign precedes λ_t in the expression above because cost "savings" are negative for an increase to period t's pollution stock.

A.2 Deriving μ_t from (2.5d)

Recall that μ_t is the shadow price of increasing the abatement capital stock in a given period. Consider the effects on Z of adding an infinitesimally small mass of capital to the WWTP's stock in period t. This mass affects costs through two channels. First, in a similar pattern as in A.1, the mass will affect current period operating costs in the amount $v'(K)/(1+r)^t$ and then affect future operating costs as the fraction $(1-\gamma)$ carries over into each subsequent period. Observe from (2.3) that each abatement capital stock subsequent to K_t can be expressed

$$\begin{aligned}
K_{t+1} &= K_t(1-\gamma) &&+ I_t, \\
K_{t+2} &= K_t(1-\gamma)^2 &&+ I_t(1-\gamma) &&+ I_{t+1}, \\
K_{t+3} &= K_t(1-\gamma)^3 &&+ I_t(1-\gamma)^2 &&+ I_{t+1}(1-\gamma) &&+ I_{t+2},
\end{aligned}$$

and so on. Differentiating the sets of $v(K_t)$, $v(K_{t+1})$, $v(K_{t+2})$, etc. that appear in Z will yield

$$\frac{v'(K_t)}{(1+r)^t} + (1-\gamma)\frac{v'(K_{t+1})}{(1+r)^{t+1}} + (1-\gamma)^2\frac{v'(K_{t+2})}{(1+r)^{t+2}} + (1-\gamma)^3\frac{v'(K_{t+3})}{(1+r)^3} + \dots$$

which constitutes the first bracketed term in (2.7).

Through a second channel, increasing K_t will cause WWTP's contemporaneous and subsequent nutrient discharges to fall and affect damage costs through S. To see this, consider that pollution stocks S_{t+1}, S_{t+2}, S_{t+3}, etc. can be expressed

$$\begin{aligned}
S_{t+1} &= (\bar{n} - K_t)\omega \\
S_{t+2} &= (\bar{n} - K_t)\omega(1-\delta) &&+ [\bar{n} - K_t(1-\gamma)]\omega \\
S_{t+3} &= (\bar{n} - K_t)\omega(1-\delta)^2 &&+ [\bar{n} - K_t(1-\gamma)]\omega(1-\delta) &&+ [\bar{n} - K_t(1-\gamma)^2]\omega
\end{aligned}$$

and so on. I've omitted new investment and agricultural load are omitted for clarity—since they do not interact at all with K_t, they will only affect the $D'(S)$ by altering the values of

S. Differentiating Z with respect to S_t (using the chain rule) will yield

$$\begin{aligned} -\omega\frac{D'(S_{t+1})}{(1+r)^{t+1}} &- \omega\frac{D'(S_{t+2})}{(1+r)^{t+2}}(1-\delta) &- \omega\frac{D'(S_{t+3})}{(1+r)^{t+3}}(1-\delta)^2 &- \ldots \\ &- \omega\frac{D'(S_{t+2})}{(1+r)^{t+2}}(1-\gamma) &- \omega\frac{D'(S_{t+3})}{(1+r)^{t+3}}(1-\delta)(1-\gamma) &- \ldots \\ & & - \omega\frac{D'(S_{t+3})}{(1+r)^{t+3}}(1-\gamma)^2 &- \ldots \end{aligned}$$

and so on. By (2.6), this collection of terms is equal to

$$\omega\lambda_{t+1} + \omega\lambda_{t+2}(1-\gamma) + \omega\lambda_{t+3}(1-\gamma)^2 + \ldots$$

which constitutes the second bracketed term in (2.7).

Appendix B

Point Source Reduction Costs

Horan and Shortle [39] specify

$$\phi(PS) \;=\; u(\bar{PS} - PS)^3$$

where reduction costs from the point source sector, ϕ, depend on the total mass of the sector's nutrient emissions, PS, relative to a baseline $\bar{PS}$. I first convert the units in this function from mass to concentration, then I split the function linearly into a capital and operating cost component, and finally I scale up capital costs to reflect the cumulative cost of building the capital stock to the level implied by the nutrient discharge level.

B.1 Unit Conversion

In this step I translate the point source cost function $\phi(PS)$ to an equivalent expression

$$\psi(n) \;=\; u'(\bar{n} - n)^3$$

for which costs depend on the nutrient concentration in the discharged wastewater, n, relative to a baseline concentration $\bar{n}$. To convert the mass-based ϕ to the concentration-based ψ, I decompose PS into the product of two components–the concentration of nitrogen in the discharged effluent, n (mg per liter), and the wastewater volume, ω (million gallons per day), conveyed through the plant. With the appropriate conversion factors, PS relates to n and ω as follows:

$$\underset{\left(\frac{\text{million lbs.}}{\text{yr.}}\right)}{PS} = \underset{\left(\frac{\text{mg}}{\text{liter}}\right)}{n} \times \underset{\left(\frac{\text{million gal.}}{\text{day}}\right)}{\omega} \times \underset{\left(\frac{\text{lbs./million gal.}}{\text{mg/liter}}\right)}{8.3454} \times \underset{\left(\frac{\text{days}}{\text{yr.}}\right)}{365} \times \underset{\left(\frac{\text{million lbs.}}{\text{lbs.}}\right)}{10^{-6}} \tag{B.1}$$

Letting $u' = u\left[\frac{\omega(8.3454)(365)}{10^6}\right]^3$, and using (B.1) to obtain $\bar{n} = \bar{PS}\frac{10^6}{\omega(8.3454)(365)}$ and $n = PS\frac{10^6}{\omega(8.3454)(365)}$, the conversion from PS to n in ϕ proceeds as follows:

$$\begin{aligned}
\phi(PS) &= u(\bar{PS} - PS)^3 \left[\frac{10^6}{\omega(8.3454)(365)}\right]^3 \left[\frac{\omega(8.3454)(365)}{10^6}\right]^3 \\
&= u\left(\bar{PS}\frac{10^6}{\omega(8.3454)(365)} - PS\frac{10^6}{\omega(8.3454)(365)}\right)^3 \left[\frac{\omega(8.3454)(365)}{10^6}\right]^3 \\
&= u'\left(\bar{n} - n\right)^3 \\
&= \psi(n)
\end{aligned}$$

For a given ω, the cost of an additional x-pound nutrient reduction may therefore be expressed as either

1. the increase in ϕ from an $\frac{x}{10^6}$-unit reduction in PS or
2. the increase in ψ due to an $\frac{x}{\omega(8.3454)(365)}$-unit reduction in n

The choice variable n can be interpreted as the minimum nutrient concentration the WWTP can achieve at a point in time, and is associated with a particular nutrient removal process that is fixed in the short run. To attain further capacity for nutrient reduction, the plant must upgrade to a new process that is associated with a lower post-treatment concentration.

B.2 Capital and O&M Costs

Building the capacity to reduce the nutrient concentration down to a particular level is distinct from actually implementing the treatment itself. The former involves making long-lived capital investments, while the latter involves purchasing the inputs required to operate and maintain the facility. I modify $\psi(n)$ to account for two distinct cost components—costs incurred during daily operation (which depend on the actual volume of treated wastewater) versus those incurred to acquire the structures and equipment that allow a WWTP to treat

incoming wastewater up to a particular standard. I define $\alpha \in (0,1)$ as the fraction of annualized costs relating to capital acquisition and split $\psi(n)$ into capital costs $\alpha u'(\bar{n}-n)^3$ and operating and maintenance costs (O&M) costs $(1-\alpha)u'(\bar{n}-n)^3$, both measured on an annual basis. Altogether, the annual cost of reducing nutrient concentration below $\bar{n}$ is given by

$$\psi(n) \;=\; \alpha u'(\bar{n}-n)^3 \;+\; (1-\alpha)u'(\bar{n}-n)^3 \tag{B.2}$$

I further modify (B.2) to explicitly account for the fact that capital costs depend on the plant's design flow (i.e. the maximum volume of wastewater a plant can receive at a point in time), whereas O&M costs depend on the actual volume of wastewater conveyed through the plant. Letting Ω denote design flow and ω denote actual flow. I construct per-unit versions of the cost coefficients in (B.2), $c^{\text{CAP}} = \frac{\alpha u'}{\Omega}$ and $c^{\text{OM}} = \frac{(1-\alpha)u'}{\omega}$, and re-express point source reduction costs as

$$\psi(n) \;=\; c^{\text{CAP}}(\bar{n}-n)^3\Omega \;+\; c^{\text{OM}}(\bar{n}-n)^3\omega \tag{B.3}$$

The parameter c^{CAP} represents annual capital costs per MGD of design flow, and c^{OM} represents annual O&M costs per MGD of treated wastewater. In reality, economies of scale may cause upgrade costs per MGD to fall as plant size or treated volume increases.

B.3 Converting Annual Capital Costs to Lump Sum

While expressing capital costs in annualized terms often make these figures more interpretable (especially when pairing them with $O\&M$ cost figures), modeling them as such abstracts away from the inherent dependencies between capacity choices at different points in time. The WWTP's nutrient removal capacity, n, is a long-run decision involving the installation or removal of particular capital goods (mixers, basins, aeration tanks). These installations typically entail large resource costs that are financed with annual installments spanning the life-cycle of the capital goods (often 20-30 years). The higher the plant's nutrient removal capacity, the more capital resources are involved and the larger the annual installments will be. The way ψ is specified in (B.2), a plant with a particular capacity n' could choose $n'' > n'$ in the following year and thereby reduce their capital costs. This formulation allows the plant to choose a new capacity each year and pay only the annualized capital costs associated with this treatment level. The plant essentially "rents" a year's worth of capacity every year. This ability to adjust n up and down each year, incurring only a year's worth of annualized costs, represents a case of perfectly malleable capital.

To capture the "lumpiness" of a treatment capacity upgrade, I define a factor L with which to scale up c^{CAP} so that the capital costs of a particular treatment level include the full present value resource costs. Including this fact, installing the capacity to achieve a nutrient concentration of n requires spending $L \cdot c^{\text{CAP}}(\bar{n} - n)^3\Omega$ up front. Once installed however, the plant's costs in each subsequent year include only the O&M costs associated with n. While treatment capacity may deteriorate over time. ϕ's current specification implicitly assumes zero depreciation. This assumption will be relaxed later.

Letting $a = L \cdot c^{\text{CAP}}$ and taking the modifications above together with (2.1), I define the total capital cost function

$$f(K) \;=\; a(K)^3\Omega \tag{B.4}$$

Capital costs increase linearly in the plant's design flow, Ω. From (B.4), (2.12), and (2.3) I derive investment costs as

$$\begin{aligned} g(I_t, K_{t-1}) \;&=\; a\Big[(1-\gamma)K_{t-1} \;+\; I_t\Big]^3\Omega \;\;-\;\; a\Big[(1-\gamma)K_{t-1}\Big]^3\Omega \\ &=\; a\Big[I_t^3 \;+\; 3I_t^2(1-\gamma)K_{t-1} \;+\; 3I_t(1-\gamma)^2K_{t-1}^2\Big]\Omega \end{aligned} \tag{B.5}$$

Bibliography

[1] Stefan Ambec and Jessica Coria. Prices vs quantities with multiple pollutants. *Journal of Environmental Economics and Management*, 66(1):123–140, 2013.

[2] Donald M Anderson, Patricia M Glibert, and Joann M Burkholder. Harmful algal blooms and eutrophication: nutrient sources, composition, and consequences. *Estuaries*, 25(4):704–726, 2002.

[3] Emmanuelle Augeraud-Véron and Marc Leandri. Optimal pollution control with distributed delays. *Journal of Mathematical Economics*, 55:24–32, 2014.

[4] William J Baumol and Wallace E Oates. The theory of environmental policy: Externalities, public outlays, and the quality of life, 1975.

[5] Brian Beavis and Martin Walker. Interactive pollutants and joint abatement costs: achieving water quality standards with effluent charges. *Journal of Environmental Economics and Management*, 6(4):275–286, 1979.

[6] Donald F Boesch. Challenges and opportunities for science in reducing nutrient overenrichment of coastal ecosystems. *Estuaries*, 25(4):886–900, 2002.

[7] E Bonsdorff, EM Blomqvist, J Mattila, and A Norkko. Coastal eutrophication: causes, consequences and perspectives in the archipelago areas of the northern baltic sea. *Estuarine, coastal and shelf science*, 44:63–72, 1997.

[8] Cyril Bourgeois and Pierre-Alain Jayet. Regulation of relationships between heterogeneous farmers and an aquifer accounting for lag effects. *Australian Journal of Agricultural and Resource Economics*, 60(1):39–59, 2016.

[9] WR Boynton, JH Garber, R Summers, and WM Kemp. Inputs, transformations, and transport of nitrogen and phosphorus in chesapeake bay and selected tributaries. *Estuaries*, 18(1):285–314, 1995.

[10] KA Cherry, M Shepherd, PJA Withers, and SJ Mooney. Assessing the effectiveness of actions to mitigate nutrient loss from agriculture: A review of methods. *Science of the Total Environment*, 406(1-2):1–23, 2008.

[11] Baltic Marine Environment Protection Commission et al. *First Periodic Assessment of the State of the Marine Environment of the Baltic Sea Area, 1980-1985*. Baltic Marine Environment Protection Commission, Helsinki Commission, 1986.

[12] Daniel J Conley, Hans W Paerl, Robert W Howarth, Donald F Boesch, Sybil P Seitzinger, Karl E Havens, Christiane Lancelot, Gene E Likens, et al. Controlling eutrophication: nitrogen and phosphorus. *Science*, 323(5917):1014–1015, 2009.

[13] Zach Corrigan. The case against water quality trading. *Nat. Resources & Env't*, 30:15, 2015.

[14] John H Dales. *Pollution, property & prices: an essay in policy-making and economics*. Edward Elgar Publishing, 1968.

[15] Victor N de Jonge, M Elliott, and E Orive. Causes, historical development, effects and future challenges of a common environmental problem: eutrophication. In *Nutrients and Eutrophication in Estuaries and Coastal Waters*, pages 1–19. Springer, 2002.

[16] Robert J Diaz and Rutger Rosenberg. Spreading dead zones and consequences for marine ecosystems. *science*, 321(5891):926–929, 2008.

[17] Taylor Dungjen and David Patch. Toledo-area water advisory expected to continue through sunday as leaders await tests; water stations to remain open. *The Blade*, August 2, 2014.

[18] Katarina Elofsson. Cost-effective control of interdependent water pollutants. *Environmental Management*, 37(1):54–68, 2006.

[19] Alfred Endres. Charges, permits and pollutant interactions. *Eastern Economic Journal*, 12(3):327–336, 1986.

[20] Yuri Ermoliev, Mikhail Michalevich, and Andries Nentjes. Markets for tradeable emission and ambient permits: a dynamic approach. *Environmental and Resource Economics*, 15(1):39–56, 2000.

[21] Ita Falk and Robert Mendelsohn. The economics of controlling stock pollutants: an efficient strategy for greenhouse gases. *Journal of Environmental Economics and Management*, 25(1):76–88, 1993.

[22] Hongli Feng, Lyubov A Kurkalova, Catherine L Kling, and Philip W Gassman. Environmental conservation in agriculture: Land retirement vs. changing practices on working land. *Journal of Environmental Economics and Management*, 52(2):600–614, 2006.

[23] Karen Fisher-Vanden and Sheila Olmstead. Moving pollution trading from air to water: potential, problems, and prognosis. *The Journal of Economic Perspectives*, 27(1):147–171, 2013.

[24] Nutrient Reduction Technology Cost Task Force. Nutrient reduction technology cost estimations for point sources in the Chesapeake Bay watershed. Technical report, Chesapeake Bay Program, 2002.

[25] Finn R Førsund and Eric Nævdal. Efficiency gains under exchange-rate emission trading. *Environmental and Resource Economics*, 12(4):403–423, 1998.

[26] Elise M Fulstone. Effluent trading: legal constraints on the implementation of market-based effluent trading programs under the Clean Water Act. *Envtl. Law.*, 1:459, 1994.

[27] James N Galloway, John D Aber, Jan Willem Erisman, Sybil P Seitzinger, Robert W Howarth, Ellis B Cowling, and B Jack Cosby. The nitrogen cascade. *AIBS Bulletin*, 53(4):341–356, 2003.

[28] Sara Gates. Toledo warns area residents not to drink water after city supply tests positive for toxin. *Huffington Post*, August 2, 2014.

[29] Patricia M Glibert, Robert Magnien, Michael W Lomas, Jeffrey Alexander, Chunlei Tan, Erin Haramoto, Mark Trice, and Todd M Kana. Harmful algal blooms in the chesapeake and coastal bays of maryland, usa: Comparison of 1997, 1998, and 1999 events. *Estuaries*, 24(6):875–883, 2001.

[30] Lawrence H Goulder. Markets for pollution allowances: what are the (new) lessons? *Journal of Economic Perspectives*, 27(1):87–102, 2013.

[31] Marie Gren, Paul Jannke, and Katarina Elofsson. Cost-effective nutrient reductions to the baltic sea. *Environmental and Resource Economics*, 10(4):341–362, 1997.

[32] Ronald C Griffin. Environmental policy for spatial and persistent pollutants. *Journal of Environmental Economics and Management*, 14(1):41–53, 1987.

[33] Arie J Haagen-Smit. Chemistry and physiology of los angeles smog. *Industrial & Engineering Chemistry*, 44(6):1342–1346, 1952.

[34] James D Hagy, Walter R Boynton, Carolyn W Keefe, and Kathryn V Wood. Hypoxia in chesapeake bay, 1950–2001: Long-term change in relation to nutrient loading and river flow. *Estuaries*, 27(4):634–658, 2004.

[35] Rob Hart. Dynamic pollution control—time lags and optimal restoration of marine ecosystems. *Ecological Economics*, 47(1):79–93, 2003.

[36] David A Hennessy and Hongli Feng. When should uncertain nonpoint emissions be penalized in a trading program? *American Journal of Agricultural Economics*, 90(1):249–255, 2008.

[37] Richard D Horan and James S Shortle. When two wrongs make a right: Second-best point-nonpoint trading ratios. *American Journal of Agricultural Economics*, 87(2):340–352, 2005.

[38] Richard D Horan and James S Shortle. Economic and ecological rules for water quality trading. *Journal of the American Water Resources Association*, 47(1):59–69, 2011.

[39] Richard D Horan and James S Shortle. Endogenous risk and point-nonpoint uncertainty trading ratios. *American Journal of Agricultural Economics*, 99(2):427–446, 2017.

[40] Richard D Horan, James S Shortle, and David G Abler. Point-nonpoint nutrient trading in the susquehanna river basin. *Water Resources Research*, 38(5), 2002.

[41] Oliver A Houck. The clean water act returns (again): Part i, tmdls and the chesapeake bay. *Envtl. L. Rep. News & Analysis*, 41:10–208, 2011.

[42] Robert W Howarth and Roxanne Marino. Nitrogen as the limiting nutrient for eutrophication in coastal marine ecosystems: evolving views over three decades. *Limnology and Oceanography*, 51(1part2):364–376, 2006.

[43] Ming-Feng Hung and Daigee Shaw. A trading-ratio system for trading water pollution discharge permits. *Journal of Environmental Economics and Management*, 49(1):83–102, 2005.

[44] Joel Beauvais. *Renewed Call To Action to Reduce Nutrient Pollution and Support for Incremental Actions to Protect Water Quality and Public Health.* Memo from the U.S. Environmental Protection Agency, Office of Water, 2016.

[45] Cy Jones, Evan Branosky, Mindy Selman, and Michelle Perez. How nutrient trading could help restore the chesapeake bay. *World Resources Institute Working Paper*, 2010.

[46] Zach Kaufman, David Abler, James Shortle, Jayson Harper, James Hamlett, and Peter Feather. Agricultural costs of the chesapeake bay total maximum daily load. *Environmental science & technology*, 48(24):14131–14138, 2014.

[47] Emmett Keeler, Michael Spence, and Richard Zeckhauser. The optimal control of pollution. *Journal of Economic Theory*, 4(1):19–34, 1972.

[48] W Michael Kemp, Walter R Boynton, Jason E Adolf, Donald F Boesch, William C Boicourt, Grace Brush, Jeffrey C Cornwell, Thomas R Fisher, Patricia M Glibert, Jim D Hagy, et al. Eutrophication of chesapeake bay: historical trends and ecological interactions. *Marine Ecology Progress Series*, 303(21):1–29, 2005.

[49] Alan J Krupnick, Wallace E Oates, and Eric Van De Verg. On marketable air-pollution permits: The case for a system of pollution offsets. *Journal of Environmental Economics and Management*, 10(3):233–247, 1983.

[50] Timo Kuosmanen and Marita Laukkanen. (In) efficient environmental policy with interacting pollutants. *Environmental and Resource Economics*, 48(4):629–649, 2011.

[51] Marita Laukkanen and Anni Huhtala. Optimal management of a eutrophied coastal ecosystem: balancing agricultural and municipal abatement measures. *Environmental and Resource Economics*, 39(2):139–159, 2008.

[52] Barbara J Lence, J Wayland Eheart, and E Downey Brill Jr. Cost efficiency of transferable discharge permit markets for control of multiple pollutants. *Water Resources Research*, 24(7):897–905, 1988.

[53] Adam H Lentz, Amy W Ando, and Nicholas Brozović. Water quality trading with lumpy investments, credit stacking, and ancillary benefits. *JAWRA Journal of the American Water Resources Association*, 50(1):83–100, 2014.

[54] Arun S Malik, David Letson, and Stephen R Crutchfield. Point/nonpoint source trading of pollution abatement: choosing the right trading ratio. *American Journal of Agricultural Economics*, 75(4):959–967, 1993.

[55] Donald W Meals, Steven A Dressing, and Thomas E Davenport. Lag time in water quality response to best management practices: A review. *Journal of environmental quality*, 39(1):85–96, 2010.

[56] Peter Michaelis. Global warming: efficient policies in the case of multiple pollutants. *Environmental and Resource Economics*, 2(1):61–77, 1992.

[57] Juan-Pablo Montero. Multipollutant markets. *RAND Journal of Economics*, pages 762–774, 2001.

[58] W David Montgomery. Markets in licenses and efficient pollution control programs. *Journal of Economic Theory*, 5(3):395–418, 1972.

[59] Ulf Moslener and Till Requate. Optimal abatement in dynamic multi-pollutant problems when pollutants can be complements or substitutes. *Journal of Economic Dynamics and Control*, 31(7):2293–2316, 2007.

[60] Ulf Moslener and Till Requate. The dynamics of optimal abatement strategies for multiple pollutants—an illustration in the greenhouse. *Ecological Economics*, 68(5):1521–1534, 2009.

[61] National Research Council (NRC). *Clean Coastal Waters:: Understanding and Reducing the Effects of Nutrient Pollution*. National Academies Press, 2000.

[62] Pennsylvania Department of Environmental Protection. Phase 2 watershed implementation plan nutrient trading supplement, 2016.

[63] National Research Council (US). Committee on the Evaluation of Chesapeake Bay Program Implementation for Nutrient Reduction to Improve Water Quality. *Achieving nutrient and sediment reduction goals in the Chesapeake Bay: An evaluation of program strategies and implementation*. National Academies Press, 2011.

[64] Hans W Paerl. Controlling eutrophication along the freshwater–marine continuum: dual nutrient (n and p) reductions are essential. *Estuaries and Coasts*, 32(4):593–601, 2009.

[65] Ann Powers. The Connecticut Nitrogen Exchange Program. *Penn St. Envtl. L. Rev.*, 14:195, 2005.

[66] Yukako Sado, Richard N Boisvert, and Gregory L Poe. Potential cost savings from discharge allowance trading: A case study and implications for water quality trading. *Water Resources Research*, 46(2), 2010.

[67] DL Saunders and J Kalff. Nitrogen retention in wetlands, lakes and rivers. *Hydrobiologia*, 443(1-3):205–212, 2001.

[68] Richard I Sedlak. *Phosphorus and nitrogen removal from municipal wastewater: principles and practice*. CRC press, 1991.

[69] James Shortle. The allocative efficiency implications of water pollution abatement cost comparisons. *Water Resources Research*, 26(5):793–797, 1990.

[70] Rajesh Singh, Quinn Weninger, and Matthew Doyle. Fisheries management with stock growth uncertainty and costly capital adjustment. *Journal of Environmental Economics and Management*, 52(2):582–599, 2006.

[71] Val H Smith, G David Tilman, and Jeffery C Nekola. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. *Environmental pollution*, 100(1):179–196, 1999.

[72] Loris Strappazzon, Arthur Ha, Mark Eigenraam, Charlotte Duke, and Gary Stoneham. Efficiency of alternative property right allocations when farmers produce multiple environmental goods under the condition of economies of scope. *Australian Journal of Agricultural and Resource Economics*, 47(1):1–27, 2003.

[73] Jordan F Suter, John M Spraggon, and Gregory L Poe. Thin and lumpy: an experimental investigation of water quality trading. *Water Resources and Economics*, 1:36–60, 2013.

[74] Thomas H Tietenberg. *Emissions trading: an exercise in reforming pollution policy*. Resources for the Future, 1985.

[75] U.S. Environmental Protection Agency. Water quality trading policy. Technical report, EPA, Office of Water, 2003.

[76] U.S. Environmental Protection Agency. A compilation of cost data associated with the impacts and control of nutrient pollution. Technical report, US EPA, Office of Water, Washington D.C., 2015.

[77] Frederick Van Der Ploeg and Cees Withagen. Pollution control and the ramsey problem. *Environmental and Resource Economics*, 1(2):215–236, 1991.

[78] George Van Houtven, Ross Loomis, Justin Baker, Robert Beach, and Casey Sara. Nutrient credit trading for the chesapeake bay: An economic study. Technical report, RTI International, 2012.

[79] Thomas von Ungern-Sternberg. Environmental protection with several pollutants: on the division of labor between natural scientists and economists. *Journal of Institutional and Theoretical Economics (JITE)/Zeitschrift für die gesamte Staatswissenschaft*, pages 555–567, 1987.

[80] Michael Wines. Behind toledo's water crisis, a long-troubled lake erie. *The New York Times*, August 4, 2014.

[81] Ralph Winkler. A note on the optimal control of stocks accumulating with a delay. *Macroeconomic Dynamics*, 15(04):565–578, 2011.

[82] Richard T Woodward. Double-dipping in environmental markets. *Journal of Environmental Economics and Management*, 61(2):153–169, 2011.

Vita

Aaron Cook

Education

Penn State University
May 2019 — Ph.D., Agricultural, Environmental, and Regional Economics

Purdue University
Aug 2013 — M.S., Agricultural Economics

University of Wisconsin
May 2011 — B.S., Agricultural and Applied Economics; B.A., Southeast Asian Studies

Publications

Water Quality Trading. James Shortle and Aaron Cook. *Routledge Handbook of Agricultural Economics*. 2018.

How do African Households Respond to Changes in Current and Past Weather Patterns? A Structural Panel Data Analysis from Malawi Juan Sesmero, Jacob Ricker-Gilbert, and Aaron Cook. *American Journal of Agricultural Economics*. 100 (1):115-144. 2017.

Presentations

Dynamic Management of Nutrient Pollution in Aquatic Environments: An Evaluation of the Chesapeake Bay TMDL
Water Insights Seminar, Penn State University, October 2018

Trade Ratio Design for Lagged Pollution
Energy and Envt'l Econ and Policy Seminar, Penn State, September 2018

Intertemporal Trading Ratios for Nutrient Pollution Control
AAEA Annual Meeting, Washington, D.C, August 2018

A Second-Best Market Design for Lagged, Persistent Pollutants
Chesapeake Research & Modeling Symposium, Annapolis, MD, June 2018

Teaching

CED 201: Environmental and Natural Resource Economics
AGBM 101: Economic Principles of Agribusiness Management

Professional Memberships

American Economic Association, **Agricultural and Applied Economics Association**

Journal Referee

Climatic Change, **American Journal of Agricultural Economics**