Selected Paper prepared for presentation at the 2018 Agricultural & Applied Economics Association Annual Meeting, Washington, DC, August 5-7, 2018

Intertemporal Trading Ratios for Nutrient Pollution Control

Aaron Cook*

James Shortle

June 21, 2018

Abstract

There is significant interest in the use of market mechanisms for controlling nutrient pollution, one of the most challenging threats to water quality in the United States and elsewhere. This type of pollution often is characterized by considerable lag times between discharge from the pollution source and delivery to the impaired waters. We investigate the implications of these lags for efficient pollution reduction markets and compare two alternative market designs: 1) forward markets where participants trade pollution deliveries directly and 2) a trading ratio system where they trade contemporaneous discharges. While a system of first-best trade ratios is complex in early periods, this system can produce the optimal steady state loads using a simple trading rule. We also find that while first-best trade ratios are 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.

Keywords: nutrient trading, water quality trading, trading ratios, Chesapeake Bay

JEL Codes: Q25, Q53, Q58

*Correspondence: Department of Agricultural Economics, Sociology, and Education, 301 Armsby Building, Penn State University, University Park, PA, 16802. Tel.: (608) 770-7982. Email: amc521@psu.edu. Web: https://sites.psu.edu/aaroncook

1 Introduction

Nutrient pollution is considered one of the most important problems facing aquatic systems globally [2]. Degradation of freshwater, estuarine, and coastal aquatic ecosystems due to excessive levels of nitrogen and phosphorus is widespread in the United States, with headline-grabbing examples in the Chesapeake Bay, the Des Moines River, the Gulf of Mexico, and Lake Erie [8]. The existence of significant threats to freshwater and coastal ecosystems from nutrient pollution in the US and other developed nations, after decades of regulations and large investments to reduce water pollution, is in large degree the result of policy architectures that have been effective in reducing point source discharges but not nonpoint source of nutrients, particularly agricultural sources, which are often the major source of nutrient loads (see [11, 1, 3, 10, 20, 24]). Policy reforms and innovations are needed to improve the effectiveness of nutrient pollution controls, and ideally, to improve the overall efficiency of water pollution control allocations given the unnecessarily high cost of current approaches (see [18, 11]).

The expectation that water quality trading provides an effective mechanism for achieving water quality goals at low cost relative to alternative mechanisms has led to much interest in trading as a means for pollution management (see [4, 7, 14, 15, 16, 17, 18, 20, 22]). Several trading programs have been established, mostly in the US but also in Canada and New Zealand. An important feature of existing markets and standard guidance on market design is an assumption that the reduction in nonpoint pollution loads delivered to target water bodies occurs in the same market period as changes in agricultural land use and installation of agricultural Best Management Practices (BMPs) intended to reduce loads. This assumption is at odds with reality for several reasons. One is that BMPs are not instantly effective. An example is a riparian buffer, which becomes effective over a period of years as the vegetative cover matures [9]. Another is that pollutants that leave fields may take long periods of time to move from farm through ground water or stream channels to the locations at which nutrient problems occur. Lags in the response of water quality condition to BMPs has often been disregarded in water quality modeling and policy design, but the topic is receiving increasing attention as understanding lags and their significance increases (e.g., [23, 13]).

The assumption that agricultural nonpoint source reductions are instantaneous in a trading design implies sources with no lags or shorts lags (e.g., point sources of water pollution) can substitute credits from implementation of agricultural BMPs that may not occur for years for contemporaneous reductions in the short lagged discharges. This implies that water quality may be immediately degraded rather than protected by trades between sources with different lag lengths. The market "fix," in economic theory, is to allow for trading across potentially lengthy periods and across space [19]. But futures markets or trading commodities over long periods are extremely complex to implement, are expensive to operate, and do not necessarily perform well economically when the commodity is complex (as is the case with water quality) and/or there is significant uncertainty about economic conditions and regulatory environments in the future.

Consequently, it can be useful to rely on the simplicity of markets designed under the assumption of contemporaneous substitution (i.e., no lags) if the smaller cost associated with the simpler design

is significant and the delay in achieving the environmental targets is acceptable. Shortle et al. [19] compared a simple market equilibrium achieved under the assumption of instantaneous effects with a dynamically efficient market equilibrium assuming zero transactions costs. They found significant pollution control savings in the simple market, but also delays in the achievement of water quality goals. This paper examines a market design that would fall between the simplicity of assuming no lags and the complexity of a complete set of futures markets. The design entails the use of trade ratios used in contemporaneous trades that offset to some degree the inefficient shift of emissions from the present to the future.

We begin by developing a conceptual framework of pollution control over sources with varying lag lengths and compare the efficiency of 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, mimic the first-best optimum, though we would expect the complexity of such a market to suppress trade considerably. The trading ratio system would be simpler to implement (market participants trade contemporaneous discharges) and can mimic the first best in the steady state, though setting trade ratios optimally prior to the steady state is much more complicated. We 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.

2 Conceptual Framework

Suppose M point sources and N nonpoint sources contribute to the instantaneous pollution levels at a point in time t. Let $WL(t)$ and $L(t)$ denote point source wasteloads and nonpoint source pollution loads, respectively. For the purposes of this study, the difference between point and a nonpoint sources lie solely in the timing of their pollution delivery relative to discharge—discharges from point sources are delivered immediately, while those from each nonpoint source i are delivered after a source-specific delay, l_i.[1] Without loss of generality, let nonpoint sources be indexed such that the source with the shortest lag takes the value $i = 1$, the source with the next-shortest takes the value $i = 2$, and so on, with the longest-lagged source taking the value $i = N$.

Let $x_i(t)$ denote the quantity of pollution *discharged* from each nonpoint source pollution at time t and delivered at $t+\ell_i$. Let $w_j(t)$ denote the quantity of pollution discharged from each point source at time t and delivered instantaneously. Since pollution control measures in the nonpoint sector are incapable of affecting delivered pollution immediately, exogenous "legacy loads" associated with their past discharges will be delivered in early periods. Let $\bar{x}_i(t)$ denote these legacy loads delivered

[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.

from source i discharged at $t - \ell_i$ and delivered at t. Finally, let $b(t)$ denote natural background loads delivered at time t, which are included in the total nonpoint loads $L(t)$. The pollution delivery structures for $WL(t)$ and $L(t)$ can be described as follows:

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

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

Point source wasteloads and nonpoint source loads combine to produce the total loads delivered at time t, $TL(t)$. Formally,

$$TL(t) \;=\; WL(t) \;+\; L(t) \tag{3}$$

Let $c_i(x)$ represent nonpoint source i's cost associated with any nonpoint pollution discharge x, where $c_i' < 0$ (because costs increase as discharges fall) and $c_i'' > 0$ (because discharges become costlier to reduce at an increasing rate as discharges fall). Similarly, let $g_j(w)$ represent point source j's cost associated with any wasteload discharge w, where $g_j' < 0$ and $g_j'' > 0$ for the same reasons as in c_i. Finally, let $D(TL_t)$ represent the ecological damage costs associated with delivered pollution loads TL, where $D' > 0$ (because damages increase with total loads) and $D'' > 0$ (because damages increase with total loads at an increasing rate).

2.1 The First-Best Optimum

Let Z represent the total costs of pollution and pollution cleanup over time in present value terms. Formally,

$$Z\big[w_j(t), x_i(t), TL(t)\big] \;=\; \int_0^T \Big\{ \sum_j g_j\big[w_j(t)\big] \;+\; D\big[TL(t)\big] \Big\} e^{-\delta t}\, dt \;+\; \sum_i \int_{\ell_i}^T c_i\big[x_i(t - \ell_i)\big]\, e^{-\delta(t-\ell_i)}\, dt \tag{4}$$

where δ denotes the discount factor. We model the problem over the continuous (but finite) time horizon $t \in [0, T]$. Due to lag times, nonpoint source i's discharges become irrelevant after $T - \ell_i$, hence the boundaries of integration for nonpoint source costs $c_i\big[x_i(t-\ell_i)\big]$ (the costs corresponding to delivered loads) in period t run from ℓ_i to T. The regulator's problem can be expressed

$$\min_{w_j(t),\, x_i(t)} Z\big[w_j(t), x_i(t), TL(t)\big] \quad \text{subject to (1)–(3)}$$

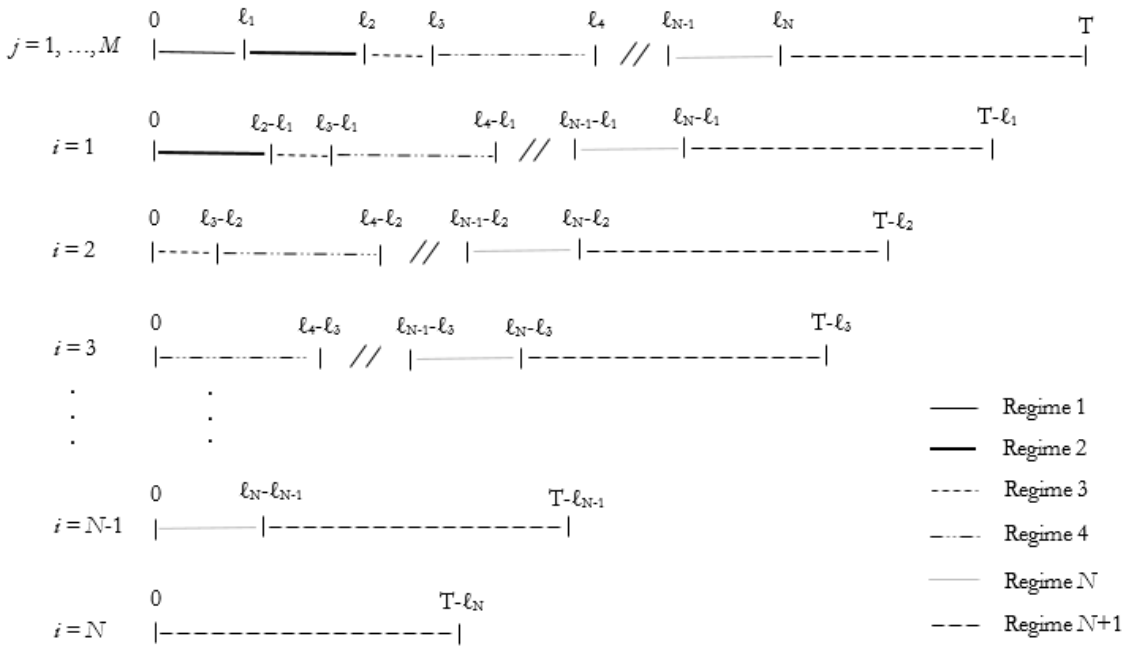

Figure 1: Regime timing for sources of different lag lengths

Since the pollutant in this analysis does not persist from one period to the next, minimizing Z merely requires minimizing the integrand at each instant. The set of cost-minimizing $w_j(t)$ and $x_i(t)$ must satisfy

$$-g_j'\big[w_j(t)\big] \;=\; D'\big[TL(t)\big] \qquad \forall j,\ t \in [0,T] \tag{5a}$$

$$-c_i'\big[x_i(t-\ell_i)\big]e^{\delta \ell_i} \;=\; D'\big[TL(t)\big] \qquad \forall i,\ t \in [\ell_i,T] \tag{5b}$$

along with (2), (1), and (3). Within these $M+N$ conditions is a set of $N+1$ unique pollution control "regimes" where particular sets of sources are optimized jointly over different time intervals. Because nonpoint sources are incapable of affecting delivered loads prior to $t=\ell_1$, the first regime consists of the M point sources allocating $w_j(t)$ optimally among themselves during $t \in [0,\ell_1]$. From the vantage point of $t=\ell_1$, the nonpoint discharges $x_1(0)$ become relevant for the solution because they are scheduled for delivery at $t=\ell_1$. The second regime therefore must involve both $w_j(t)$ during $t \in [\ell_1,\ell_2]$ together with $x_1(t)$ during $t \in [0,\ell_2-\ell_1]$. In the same way, discharges $x_2(0)$ become relevant by $t=\ell_2$, and so the third regime involves optimizing $w_j(t)$ during $t \in [\ell_2,\ell_3]$ together with $x_1(t)$ during $t \in [\ell_2-\ell_1,\ell_3-\ell_1]$ and $x_2(t)$ 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. In general, the k^{th} regime involves all M point sources together with the nonpoint sources $i \in \{1,\ldots,k-1\}$. Figure 1 illustrates how the timing of the various regimes are staggered for sources with different lag lengths.

Together with (1)-(3), optimal discharges in the first regime must satisfy

$$-g_j'\big[w_j(t)\big] \;=\; D'\big[TL(t)\big] \qquad \forall j,\ t \in [0, \ell - 1] \tag{6}$$

while for $k \in \{2, \ldots, N\}$, optimal discharges in the k^{th} regime must satisfy

$$-g_j'\big[w_j(t)\big] \;=\; D'\big[TL(t)\big] \qquad \forall j,\ t \in [\ell_k, \ell_{k+1}) \tag{7a}$$

$$-c_i'\big[x_i(t-\ell_i)\big]e^{\delta \ell_i} \;=\; D'\big[TL(t)\big] \qquad \text{for } i \in \{1, \ldots, k\},\ t \in [\ell_k, \ell_{k+1}) \tag{7b}$$

The final regime, $N+1$, will account for deliveries from the longest-lagged source and so its solution will optimize all pollution sources jointly. Optimal discharges in the final regime must satisfy

$$-g_j'\big[w_j(t)\big] \;=\; D'\big[TL(t)\big] \qquad \forall j,\ t \in [\ell_N, T] \tag{8a}$$

$$-c_i'\big[x_i(t-\ell_i)\big]e^{\delta \ell_i} \;=\; D'\big[TL(t)\big] \qquad \forall i,\ t \in [\ell_N, T] \tag{8b}$$

The costs of reducing point source discharges during $t \in [0, \ell_1]$ are balanced against the contemporaneous damages they prevent according to condition (6). For $t \geq \ell_1$ point source discharges are chosen jointly with nonpoint sources according to conditions (8a) and (8b), where reduction costs in each sector are balanced against pollution damage costs being prevented, adjusted for when pollution control costs occur relative to deliveries. These conditions imply that between any two firms i and j,

$$-g_j'\big[w_j(t)\big] \;=\; -c_i'(t)\big[x_i(t-\ell_i)\big]e^{\delta \ell_i} \tag{9}$$

Condition (9) states that the optimal allocation will equalize the present value marginal abatement costs for any point source j and nonpoint source i. An alternative way to think about the optimal allocation is in terms of ratios of marginal abatement costs and marginal damage costs. At the optimum,

$$\frac{g_j'\big[w_j(t)\big]}{c_i'\big[x_i(t)\big]} \;=\; \frac{D'\big[TL(t)\big]}{D'\big[TL(t+\ell_i)\big]e^{-\delta \ell_i}}$$

which says that the ratio of contemporaneous marginal costs between point source j and nonpoint source i must equal the present value of the marginal damages corresponding to each firm's emissions. In a first-best setting, pollution must be allocated between sources according to both abatement cost and damage cost criteria. This expression hints at the rationale for the trade ratio framework developed later in the paper where the trade ratios are intended to correct for the imperfect substitution between point and nonpoint loads in terms of damage costs they imply.

Finally, we consider the time structure of optimal delivered loads in this lagged pollution setting. Let $w_j^*(t)$ and $x_i^*(t)$ represent the point and nonpoint source discharges that solve (1)-(3) and (6)-(8). Point source discharges $w_j^*(t)$ will span the full time period $t \in [0, T]$ while $x_i^*(t)$ will only be relevant during the period $t \in [0, T - \ell_i]$. Optimal deliveries, $TL^*(t)$, in each period are by

definition

$$TL^*(t) \;=\; \begin{cases} \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=1}^{N} \bar{x}_i(t) & \text{for } t \in [0, \ell_1) \\ \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=k}^{N} \bar{x}_i(t) \;+\; \sum_{i=1}^{k-1} x_i^*(t-\ell_i) & \text{for } k \in \{2,\ldots,N\}, t \in [\ell_{k-1}, \ell_k) \\ \displaystyle\sum_{j=1}^{M} w_j^*(t) \;+\; b(t) \;+\; \sum_{i=1}^{N} x_i^*(t-\ell_i) & \text{for } t \in [\ell_N, T] \end{cases} \tag{10}$$

2.2 Markets for Pollution Deliveries (A Forward Market Approach)

One way of achieving the optimal delivered loads given by (10) 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. Permits 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 trading 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 trading eliminates gains from trade, the permit market equilibrium under this market design is found by solving

$$\min_{x_i(t),\, w_j(t)} \;\; \int_0^T \sum_j g_j\big[w_j(t)\big]\, e^{-\delta t}\, dt \;+\; \sum_{i=1}^{N} \int_{\ell_i}^{T} c_i\big[x_i(t-\ell_i)\big]\, e^{-\delta(t-\ell_i)}\, dt$$

subject to

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

where the constraints of this cost-minimization problem restrict actual deliveries in each t to be less than or equal to the levels implied in (10). The Lagrangian for this problem is

$$\begin{aligned}
\mathcal{L} \;=\; & \int_0^T \sum_j g_j\big[w_j(t)\big]\, e^{-\delta t}\, dt \;\;+\;\; \sum_i \int_{\ell_i}^T c_i\big[x_i(t-\ell_i)\big]\, e^{-\delta(t-\ell_i)}\, dt \\
& + \int_0^{\ell_i} \lambda(t)\Big[\sum_j w_j(t) \,+\, b(t) \,+\, \sum_{i=1}^{N} \bar{x}_i(t) \,-\, TL^*(t)\Big]\, dt \\
+ \sum_{k=2}^{N} \int_{\ell_{k-1}}^{\ell_k} & \lambda(t)\Big[\sum_j w_j(t) \,+\, b(t) \,+\, \sum_{i=k}^{N} \bar{x}_i(t) \,+\, \sum_{i=1}^{k-1} x_i(t-\ell_i) \,-\, TL^*(t)\Big]\, dt \\
& + \int_{\ell_N}^T \lambda(t)\Big[\sum_j w_j(t) \,+\, b(t) \,+\, \sum_{i=1}^{N} x_i(t-\ell_i) \,-\, TL^*(t)\Big]\, dt
\end{aligned}$$

with the optimal discharges satisfying the first order conditions

$$-g_j'\big[w_j(t)\big]e^{-\delta t} \;=\; \lambda(t) \qquad \forall j,\; t \in [0,T] \tag{11a}$$

$$-c_i'\big[x_i(t-\ell_i)\big]e^{-\delta(t-\ell_i)} \;=\; \lambda(t) \qquad \forall i,\; t \in [\ell_i,T] \tag{11b}$$

Provided the time-specific caps on aggregate delivered loads match those in (10), the discharge levels that satisfy equilibrium under this market structure will match those in (5). Note that the allocation of discharges across point and nonpoint sources will be efficient under this market design since (11) implies

$$-g_j'\big[w_j(t)\big] \;=\; -c_i'\big[x_i(t-\ell_i)\big]e^{\delta\ell_i}$$

which is identical to (9).

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 for large N. 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 today estimated to deliver x pounds of pollution reduction at some future date t (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 this later date.

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. [12]) 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). We discuss this system next.

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 we allow these trades 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, $\hat{TD}_t$, be based on the optimal discharges from section 2.1, where

$$\hat{TD}(t) \;=\; \sum_j w_j^*(t) \;+\; \sum_i a_i^*(t)$$

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

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

subject to

$$\sum_j w_j(t) \;+\; \sum_i x_i(t) \;\le\; \hat{TD}(t) \quad \forall t \tag{12}$$

with corresponding Lagrangian expression

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

Optimal discharges satisfy

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

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

implying that $-g_j'\big[w_j(t)\big] = -c_i'\big[x_i(t)\big]$. This outcome differs from (9) in two ways. First, marginal costs between the two sectors are being evaluated on contemporaneous discharges, whereas in (9) marginal point source reduction costs at t are being compared with marginal nonpoint source reduction costs at $t-\ell_i$. Second, nonpoint marginal reduction costs in (9) are inflated by the continuous time discount factor $e^{\delta\ell_i}$, whereas 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 each nonpoint source's discharges equal to $e^{\delta\ell_i}$, meaning that for every pound of pollution increased at a point source, its nonpoint trading partner would have to reduce its discharges $e^{\delta\ell_i}$ pounds. Formally, this design choice would require modifying the cap on discharges, turning (12) into

$$\sum_j w_j(t) \;+\; \sum_i x_i(t)e^{-\delta\ell_i} \;\le\; T\hat{D}(t) \quad \forall t$$

The first order conditions for the new Lagrangian expression are

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

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

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

$$-g_j'\big[w_j(t)\big] \;=\; -c_i'\big[x_i(t)\big]e^{\delta\ell_i} \tag{15}$$

This will match (9) provided that $x_i(t) = x_i(t-\ell_i)$ for all i, which will be true for $t\in[\ell_N,T]$. The time frame during which this approach would be valid corresponds to the "final regime" during which loads reach steady state levels. However, because (15) does not, in general, match (9) during the period $t\in[0,\ell_N]$, allowing sources with lag lengths that differ by ℓ trade with one another at the rate $e^{\delta\ell}$-to-1 is not guaranteed to mimic the first-best solution. Regime-specific trade ratios would have to be applied prior to $t=\ell_N$. Given the enormous complexity of the actual nonpoint pollution delivery process, these would be enormously difficult to compute. In the next section, we provide a framework for how trade ratios to address lags would be designed in theory.

3 Two Polluter, Two Period Problem

To gain intuition about applying trade ratios to load reallocations between sources with different lag lengths, consider a two polluter model where one source's discharges are delayed by ℓ periods.

Let w_t and x_t denote loads discharged in period t from wastewater and agriculture, respectively. As before, let the costs associated with these discharge levels be given by $g(w_t)$ and $c(x_t)$ and let L_t and $L_{t+\ell}$ denote the loads delivered in period t and $t+\ell$, which are defined as follows:

$$L_t \;=\; x_{t-\ell} \;+\; w_t \tag{16a}$$

$$L_{t+\ell} \;=\; x_t \;+\; w_{t+\ell} \tag{16b}$$

where $x_{t-\ell}$ represents exogenous "legacy loads" from agricultural discharges in period $t-\ell$ and $w_{t+\ell}$ represents exogenous wasteloads discharged in period $t+\ell$. The regulator's problem is

$$\min_{w_t,x_t} \; g(w_t)\delta^t \;+\; c(x_t)\delta^t \;+\; D(L_t)\delta^t \;+\; D(L_{t+\ell})\delta^{t+\ell} \qquad \text{subject to (16)}$$

The optimal w_t and x_t must satisfy $-g'(w_t) = D'(x_{t-l}+w_t)$ and $-c'(x_t) = D'(x_t+w_{t+\ell})\delta^\ell$, 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_t^* and x_t^*. Note that x_t^* depends on lag length ℓ, whereas w_t^* is independent of ℓ. The trade ratio and discharge cap that replicates this first-best solution under a particular lag length must induce the lag-specific optimal nonpoint source load while maintaining the optimal point source load that would prevail under any lag length.

Consider a regulatory mechanism that establishes a trade ratio (to dictate 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

$$\min_{w_t,x_t} \; g(w_t) \;+\; c(x_t) \quad \text{subject to} \quad x_t \;+\; w_t\psi \;=\; \hat{L}_t \tag{17}$$

where ψ represents the trade ratio (denominated in pounds of agricultural loads per pound of wasteloads) and $\hat{L}_t$ represents the discharge cap (denominated in agricultural loads). The Lagrangian expression that corresponds to (17) is

$$\mathcal{L} \;=\; g(w_t) \;+\; c(x_t) \;+\; \lambda\Big[x_t \;+\; w_t\psi \;-\; \hat{L}_t\Big]$$

Sources will reallocate discharges between them until

$$-g'(w_t) = -c'(x_t)\psi$$

Let $\tilde{w}_t(\psi,\hat{L})$ and $\tilde{x}_t(\psi,\hat{L})$ denote each polluter's equilibrium discharges for any ψ and $\hat{L}$. Using this trading equilibrium condition $g'(w_t)-c'(x_t)\psi = 0$, the credit balancing condition $x_t+w_t\psi-\hat{L}_t = 0$, and the implicit function theorem (Mas Colell, Whinston, and Green, 1995) the changes in the

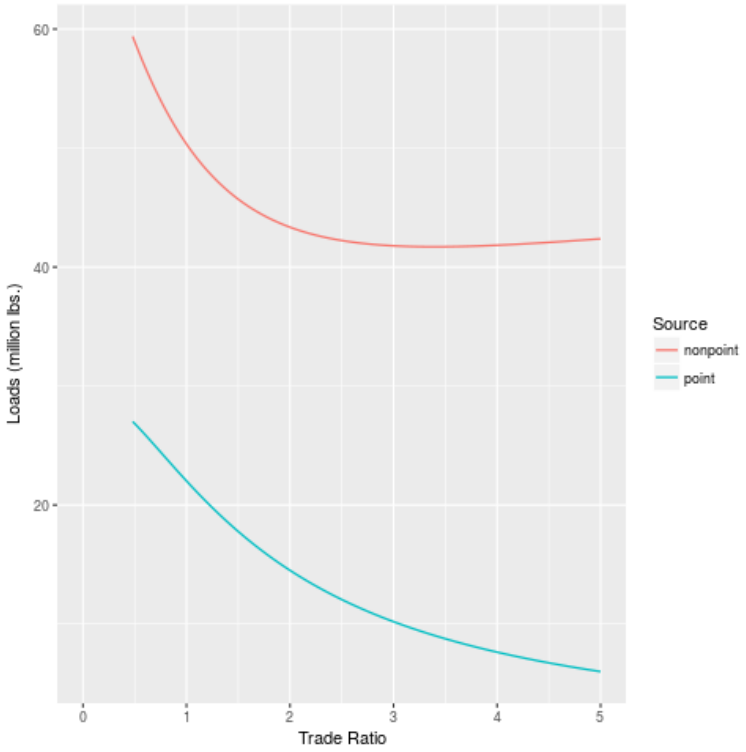

Figure 2: Loads under the optimal cap for various trade ratios

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 $\hat{L}$ are

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

The derivatives of equilibrium loads with respect to $\hat{L}$ are unambiguously positive and the derivative of wasteloads with respect to ψ is unambiguously negative. Howerver, since $-c'(\tilde{x})\, \psi$ and $\tilde{w}\, g''(\tilde{w})$ are both positive, the sign of $\frac{\partial \tilde{x}}{\partial \psi}$ is ambiguous. To illustrate these relationships more clearly Figure 2 provides a graphical example. In this particular case, 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 $\hat{L}$. Recall the constraint in problem (17) (the discharge cap) and note how the point source sector's usage of the cap is given by the product of w_t and ψ. For large values of ψ, point source loads may shrink such that the overall size of $w_t \psi$ may decrease, leaving a larger share of cap left for nonpoint loads.

3.1 The First Best Regulation

Given the equilibrium outcome of a discharge trading system under any choice of trading ratio ψ and endowment $\hat{L}$, consider next how to choose the optimal ψ and $\hat{L}$. Formally, a regulator would choose these parameters to

$$\min_{\psi,\,\hat{L}} \; g\big[\tilde{w}(\psi,\hat{L})\big] \;+\; c\big[\tilde{x}(\psi,\hat{L})\big] \;+\; D\Big[x_0+\tilde{w}(\psi,\hat{L})\Big] \;+\; D\Big[\tilde{x}(\psi,\hat{L})+w_2\Big]\delta$$

The optimal values of ψ and $\hat{L}$ must jointly satisfy

$$\left\{g'\big[\tilde{w}(\psi,\hat{L})\big] \;+\; D'\Big[x_{t-\ell}+\tilde{w}(\psi,\hat{L})\Big]\right\}\frac{\partial\tilde{w}}{\partial\psi} \;+\; \left\{c'\big[\tilde{x}(\psi,\hat{L})\big] \;+\; D'\Big[\tilde{x}(\psi,\hat{L})+w_{t+\ell}\Big]\delta^{\ell}\right\}\frac{\partial\bar{x}}{\partial\psi} \;=\; 0$$

$$\left\{g'\big[\tilde{w}(\psi,\hat{L})\big] \;+\; D'\Big[x_{t-\ell}+\tilde{w}(\psi,\hat{L})\Big]\right\}\frac{\partial\tilde{w}}{\partial\hat{L}} \;+\; \left\{c'\big[\tilde{x}(\psi,\hat{L})\big] \;+\; D'\Big[\tilde{x}(\psi,\hat{L})+w_{t+\ell}\Big]\delta^{\ell}\right\}\frac{\partial\tilde{x}}{\partial\hat{L}} \;=\; 0$$

Since $\frac{\partial\tilde{w}}{\partial\psi}$ will not equal $\frac{\partial\tilde{w}}{\partial\hat{L}}$ in general and $\frac{\partial\tilde{x}}{\partial\psi}$ will not equal $\frac{\partial\tilde{x}}{\partial\hat{L}}$ in general, the terms in brackets must equal zero to guarantee these conditions are satisfied. Recall that $-g'(w_t) = D'(x_{t-\ell}+w_t)$ and $-c'(x_t) = D'(x_t+w_{t+\ell})\delta^{\ell}$ corresponds to the first best solution.

With two policy levers at their disposal, a regulator can, in theory, adjust both the cap and the trade ratio to reproduce the first-best loads w_t^* and x_t^*. We illustrate this with a numerical example (see Figure 3), taking abatement costs and damage costs as given and solving for the optimal cap-trade ratio pairs for lag lengths ranging from 1 to 30 years. From the regulator's perspective, to accommodate the increase in optimal nonpoint loads as lag lengths get longer, they must first increase the cap, making room for more nonpoint loads, and then increase the trade ratio to shift point source loads back to their previous level. The result is higher nonpoint loads, but constant point loads at each new lag length.

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. Along these lines, consider the case in which the discharge cap $\hat{L}$ is given exogenously and the regulator seeks to minimize social costs solely through its choice of ψ. The optimal ψ must satisfy

$$\left\{g'\big[\tilde{w}(\psi,\hat{L})\big] \;+\; D'\Big[x_{t-\ell}+\tilde{w}(\psi,\hat{L})\Big]\right\}\frac{\partial\tilde{w}}{\partial\psi} \;+\; \left\{c'\big[\tilde{x}(\psi,\hat{L})\big] \;+\; D'\Big[\tilde{x}(\psi,\hat{L})+w_{t+\ell}\Big]\delta^{\ell}\right\}\frac{\partial\bar{x}}{\partial\psi} \;=\; 0$$

Here, the first-best outcome can be achieved only if the cap was initially set at the optimal level. In the event that the cap is sub-optimal, the terms in brackets will not be equal to one another and the rate of substitution between point and nonpoint sources at the (second-best) optimal ψ will be given by

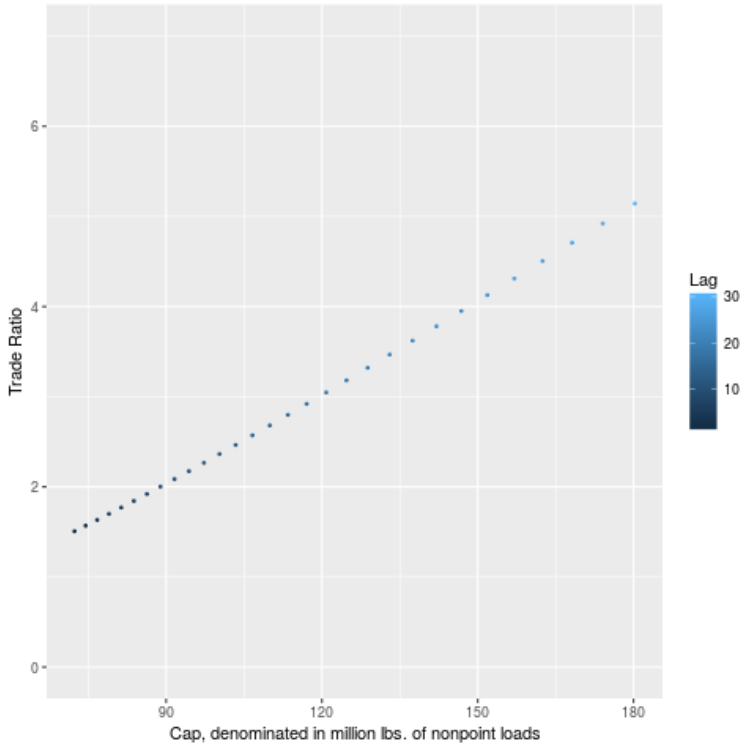

Figure 3: Optimal pair of trade ratio and cap for lag lengths of 1 to 30 years

$$\frac{\partial \tilde{x}}{\partial \tilde{w}} = -\frac{g'(\tilde{w}) + D'\left[x_{t-\ell} + \tilde{w}\right]}{c'(\tilde{x}) + D'\left[\tilde{x} + w_{t+\ell}\right]\delta^{\ell}} \tag{18}$$

Expression (18) says that the optimal rate at which nonpoint loads should substitute for point loads is equal to the ratio of the net marginal costs[2] of point source emissions to the net marginal costs of nonpoint source emissions. Rather than directly equating marginal abatement costs and marginal damage costs for each type of load separately 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 point source reductions, thereby delivering more immediate ecological benefits).

Figure 4 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

[2]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 (18) is therefore the damage costs associated with each type of load *net of the cost savings* associated with discharging loads of that size

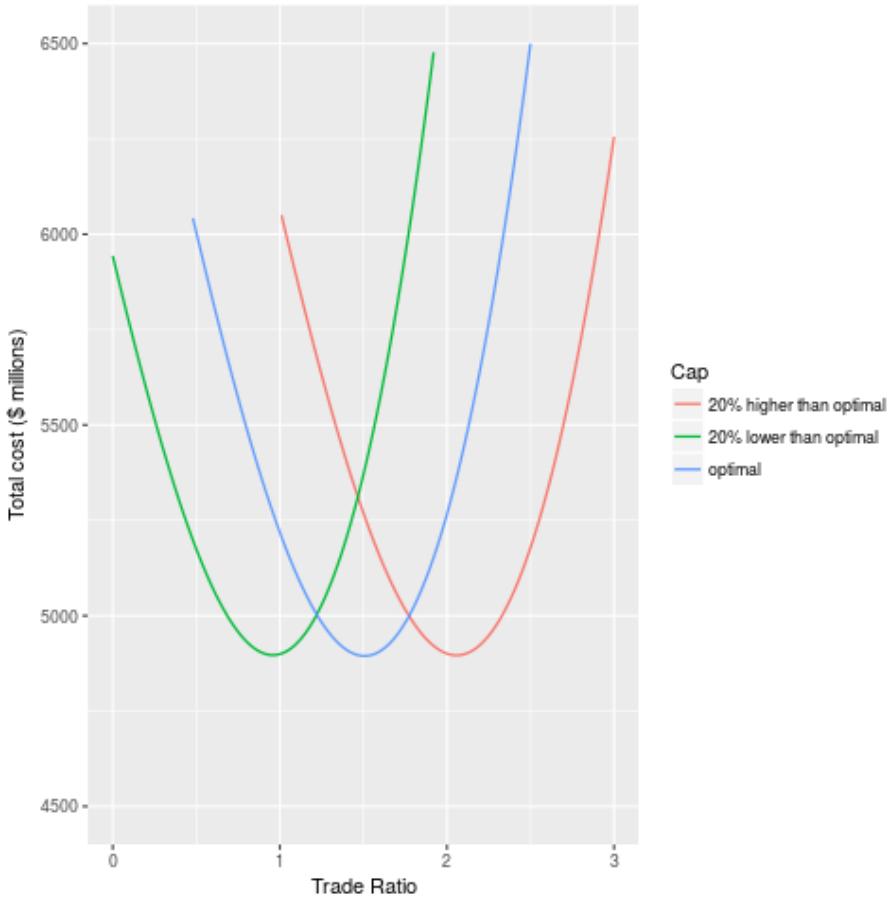

Figure 4: Total costs (abatement plus damage costs) for various trade ratios

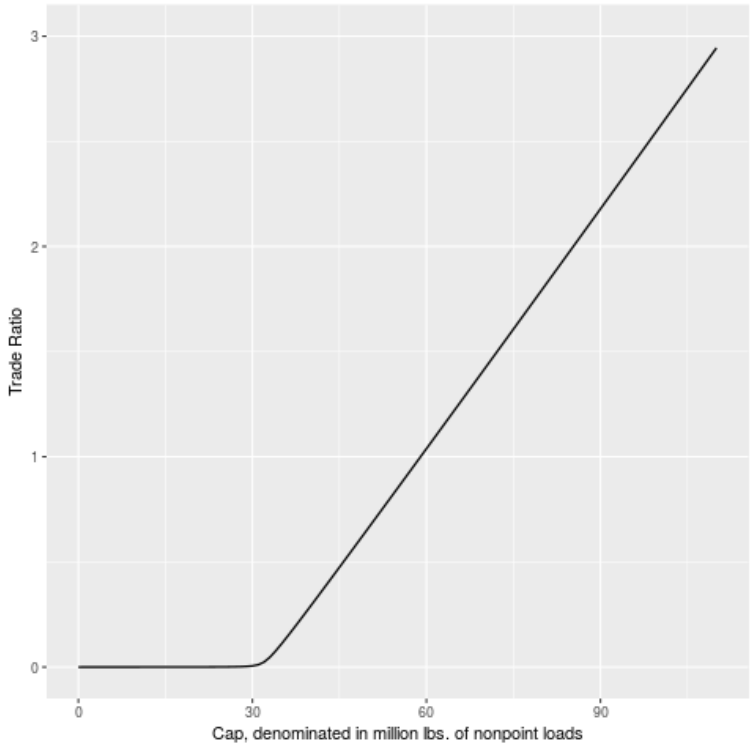

Figure 5: Optimal trade ratio given the cap (for 1-year lag length)

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 4, 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 pollution context can be seen from equation (18) where both abatement costs and damage costs factor into the solution. 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 5 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 [21] and Horan and Shortle [6] 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 [5], 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.

4 Conclusion

We develop a theory of lagged pollution control, noting first 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 1). The number of regimes corresponds to the number of distinct lag lengths that exist among the polluters under regulation. Second, we observe that 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 [4], we could expect this design to suppress trading activity still further.

Next, we examined an alternative market design 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 provided the trading occurs during $t \in [\ell_N, T]$ (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. However, this trading rule will not generally reproduce the first-best loads for $t < \ell_N$.

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, we 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). 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 [5]; Horan and Shortle [6]), regulators must account for the size of the cap when designing trade ratios that account for lag length.

References

[1] Jose Albiac. Nutrient imbalances: pollution remains. *Science*, 326(5953):665–665, 2009.

[2] Millennium Ecosystem Assessment. Ecosystems and human well-being: synthesis. millennium ecosystem assessment series. *World Resources Institute, Washington, DC*, page 155, 2005.

[3] Stephen R Carpenter, Nina F Caraco, David L Correll, Robert W Howarth, Andrew N Sharpley, and Val H Smith. Nonpoint pollution of surface waters with phosphorus and nitrogen. *Ecological Applications*, 8(3):559–568, 1998.

[4] Karen Fisher-Vanden and Sheila Olmstead. Moving pollution trading from air to water: potential, problems, and prognosis. *Journal of Economic Perspectives*, 27(1):147–72, 2013.

[5] 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.

[6] 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.

[7] Antti Iho, Marc Ribaudo, and Kari Hyytiäinen. Water protection in the baltic sea and the chesapeake bay: Institutions, policies and efficiency. *Marine pollution bulletin*, 93(1-2):81–93, 2015.

[8] Madhu Khanna and James S Shortle. (theme overview) preserving water quality: Challenges and opportunities for technological and policy innovations. *Choices*, 32(4):1–4, 2017.

[9] 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.

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

[11] OECD. *Water Quality and Agriculture*. 2012.

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

[13] Lucy O'Shea and Andrew Wade. Controlling nitrate pollution: An integrated approach. *Land Use Policy*, 26(3):799–808, 2009.

[14] Marc O Ribaudo and Jessica Gottlieb. Point-nonpoint trading–can it work? *JAWRA Journal of the American Water Resources Association*, 47(1):5–14, 2011.

[15] Kathleen Segerson and Dan Walker. Nutrient pollution: an economic perspective. *Estuaries*, 25(4):797–808, 2002.

[16] Mindy Selman, Evan Branosky, and Cy Jones. Water quality trading programs: An international overview. *WRI Issue Brief Water Quality Trading*, 1, 2009.

[17] James Shortle. Economics and environmental markets: Lessons from water-quality trading. *Agricultural and Resource Economics Review*, 42(1):57–74, 2013.

[18] James Shortle. Policy nook: "economic incentives for water quality protection". *Water Economics and Policy*, 3(02):1771004, 2017.

[19] James Shortle, David Abler, Zach Kaufman, and Katherine Y Zipp. Simple vs. complex: Implications of lags in pollution delivery for efficient load allocation and design of water-quality trading programs. *Agricultural and Resource Economics Review*, 45(2):367–393, 2016.

[20] James Shortle and Richard D Horan. Policy instruments for water quality protection. *Annu. Rev. Resour. Econ.*, 5(1):111–138, 2013.

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

[22] Hale W Thurston, Haynes C Goddard, David Szlag, and Beth Lemberg. Controlling stormwater runoff with tradable allowances for impervious surfaces. *Journal of Water Resources Planning and Management*, 129(5):409–418, 2003.

[23] Kimberly J Van Meter and Nandita B Basu. Catchment legacies and time lags: a parsimonious watershed model to predict the effects of legacy storage on nitrogen export. *PloS one*, 10(5):e0125971, 2015.

[24] Peter M Vitousek, Rosamond Naylor, Timothy Crews, MB David, LE Drinkwater, E Holland, PJ Johnes, J Katzenberger, LA Martinelli, PA Matson, et al. Nutrient imbalances in agricultural development. *Science*, 324(5934):1519–1520, 2009.