When Two Wrongs Make a Right: Second-Best Point–Nonpoint Trading Ratios

Richard D. Horan and James S. Shortle

Most research on point–nonpoint trading focuses on the choice of trading ratio (the rate point source controls trade for nonpoint controls), although the first-best ratio is jointly determined with the optimal number of permits. In practice, program managers often do not have control over the number of permits—only the trading ratio. The trading ratio in this case can only be second-best. We derive the second-best trading ratio and, using a numerical example of trading in the Susquehanna River Basin, we find the values are in line with current ratios, but for different reasons than those that are normally provided.

Key words: agricultural pollution, environmental policy, permit trading, water quality.

Among the most important EPA initiatives to address agricultural and other nonpoint source contributions to water quality problems is the Total Maximum Daily Load (TMDL) program. The program requires states to develop and implement watershed-based plans for water resources that are too polluted to meet designated uses. In many watersheds, achieving designated uses will require that states tackle long unregulated nonpoint sources. As the leading nonpoint source, agriculture will likely be a major target of TMDL initiatives (USDA and USEPA).

There is substantial interest in using point–nonpoint trading to achieve nonpoint source reductions (GLTN, Faeth, U.S. EPA). Several fully implemented and pilot point–nonpoint trading programs have emerged over the past decade, the best-known being Tar-Pamlico (NC), Cherry Creek (CO), Dillon Reservoir (CO), and Fox River Basin (WI) (Horan). In January 2003, the EPA announced rules for trading programs, with funding for eleven pilot programs, including one for the Chesapeake Bay region (U.S. EPA).

Richard D. Horan is associate professor, Department of Agricultural Economics, Michigan State University. James S. Shortle is distinguished professor of Agricultural and Environmental Economics, Department of Agricultural Economics and Rural Sociology, The Pennsylvania State University.

The authors appreciate the helpful comments of Stephen Swallow and three anonymous reviewers. The usual disclaimer applies. Horan acknowledges the support of the Michigan Agricultural Experiment Station. Shortle acknowledges the support of the USEPA National Center for Environmental Research STAR Program Grant No. EPA/R-8286841.

Point–nonpoint trading works as follows: Pollution sources are required to hold permits that define their allowable discharges. For metered point sources, the permits define allowable measured discharges. Because nonpoint discharges are generally unobservable, the permits define allowable "estimated" discharges, where the estimates are derived from models linking observable land use and management practices to nonpoint loads. With tradeable permits, each source can adjust its allowances by buying or selling permits subject to rules governing trades. Among these rules is a trading ratio that defines how many nonpoint source permits trade for one point source permit.

The current interest in point–nonpoint trading originates in large part from the expectation that trading will achieve water quality improvements at lower cost than would be possible with traditional regulatory approaches (GLTN, Faeth, U.S. EPA). However, as with any trading program, the magnitude of the efficiency gains (EGs) from trading will clearly depend on how well the trading program is designed and implemented, i.e., the choice of trading ratios, permit allocations, etc. (e.g., Stavins). Theoretical research on the design of point–nonpoint trading schemes has focused on the choice of the trading ratio (Shortle, Horan, Woodward), and there has been substantial attention to the choice of this parameter in the design of actual programs (GLTN).

For economic efficiency, the trading ratio should reflect the relative expected marginal environmental (damage) impacts from each

Amer. J. Agr. Econ. 87(2) (May 2005): 340–352

source, the relative uncertainty (risk) created by each source, and the relative marginal transactions costs associated with a trade. Risk is often the major focus of debate. Research on optimal point–nonpoint trading indicates that the trading ratio should be set to encourage more control from the source whose emissions generate the most risk—a smaller ratio (possibly less than unity) is optimal if, at the margin, nonpoint controls result in greater risk reduction than do point source controls, and a larger ratio is optimal when the opposite condition holds. In practice, trading ratios are all greater than one (see table 1 in Horan) with the standard argument for these large ratios being that it is less risky to encourage more control from point sources, as nonpoint controls may have uncertain effectiveness. This argument only accounts for part of the risk; however, as failure to provide sufficient nonpoint controls also results in risk due to the inherent randomness of nonpoint pollution loads (e.g., due to weather) (Horan), and the optimal response to this risk is a reduction in the trading ratio, under reasonable assumptions (Shortle). For instance, risk due to stochastic weather processes may result in optimal trading ratios for the Susquehanna River Basin (SRB) being less than one (Horan, Shortle, and Abler 2002; Horan et al.).

So, does prior research imply that actual trading ratios are too high? The answer is "perhaps not." The economics research on trading ratios has largely ignored a key policy tool parameter in a trading program—the number of permits. In theory, permit numbers and the trading ratio must be chosen simultaneously to achieve economic efficiency. But in practice, permit numbers have been exogenous to the trading program authority. Agriculture's participation is voluntary in existing programs, with the sector essentially having a presumptive right to pollute at historical levels.[1] All enforceable regulations are placed on point sources, with administration usually at the national level (e.g., through National Pollution Discharge Elimination System [NPDES] permits).

We examine the trading authority's (e.g., a state agency developing TMDL strategies) choice of the trading ratio when allowable emissions are allocated at suboptimal levels by a superseding authority (e.g., U.S. EPA). The trading ratio in this case can only be second-best, as it becomes the only available instrument to address multiple distortions (Tinbergen). We find the second-best trading ratio depends on the initial permit allocation, and it is likely to be larger than the first-best or optimal value. This result is investigated with a numerical analysis of point–nonpoint nutrient trading in the Pennsylvania portion of the SRB, and we find that optimal trading ratios lie in the range of those often applied. This provides support for the use of current ratios, but for different reasons than those that are normally provided. The results also indicate that gains could arise if states were to break with tradition and impose enforceable requirements on agricultural sources through the TMDL process and link these choices with that of the trading ratio.

A Model of Point–Nonpoint Trading

Consider a very simple model involving a single point source (a firm) and a single nonpoint source (a farm). The firm's emissions are denoted by e, and the firm can control these emissions with certainty at a cost of $c(e)$ (with $c'(e) < 0$). Farm emissions or loadings are given by $r(x, \theta)$, where x represents input use (e.g., production and pollution control choices) and θ is a random variable influenced by weather and other stochastic environmental drivers. The farmer cannot control loadings with certainty, but rather influences the distribution of loadings through input choices. For simplicity, we take x to be a scalar although in principle it would be a vector (Horan, Shortle, and Abler 2002). The farmer's profit from this input choice is $\pi(x)$. In the unregulated equilibrium, the farmer sets input use at the level x^0, earning profit $\pi(x^0)$, and producing pollution loads $r(x^0, \theta)$. The farmer's pollution control costs, g, are simply the reduction in profits, or $g(x) = \pi(x^0) - \pi(x)$. Pollution from the two sources causes economic damages, denoted by $D(e, r)$. We assume that society is risk-neutral so that the regulatory authority seeks to minimize the expected social

[1] For trades in a watershed where a TMDL is not yet established (pre-TMDL trading), "nonpoint source baselines are the level of pollutant load associated with existing land uses and management practices," and point source baselines are defined by their NPDES permit or other applicable effluent limitation (U.S. EPA). For watersheds with a TMDL in place, trading "should be consistent with the assumptions and requirements upon which the TMDL is established" (U.S. EPA). States have flexibility in implementing TMDLs and in defining nonpoint source accountability, as EPA has no Clean Water Act oversight authority for nonpoint sources. Our analysis is based on an aggregate emissions cap, with initial enforceable requirements placed on point sources. This is consistent with established programs (GLTN 2000) and with the example that EPA Administrator Whitman used in her speech introducing the national trading program (Whitman 2003).

costs (TC) from pollution and its control, i.e., $TC = c(e) + g(x) + E\{D(e, r)\}$.[2]

The Market Equilibrium

Conventionally, such as in markets for SO_2 permits, pollution permits define allowable emissions for the permit holder. However, nonpoint loads cannot be directly traded because they cannot be routinely and accurately metered at reasonable cost and they have a significant random component (Shortle). Accordingly, an alternative basis for nonpoint trades is required. The option we consider entails trading changes in point source emissions for changes in estimated nonpoint loadings. In this case, data on agricultural land uses, and geophysical and climatic factors are input into models (e.g., SWAT or AGNPS) that estimate nonpoint loads. Existing point–nonpoint trading programs are of this type (Hoag and Hughes-Popp; Shortle and Abler).

The trading program works as follows. Two categories of permits are required: point source permits, $\hat{e}$, and nonpoint source permits, $\hat{r}$. The former are denominated in terms of emissions while the latter are denominated in terms of expected loadings. Firms must have a combination of both types at least equal to their emissions, in the case of point sources, or expected loadings in the case of nonpoint sources. The cross-category trading ratio is denoted by t, i.e., $t = |d\hat{r}/d\hat{e}|$.

Denote the market price of expected loadings permits as p and the price of emissions permits as q. The point source will choose emissions levels, point source permit holdings, $\hat{e}_{ps}$, and nonpoint source permit holdings, $\hat{r}_{ps}$, to minimize costs, $C = c(e) + q[\hat{e}_{ps} - \hat{e}^0_{ps}] + p[\hat{r}_{ps} - \hat{r}^0_{ps}]$, given that its total emissions cannot be greater than its permit holdings, $e \leq \hat{e}_{ps} + (1/t)\hat{r}_{ps}$, where $\hat{e}^0_{ps}$ and $\hat{r}^0_{ps}$ are initial point and nonpoint source permits held by the firm, respectively. The term $(1/t)\hat{r}_{ps}$ represents the emissions the firm can generate based on its expected loadings permits. Assuming as in existing trading programs that firms do not initially hold nonpoint source permits (i.e., $\hat{r}^0_{ps} = 0$, so that aggregate nonpoint permits are $\hat{r}^0 = \hat{r}^0_{nps}$) and also assuming that the emissions constraint is satisfied as an equality, then $\hat{e}_{ps}$ can be eliminated as a choice variable so that $C = c(e) + q[e - (1/t)\hat{r}_{ps} - \hat{e}^0_{ps}] + p[\hat{r}_{ps}]$. The resulting first-order conditions are

$$(1) \qquad \frac{\partial C}{\partial e} = c'(e) + q = 0$$

$$(2) \qquad \frac{\partial C}{\partial \hat{r}_{ps}} = -\left(\frac{1}{t}\right) q + p = 0$$

where the second equality in equation (2) emerges in a competitive market equilibrium. This condition indicates indifference between point and nonpoint permits at the margin, implying $t = q/p$. Using this relation and substituting the permit constraint into the cost function, we have $C = c(e) + q[e - \hat{e}^0_{ps}]$.

Similarly, the nonpoint source will choose input use, nonpoint source permit holdings, $\hat{r}_{nps}$, and point source permit holdings, $\hat{e}_{nps}$, to minimize costs. Assume, as in existing trading programs, that farms do not initially hold point source permits (i.e., $\hat{e}^0_{nps} = 0$, so that aggregate emissions permits are $\hat{e}^0 = \hat{e}^0_{ps}$). Then following the steps used above for point sources, the farmer's control costs can be written as $G = g(x) + p[E\{r(x, \theta)\} - \hat{r}^0]$. The farm's necessary condition for optimal input use is

$$(3) \qquad \frac{\partial G}{\partial x} = \frac{\partial g}{\partial x} + pE\left\{\frac{\partial r}{\partial x}\right\} = 0.$$

The market solution is characterized by equation (3) along with the firm's necessary conditions (1) and (2), and the market clearing condition

$$(4) \qquad \hat{e}^0 + \left(\frac{1}{t}\right)\hat{r}^0 \geq e + \left(\frac{1}{t}\right) E\{r(x, \theta)\}.$$

The Economic Optimum

A first-best or optimal trading program is designed by choosing the aggregate number of permits (in either denomination) and the trading ratio to minimize TC subject to polluters' behavior in the market as given by conditions (1)–(4). The solution is depicted in figure 1. The optimal trading ratio (derived formally below and also in Shortle 1987) is $t = E\{\partial D^{\#}/\partial e\}/E\{\partial D^{\#}/\partial x\}$, where the superscript # indicates that damages are evaluated

[2] For simplicity, we assume that there is no asymmetric information and that the regulator can easily observe the choices made by point and nonpoint sources. Johansson (2002) provides a model in which there is asymmetric information between the regulator and nonpoint sources, although his analysis is not concerned with the trading ratio.

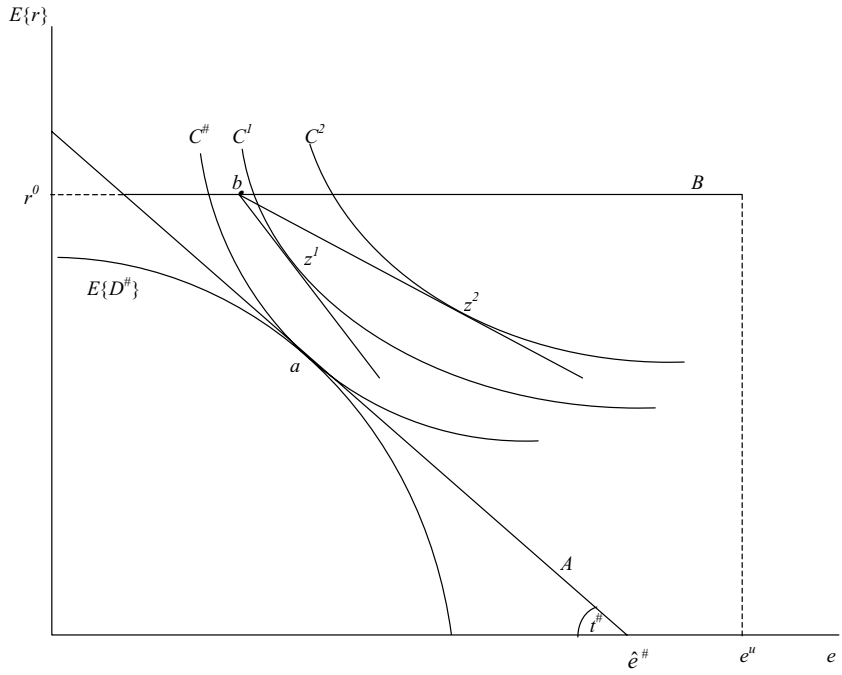

Figure 1. Comparing the optimal and conditionally optimal trading markets

given optimal values for e and x.[3] When drawn in $(e, \mathrm{E}\{r\})$ space, this optimal ratio is the slope of the tangent between the optimal iso-cost curve, $C^{\#}$, and the optimal iso-expected damage curve, $\mathrm{E}\{D^{\#}\}$. The tangent, labeled A, is the locus of feasible posttrade allocations that fully utilize the available permits under the design rules. The optimal aggregate number of permits, denominated in terms of emissions, is $\hat{e}^{\#}$. However, the optimal equilibrium a is attainable from any initial allocation that lies on A, given the trading ratio $t^{\#}$—that is, given no transactions costs associated with trading. So the initial allocation does not matter for the optimum given our assumptions.

The Conditional Optimum

Now consider the more realistic choice of a second-best or "conditionally optimal" trading ratio by the trading authority, given that the allocation of permits has been explicitly or implicitly pre-specified by a superseding authority. Specifically, the farm's initial permit allocation is equal to its expected initial loads, $\hat{r}^0_{\mathrm{nps}} = \mathrm{E}\{r(x^0, \theta)\}$, while the firm's permits are determined by the EPA under the NPDES permit system. This is consistent with existing trading programs that are usually administered by the states and set up independently of the EPA's permit choice for the firm.[4] The choice of trading ratio in this situation can only be second-best, as a first-best outcome requires two instruments to address the two externalities (point and nonpoint) (Tinbergen). As we show below, the optimal program design emerges as a special case of the restricted case we consider, in which the exogenously chosen permits are set at the optimal level.

There are two ways to derive the conditionally optimal outcome. A primal approach would be to use conditions (1)–(4) to derive firms' response functions to the endogenous trading ratio and the exogenous permit allocation, plug these into TC, and then choose the trading ratio that minimizes TC. In contrast,

[3] Woodward derives a trading ratio that maximizes environmental gains from trade, irrespective of costs, given a fixed permit supply and perfect substitutability of emissions. He finds a large ratio is optimal, although only ratios >1 produce gains in his model.

[4] Even if states could choose point source permit levels, they may still set this value too large to lie on the optimal trading locus, given initial nonpoint permits equal to historical levels. For instance, there may be political barriers to setting a low enough value because, under existing programs, point sources are ultimately held responsible for meeting water quality goals if nonpoint sources do not meet their obligations under a trade (GLTN; U.S. EPA). There may also be physical barriers if the required value is negative (e.g., if segments A and B intersected to the left of the vertical axis in figure 1).

a dual approach is to take as given the farm's input demand function ($x(p)$) and the firm's emissions demand function ($e(q)$) that result from the firms'/farms' first-order conditions (1) and (3), and then choose permit prices optimally subject to the constraint that the initial allocation of permits is set exogenously. The regulator has one choice variable under the unconstrained primal approach outlined here, but two choice variables under the constrained dual approach. The constraint on the dual approach makes it possible to maximize over two choice variables, although it essentially renders one of the choices to be trivial. Still, insight is gained by examining how the shadow value of the constraint affects equilibrium prices.

We adopt the dual approach. Plugging the derived demand functions into TC, p and q are chosen to minimize $\text{TC} = c(e(q)) + g(x(p)) + \text{E}\{D(e(q), r(x(p), \theta)\}$, subject to condition (4). After substituting $q/p = t$ into condition (4), the Lagrangian is $L = c(e(q)) + g(x(p)) + \text{E}\{D(e(q), r(x(p), \theta)\} + \lambda[(\hat{e}^0 - e) + (p/q)(\hat{r}^0 - \text{E}\{r(x, \theta)\})]$, where λ is the shadow value of increased permits. The shadow value λ is positive assuming too many permits are administered (since the farm is not regulated initially), so that a decrease in total permits would reduce TC while an increase in permit levels would increase TC. If the number of permits were to equal the number that would be chosen in a first-best optimum, then λ would equal zero since p and q would optimally be set equal to the values that would minimize TC in the unconstrained, first-best case.

The necessary conditions for the conditionally optimal problem are equation (4) and

$$\frac{\partial L}{\partial q} = c'(e)\frac{de}{dq} + \text{E}\left\{\frac{\partial D}{\partial e}\right\}\frac{de}{dq} - \lambda\frac{de}{dq} + \lambda\frac{\partial(1/t)}{\partial q}\hat{r}^0 - \lambda\frac{\partial(1/t)}{\partial q}\text{E}\{r\} = 0 \tag{5}$$

$$\frac{\partial L}{\partial p} = g'(x)\frac{dx}{dp} + \text{E}\left\{\frac{\partial D}{\partial r}\frac{dr}{dx}\right\}\frac{dx}{dp} - \lambda\frac{p}{q}\text{E}\left\{\frac{dr}{dx}\right\}\frac{dx}{dp} + \lambda\frac{\partial(1/t)}{\partial p}(\hat{r}^0 - \text{E}\{r\}) = 0 \tag{6}$$

where $\partial(1/t)/\partial q = -p/q^2$ and $\partial(1/t)/\partial p = 1/q$. Consider equation (5). An increase in the permit price q has two effects: (*a*) it leads to a decrease in point source emissions and (*b*) it increases the trading ratio. The first three right-hand-side (RHS) terms (relative to the first equality) reflect the impact of a decrease in emissions, with the first two terms reflecting standard trade-offs: a decrease in emissions increases abatement costs (increasing L) and decreases expected damages (decreasing L). The third RHS term reflects the impact of decreased emissions on the constraint: at the margin, a further reduction in point source emissions is socially costly (increasing L). This term would not arise in a first-best optimum because $\lambda = 0$ in that case.

The final two RHS terms in equation (5), which would also not arise in a first-best optimum, reflect the impact of an increase in q on the trading ratio—or rather $1/t$ in this case. The trading ratio has two effects. On the one hand, it is partially responsible for defining the total number of permits (expressed in terms of emissions) in the market (the permit quantity effect): $\hat{e}^0 + (1/t)\hat{r}^0$. This is not true in a first-best optimum where $\hat{e}^0$ and $\hat{r}^0$ are chosen optimally—the choice of t is irrelevant for defining the number of permits in this case because $\hat{r}^0$ can be adjusted accordingly. But in the conditional optimum, $\hat{r}^0$ is exogenous and so the choice of t matters. The total number of permits is clearly reduced as $(1/t)$ is diminished. A marginal increase in q reduces the number of permits by $[\partial(1/t)/\partial q]\hat{r}^0$, which reduces social costs (decreasing L) assuming too many permits are initially administered. Hence, the initial endowment matters.

The second effect of the trading ratio (the permit price effect) is to influence how much trading occurs, as a larger t (or a smaller $1/t$) makes it more expensive for point sources to purchase nonpoint permits. A marginal increase in q therefore reduces the number of trades by $[\partial(1/t)/\partial q]\text{E}\{r\}$, which is costly at the margin (increasing L) due to the oversupply of permits. With $\hat{r}^0 > \text{E}\{r\}$, the permit quantity effect dominates the permit price effect, so that the net effect of a larger q on t is a reduction in social costs (decreased L). Combining these last two terms, $\lambda[\partial(1/t)/\partial q][\hat{r}^0 - \text{E}\{r\}] < 0$, the net effect is larger. The larger is the excess supply of nonpoint permits, $[\hat{r}^0 - \text{E}\{r\}]$ (or the larger is the excess demand for point source permits, since $[\hat{r}^0 - \text{E}\{r\}]/t = [\hat{e}^0 - e]$ in equilibrium), so that there are net social benefits from

increasing q and hence lowering t in this case to effectively reduce total permit levels.[5]

The trading ratio terms arise in the conditional optimum because t must perform two tasks in this case—determine the quantity of nonpoint permits and determine the relative price of these permits—and it cannot do both optimally. This is not an issue in the first-best optimum because the additional degree of freedom associated with the choice of nonpoint permits takes the burden off of the trading ratio so that it can be chosen only to affect the relative price of the permits. Note that these results do not necessarily vanish when point and nonpoint emissions are perfect substitutes because the trading ratio still performs double-duty in the conditional optimum. Finally, equation (6) has an identical interpretation, although with opposite effects on the trading ratio since a larger value of p decreases t (and hence increases $1/t$).

The conditionally optimal permit prices are derived by plugging the results of expressions (1) and (3), along with $t = q/p$, into conditions (5) and (6)

$$(7)\quad q = \mathrm{E}\left\{\frac{\partial D^*}{\partial e}\right\} - \lambda^* - \lambda^* \frac{p}{q^2}[\hat{r}^0 - \mathrm{E}\{r\}]\frac{dq}{de} = \mathrm{E}\left\{\frac{\partial D^*}{\partial e}\right\} - \lambda^* - \lambda^* s_e \varepsilon^*_{qe}$$

$$(8)\quad p = \mathrm{E}\left\{\frac{\partial D^*}{\partial r}\right\} + \frac{\mathrm{cov}\{\partial D^*/\partial r, \partial r^*/\partial x\}}{\mathrm{E}\{\partial r^*/\partial x\}} - \frac{\lambda^*}{t^*} + \frac{\lambda^*}{t^*} s_r \varepsilon^*_{pr}$$

where $\varepsilon^*_{qe} < 0$ is the firm's inverse elasticity of demand for emissions, $\varepsilon^*_{pr} < 0$ is the farm's inverse elasticity of demand for expected pollution loads, $s_e = (e^* - \hat{e}^0)/e^*$ is the firm's proportional excess demand for point source permits, $s_r = (\hat{r}^0 - \mathrm{E}\{r^*\})/\mathrm{E}\{r^*\}$ is the farm's proportional excess supply of nonpoint permits, and the superscript $*$ indicates that all variables are set at their conditionally optimal values. Note that the second equality in equation (7) comes from setting (4) as an equality: $(\hat{r}^0 - \mathrm{E}\{r^*\})/t = (e^* - \hat{e}^0)$.

These two prices are designed to address two externalities: one by the point source and the other by the nonpoint source. However, due to the constraint (4), it is not possible to fully address either externality. So in this sense the prices in equations (7) and (8) are only second-best. As described above, the economically optimal permit price for each source equals the expected marginal damages created by the source. This result emerges from conditions (7) and (8) when the aggregate number of permits is chosen optimally and hence $\lambda = 0$. But, in the present case, an inefficient number of permits and a suboptimal number of policy tools reduce the ability of the market to create efficient incentives to reduce expected damages. In consequence, the conditionally optimal permit prices are modified by two additional terms to account for these two sources of inefficiency. Additional terms typically arise for second-best incentives to reflect the impact of the incentive on externalities or distortions that the regulator is unable to perfectly target in a second-best world (e.g., Baumol and Oates).[6]

Consider the point source permit price, q. The second RHS term in equation (7) (relative to the second equality sign), $-\lambda^* < 0$, is an adjustment that reduces the emissions permit price relative to expected marginal damages. This adjustment accounts for the oversupply of permits in the market, which would tend to decrease the permit price relative to the optimum value. The third RHS term, $-\lambda^* s_e \varepsilon^*_{qe} > 0$, accounts for the fact that q plays a role (through

[5] If too few permits were initially allocated (i.e., over control), then an increase in permits decreases TC; hence $\lambda < 0$ and the signs of the three final terms in equations (5) and (6) are reversed.

[6] Lipsey and Lancaster first formalized the concept of second-best, finding that "first-best" rules (e.g., marginal cost pricing) in one sector of the economy may be sub-optimal if distortions (i.e., prices not equal to marginal cost) remain in one or more other sectors. In the environmental economics literature, the classic example is of a monopolist, where two distortions exist: pollution and an inefficient output price. An emissions tax cannot efficiently address both distortions, so a Pigouvian tax is not optimal. Rather, the second-best emissions tax in this case equals marginal damages plus a term that accounts for the impact of the tax on reducing the pre-existing monopoly distortion. More generally, the concept of second-best is applicable when legal, institutional, or informational constraints restrict policy makers' choice or design of policy instruments in a way that prevents them from achieving first-best allocations (i.e., usually meaning Pareto Optimal) (Mas-Colell, Whinston, and Green; Boadway). For example, when polluting discharges are imperfectly mixed, an optimal emissions tax is source-specific to account for the imperfect substitutability of emissions from alternative sources (Baumol and Oates). Similarly, tradable discharge permits should not generally be traded one to one (1:1) under such conditions (Tietenberg) (although 1:1 trading is standard in most point–nonpoint programs). In such contexts, a restriction requiring uniform tax rates or 1:1 trading creates second-best problems.

its impact on t) in affecting the number of available permits. As described above, since there are too many permits in equilibrium, there are benefits from increasing q in order to increase t and hence decrease total permit numbers. The benefits of this are greater, the larger is the excess demand for permits by point sources (i.e., the larger is s_e), so essentially this term is a corrective tax based on the firm's excess demand for emissions; the initial allocation matters. Prior models of trading have only found the initial allocation to matter when there are transactions costs associated with trades (which could implicitly arise if there are restrictions on how trades could take place) or when firms behave strategically (e.g., due to imperfect competition) (see Hanley, Shogren, and White for an overview). In the present model, the initial allocation matters primarily because it affects the *path* to the final allocation by way of the trading ratio, which is determined by the permit price q and also p. We describe the impact of the initial allocation in more detail below.

The same sorts of results emerge with respect to the nonpoint permit price, p, with three exceptions. First, there is an additional covariance term in the nonpoint permit price to account for the risk created by nonpoint source loads. Assuming damages are convex in r, then the covariance term is positive: the stochasticity of nonpoint loads creates socially costly risk and so there are incentives to reduce nonpoint loads in order to reduce this risk (Shortle). Second, λ^* is scaled by $1/t^*$ to reflect the fact that nonpoint source permits are denominated in terms of expected nonpoint loads as opposed to emissions. Third, the final term in equation (8), $(\lambda^*/t^*)s_r\varepsilon^*_{pr}$, represents a corrective subsidy instead of a tax. The reason is that the associated reduction in p increases t and hence reduces the total number of permits. The benefits of reducing p for this purpose are greater, the larger is the excess supply of nonpoint permits, s_r.

The Conditionally Optimal Trading Ratio

Using the result that $tp = q$ in a market equilibrium, we can substitute in expressions (7) and (8) and solve for t to obtain the following equilibrium condition for the trading ratio

$$t = \frac{\mathrm{E}\{\partial D^*/\partial e\} - \lambda^* s_e^* \varepsilon_{qe}^* - \lambda^* s_r^* \varepsilon_{pr}^*}{\mathrm{E}\{\partial D^*/\partial r\} + \left(\mathrm{cov}\{\partial D^*/\partial r,\, \partial r^*/\partial x\} \big/ \mathrm{E}\{\partial r^*/\partial x\}\right)}. \tag{9}$$

For the special case in which permit levels are set optimally, $\lambda = 0$ and t reduces to

$$t = \frac{\mathrm{E}\{\partial D^*/\partial e\}}{\mathrm{E}\{\partial D^*/\partial r\} + \left(\mathrm{cov}\{\partial D^*/\partial r,\, \partial r^*/\partial x\} \big/ \mathrm{E}\{\partial r^*/\partial x\}\right)}. \tag{10}$$

This is the standard result in the literature (Shortle). The trading ratio depends on the relative expected marginal impacts from each source, adjusted for the additional risk created by nonpoint sources. The effect of the covariance term is to reduce t. In consequence, t will be less than one if nonpoint loads and point source emissions have similar expected marginal impacts or if point source emissions have smaller expected marginal impacts. Otherwise, the ratio may be greater than one.

Two additional terms emerge in the numerator when the number of permits is chosen exogenously (see equation (9)). These terms account for the additional role that t now plays—defining the number of total nonpoint permits. Consistent with the discussion above, both terms have the effect of increasing the trading ratio so as to decrease the number of permits, as would be expected when the permit supply is inefficiently large.

Figure 1 illustrates the intuition behind this result. As existing trading programs are structured, the initial permit allocation is not likely to lie on A. Rather, the initial allocation will lie to the right of A at a point along segment B, which defines the possible initial permit combinations given initial nonpoint permits equal to r^0 and given initial emissions permits less than some unregulated level, e^{u}, but greater than the value that would place society on the optimal trading locus A (i.e., too many initial permits relative to the optimum). The regulator's objective is to choose a trading ratio so that polluters can trade from the initial allocation on B and move toward a more efficient equilibrium. Although a is optimal when both the trading ratio and the number of permits can be chosen, a is unattainable from any point on

B (except the intersection of B and A) for any constant trading ratio. This is because in a decentralized trading market the polluters will minimize their control costs by choosing the pollution allocation where the trading curve is tangent with their iso-cost curve, the optimal tangent A only intersects B at a single point. So another equilibrium must be pursued.

As an example, suppose the initial allocation was at b in figure 1. Implementing a trading ratio greater than the optimal ratio $t^{\#}$ would lead to an equilibrium such as z^1 while a trading ratio smaller than $t^{\#}$ would lead to an equilibrium such as z^2. Visually, z^1 is "closer" to the optimum a, with control costs higher and expected damages lower at z^1 than at z^2. Given too many initial permits—or too little pollution control—and given convex cost and expected damage functions, we would expect that the expected marginal benefits of pollution control (the reduction in expected damages) are greater than the marginal costs of control. In this case it is beneficial to pursue a lower iso-expected damage curve at the expense of control costs, and this is accomplished by choosing a relatively steep trading ratio so as to reduce the number of permits.[7] We now turn to a numerical example to provide additional insights.

Numerical Example: The Susquehanna River Basin

We analyze optimal and conditionally optimal trading programs for the Pennsylvania portion of the SRB. The SRB is the major source of freshwater and also nutrients entering the Chesapeake Bay (Chesapeake Bay Program). Approximately, 31 million kg of nitrogen annually load into SRB water resources, with about 87% coming from (primarily agricultural) nonpoint sources (Nizeyimana et al.). Reducing SRB nutrient loads is a key challenge for state and federal agencies developing programs to improve water quality in the Susquehanna River and the Chesapeake Bay, and a major reason for the EPA's new trading program in this area. The simulation model for this region is taken from Horan, Shortle, and Abler (2002) and is outlined in the Appendix.

[7] If the permit supply was inefficiently small (so that $\lambda < 0$), then t would be smaller in the conditional optimum. The reason is that marginal costs of pollution control would exceed the expected marginal benefits, so society would gain from reducing t to increase permit numbers and reduce control costs at the expense of increased expected damages.

Results for the baseline model are reported under Scenario I in table 1 for an optimally designed trading program and several conditionally optimal programs. For the conditionally optimal scenarios, total permit levels depend on historical nonpoint pollution levels and allowable point source emissions, denoted ϕe^{u}, where e^{u} represents baseline emissions levels and ϕ is a percentage reduction from these levels. Results are reported for various values of ϕ.

The first column of results reports an efficiency gain index (EGI). An index is used due to our interest in the relative (as opposed to absolute) performance of the trading programs and also to overcome some scaling effects. EG for a particular scenario is calculated as the percentage reduction in TC relative to the baseline data that consists of an unregulated equilibrium for agricultural sources and some prior degree of controls for point sources, i.e., $\text{EG} = (\text{TC}^{\text{B}} - \text{TC}^{\text{s}})/\text{TC}^{\text{B}}$, where TC^{s} represents expected social costs in scenario s and TC^{B} represents baseline expected social costs. The maximum EG occurs in an optimal trading program. We therefore divide each scenario's EG by that of the optimal trading program; hence, the EGI for the optimal program is 100.

The next four columns, respectively, report nonpoint control costs, nonpoint gains after selling its excess permits to point sources, point source control costs, and expected damage costs for the various scenarios. These costs are represented by indices, with the base in each case being the efficient level of total costs, $\text{TC}^{\#}$. Net sales are given by $p(\hat{r}^0 - \text{E}\{r^*\})/\text{TC}^{\#}$. In the conditionally optimal scenarios, initial nonpoint permits equal mean unregulated nonpoint loads, i.e., $\hat{r}^0 = \text{E}\{r(x^0)\}$. This is not the case in the optimal trading scenario, however, because trades must occur along the trading locus (e.g., tangent A in figure 1) and this would involve negative point source permits for $\hat{r}^0 = \text{E}\{r(x^0)\}$ (e.g., while segments A and B intersect to the right of the vertical axis in figure 1, they would intersect to the left of the axis in the context of the current simulation). We therefore define sales in the optimal scenario to be from an initial allocation with zero point source permits and nonpoint permits arising at the intersection between the trading locus and the vertical axis (i.e., $\hat{r}^0 = \text{E}\{r(x^*)\} + t^* e^*$). Of course, any combination of nonnegative initial permit levels along the trading locus would be feasible in the optimal solution.

The last three columns in table 1 report the trading ratios as well as equilibrium permit

Table 1. Numerical Results for the SRB

Scenario	Economic Outcomes					Policy Variable	Equilibrium Permit Levels	
	EGI	NPS Pollution Control Cost Index	NPS Gains after Permit Sales Index	PS Pollution Control Cost Index	Expected Damages Index	Trading Ratio	Expected NPS Loads	PS Emissions
I. Baseline scenario								
Optimal trading	100	18.8	6.3	7.0	74.2	0.89	69.5	94.2
Conditionally optimal								
$\phi = 0.6$	79.7	5.0	5.7	9.6	90.8	2.9	83.3	86.5
$\phi = 0.2$	96.6	12.5	28.0	10.0	78.3	1.8	74.7	85.5
II. Greater risk[a]								
Optimal trading	100	25.5	4.7	8.3	66.2	0.85	66.1	92.1
Conditionally optimal								
$\phi = 0.6$	78.3	5.8	6.7	12.0	90.6	3.3	81.3	83.9
$\phi = 0.2$	95.7	16.0	20.2	13.2	72.6	2.06	72.3	81.2
III. Smaller nonpoint loads[b]								
Optimal trading	100	7.8	9.4	16.0	76.2	0.90	79.3	74.5
Conditionally optimal								
$\phi = 0.6$	99.3	5.3	6.1	16.1	78.8	1.12	82.8	74.4
$\phi = 0.2$	89.3	15.2	19.4	12.6	74.5	0.44	71.8	79.5

[a]The elasticity of damages is increased to two and the variance of precipitation is increased four-fold.
[b]Baseline nonpoint source loads are divided by five.

levels in each scenario. Equilibrium permit levels are defined as an index, with the base being baseline point source emissions and nonpoint loads.

Results for the optimal trading program indicate that nonpoint sources incur almost three times the control costs of point sources. This is in accordance with the fact that t is less than unity, encouraging more nonpoint controls at the margin (and is consistent with prior work, e.g., Horan, Shortle, and Abler 2002; Horan et al.). This result is due mainly to the risk associated with nonpoint pollution and it stands in stark contrast to the ratios of 2 or 3 that are often used in practice (see table 1 in Horan).

Results for the conditionally optimal program are presented for two percentage reductions $(1 - \phi)$ in point source emissions from baseline levels. Note that the trading locus would involve negative point source permits when $\hat{r}^0 = \mathrm{E}\{r(x^0)\}$ because nonpoint loads are a large proportion of aggregate initial discharges, and so no positive value of ϕ can yield the optimal outcome. When $\phi = 0.6$ (so that $\hat{r}^0 = 100$ and $\hat{e}^0 = 60$), the optimal trading ratio is 2.9 and this coincides with an EGI of 79.7, or roughly a 20% loss in the EG that could have been achieved under an optimal program. The efficiency loss arises because nonpoint sources pollute too much (loads of 83.3 as compared to 69.5 under optimal trading) and point sources pollute too little (emissions of 86.5 as compared to 94.2 under optimal trading). This results in inefficiently low nonpoint costs (an index of 5.0 as compared to 18.8 under optimal trading), excessive point source control costs (9.6 versus 7.0), and excessive expected damages (90.8 versus 74.2). In terms of actual control costs, nonpoint sources benefit at the expense of point sources. This makes sense given the initial permit entitlement that nonpoint sources are provided within the conditional optimum. However, nonpoint sources are strictly better off (and point sources worse off) in the efficient scenario after permit sales are accounted for, since point sources purchase more nonpoint controls in the optimal trading program (at least, given the assumed permit allocation).

When ϕ is decreased to $\phi = 0.2$ (so that $\hat{r}^0 = 100$ and $\hat{e}^0 = 20$), both the trading ratio and the efficiency loss decrease, primarily due to additional reductions in nonpoint loads. In the context of figure 1, a smaller ϕ corresponds to an initial allocation that lies farther to the left along segment B. When starting from such initial allocations, smaller trading ratios are capable of guiding trades to a lower iso-expected damage curve. Finally, relative to the optimal trading scenario, nonpoint sources again incur lower control costs at the expense of point sources, although not as much as when $\phi = 0.6$. But in contrast to the case where $\phi = 0.6$, nonpoint sources are strictly better off (and point sources worse off) in the conditional optimum after sales are accounted for. This is because nonpoint sources receive a greater entitlement in the conditional optimum: the final allocations in both the optimum and conditional optimum are similar, but initial nonpoint permits in the conditional optimum equal unregulated levels while, as explained above, initial nonpoint permits are less than this in the optimum.

Comparative Statics

Scenarios II and III in table 1 represent two alternative specifications that are used to investigate the impact of some key model features (i.e., risk, heterogeneity, and initial loads) on the results. In each case, costs are still expressed relative to the optimum for that scenario, so that costs are not generally comparable across scenarios. Our main focus is on the trading ratio and relative permit levels, which are generally comparable.

In Scenario II, risk is increased by making damages more convex and by increasing the variance of precipitation. The optimal trading ratio falls slightly, as is expected from equation (10), to increase the relative value of nonpoint source permits and encourage more nonpoint controls. Moreover, risk makes pollution more socially costly and so there is also a move to a lower iso-expected damage curve, as evidenced by smaller equilibrium permit levels for both sources relative to Scenario I. In contrast, the conditionally optimal ratios increase, as this is the only way to proceed to a lower iso-expected damage curve given the initial allocation. Indeed, the conditional optima exhibit reduced equilibrium permit levels for both sources relative to the analogous conditional optima in Scenario I. Mathematically, increased risk affects the conditionally optimal trading ratios in two ways: (*a*) it diminishes the ratio by increasing the covariance term in equation (9) (i.e., increasing the ratio's permit price effect) and (*b*) it increases the ratio by increasing the ratio's permit quantity effect. The conditionally optimal trading ratios are larger in Scenario II than in Scenario I because the effect of increased risk on the ratio's permit quantity effect dominates.

In Scenario III, we divided baseline nonpoint loads by five (to make initial point and nonpoint loads more comparable) to consider how changes in baseline loads affect the results. As expected, the optimal ratio changes minimally and only to reflect a recalibration of the nonpoint loadings functions given the new baselines. The optimal allocation of controls does change, however, placing more weight on point source controls. This is because the exogenous reduction of initial nonpoint loads increases nonpoint marginal abatement costs, as nonpoint sources now have fewer loads that they can abate. Now consider the conditionally optimal outcomes. When $\phi = 0.6$, the trading ratio is significantly reduced relative to Scenario I. Graphically, the segment B in figure 1 has shifted down and a flatter trading locus can lead to the same iso-expected damage curve. Mathematically, the permit quantity effect as impacted by the term s_r is reduced, reducing the ratio (see equation (9)). When $\phi = 0.2$, the trading ratio is even lower than the optimal ratio! This is because the reduction in baseline nonpoint loads combined with the small initial number of point source permits has resulted in too few permits relative to the optimum. Hence, $\lambda < 0$ and all the mathematical results are reversed—the permit quantity effect in equation (9) reduces the trading ratio.

Finally, although not reported in table 1, we also investigated the implications of imposing a homogeneous mean and variance of delivery coefficients across all sources of pollution (and also significantly increasing the heterogeneity of delivery). The impacts of this are quite small and do not significantly alter any of the trading ratios or permit levels.

Conclusion

Existing point–nonpoint trading programs apply trading ratios in excess of 1:1, often two to three times that, to account for the risk associated with nonpoint sources. However, prior theory and empirical research suggest that water quality risks associated with the inherent randomness of nonpoint sources would be better managed by comparatively smaller trading ratios that would encourage greater substitution of nonpoint emissions reductions for point source emissions reductions. There is, however, an important difference between theory and practice. In the theoretical optimum, the choice of trading ratio and permits is simultaneous and endogenous. In practice, as trading programs are often implemented in the United States, the permits are exogenous to the local trading authority. When the number of permits is exogenously specified at inefficiently high levels, we find both theoretical and numerical support for the use of larger ratios. This means that the current large ratios might be justified, but for different reasons than those that are normally provided.

The results also indicate the importance of correctly matching program development and implementation with theory. Program planners claim to set their trading ratios in accordance with efficiency principles derived from theories of first-best program design. But in practice they do not control the choice of the other major policy variable—permit numbers, which are jointly determined along with the trading ratio in an efficient model. In the present model, the "efficient" trading ratio could be used, but it would be a poor choice because the conditionally optimal ratio is vastly different. There may be additional considerations that we have not analyzed here, such as the impact of agri-environmental and/or other farm payments, which could also have important impacts on the optimal choice of trading ratio (Horan, Shortle, and Abler 2004). Clearly, an à la carte approach to policy development, in which program designers pick and choose some aspects of an efficient program while ignoring others, is ill-advised. Theory is needed to guide choices in the context that these choices can actually be made.

[Received October 2003; accepted May 2004.]

References

Baumol, W.J., and W.E. Oates. *The Theory of Environmental Policy*. Cambridge: Cambridge University Press, 1988.

Boadway, R. "The Role of Second Best in Public Economics." Working paper, Economic Policy Research Unit, Copenhagen Business School, 1995.

Edwards, R.E., and T.W. Stoe. "Nutrient Reduction Cost Effectiveness Analysis, 1996 Update." Susquehanna River Basin Commission Publication No. 195, Harrisburg, PA, 1998.

Faeth, P. *Fertile Ground: Nutrient Trading's Potential to Cost-Effectively Improve Water Quality*. Washington, DC: World Resources Institute, 2000.

Great Lakes Trading Network (GLTN). "Summary of Program and Project Drivers and Other Presentation Materials." Markets for the New Millennium: How Can Water Quality Trading Work for You? Conference and Workshop, Chicago, IL, 18–19 May, 2000.

Hanley, N., J.F. Shogren, and B. White. *Environmental Economics: In Theory and Practice*. New York: Oxford, 1997.

Hoag, D.L., and J.S. Hughes-Popp. "Theory and Practice of Pollution Credit Trading in Water Quality Management." *Review of Agricultual Economics* 19(1997):252–62.

Horan, R.D. "Differences in Social and Public Risk Perceptions and Conflicting Impacts on Point/Nonpoint Trading Ratios." *American Journal of Agricultural Economics* 83(2001): 934–41.

Horan, R.D., D.G. Abler, J.S. Shortle, and J. Carmichael. "Cost-Effective Point–Nonpoint Trading: An Application to the Susquehanna River Basin." *Journal of the American Water Resources Association* 38(2002):467–78.

Horan, R.D., J.S. Shortle, and D.G. Abler. "The Coordination and Design of Point–Nonpoint Trading Programs and Agri-Environmental Policies." *Agricultural and Resource Economics Review* 33(2004):61–78.

Horan, R.D., J.S. Shortle, and D.G. Abler. "Nutrient Point–Nonpoint Trading in the Susquehanna River Basin." *Water Resources Research* 38(2002):8-1–8-13, 10.1029/2001WR000853.

Johansson, R.C. "Watershed Nutrient Trading Under Asymmetric Information." *Agricultural and Resource Economics Review* 31(2002): 221–32.

Lipsey, R., and K. Lancaster. "The General Theory of Second Best." *Review of Economic Studies* 25(1956):11–32.

Mas-Colell, A., M. Whinston, and J. Green. *Microeconomic Theory*. New York: Oxford University Press, 1995.

Nizeyimana, E., B. Evans, M. Anderson, G. Peterson, D. DeWalle, W. Sharpe, J. Hamlett, and B. Swistock. *Quantification of NPS Pollution Loads within Pennsylvania Watersheds*, Final Report to the Pennsylvania Department of Environmental Protection, Environmental Resources Research Institute, The Pennsylvania State University, 1997.

Pennsylvania Agricultural Statistics Service (PASS). *1997–1998 Statistical Summary & Pennsylvania Department of Agriculture Annual Report*. Pennsylvania Department of Agriculture. Harrisburg, PA, 1988.

Sato, K. "A Two-Level Constant-Elasticity of Substitution Production Function." *Review of Economic Studies* 34(1967):210–18.

Shortle, J.S. "Allocative Implications of Comparisons Between the Marginal Costs of Point and Nonpoint Source Pollution Abatement." *Northeast Journal of Agricultural and Resource Economics* 16(1987):17–23.

Shortle, J.S., and D. G. Abler. "Nonpoint Pollution." In H. Folmer and T. Tietenberg, eds. *The International Yearbook of Environmental and Resource Economics 1997/98*. Cheltenham, UK: Edward Elgar, 1997, pp. 156–204.

Smith, R.A., G.E. Schwarz, and R.B. Alexander. "Regional Interpretation of Water Quality Monitoring Data." *Water Resources Research* 33(1997):2781–98.

Stavins, R.N. "Transaction Costs and Tradeable Permits." *Journal of Environmental Economics and Management* 29(1995):133–48.

Tietenberg, T.H. "Tradeable Permits for Pollution Control When Emission Location Matters: What Have We Learned?" *Environmental and Resource Economics* 5(1995):95–113.

Tinbergen, J. *On the Theory of Economic Policy*. Amsterdam: North-Holland, 1952.

USDA, Economic Research Service, Corn Costs and Returns, 2000. Available at http//www.ers.usda.gov/briefing/farmincome/car/corn2.htm.

USDA and USEPA. *Clean Water Action Plan: Restoring and Protecting America's Waters*. U.S. Environmental Protection Agency EPA-840-R-98-001, Washington DC, 1998.

U.S. Environmental Protection Agency (U.S. EPA). "Water Quality Trading Policy; Issuance of Final Policy," *Federal Register*. January 13, 2003 68(8):1608–13. (From the Federal Register Online via GPO Access [wais.access.gpo.gov] [DOCID: fr13ja03-53]).

Whitman, C.T. "Remarks of Governor Christine Todd Whitman, Administrator of the U.S. Environmental Protection Agency, on Announcing the Water Quality Trading Policy." U.S. Environmental Protection Agency, Washington DC, January 13, 2003.

Woodward, R.T. "The Environmentally Optimal Trading Ratio." Paper presented at the Annual Meeting of the American Agricultural Economic Association, Chicago, August 2001.

Appendix

The Simulation Model

Nonpoint sources. Most SRB nonpoint loads are due to agriculture, with corn production being the most important contributing agricultural activity. Corn production is modeled as a two-level, constant elasticity of scale technology (Sato). All input and output prices, except land prices, are fixed. Land supply is defined at the watershed level to reflect

the opportunity cost of this input, which is likely to differ in each region. The economic model is calibrated for each region using cost shares and production shares developed from USDA and Pennsylvania data (PASS). For the production and supply elasticities, we adopt the mean of the values used by Horan, Shortle, and Abler (2002).

Regional nonpoint loadings functions (defined as the amount of nitrogen entering the Susquehanna River or its tributaries from that region) are derived from the results of a research team that used the simulation model Generalized Watershed Loadings Function to develop TMDL recommendations for Pennsylvania. The loadings functions are stochastic due to stochastic precipitation.

Point sources. Point source abatement cost functions are derived using Susquehanna River Basin Commission data (Edwards and Stoe) for the most important point sources of nitrogen in the SRB. The data includes base-level emissions and estimated costs and abatement levels for various nutrient control technologies.

Nutrient delivery. We model the fraction of the pollution from each watershed that is delivered to the Chesapeake Bay as a stochastic delivery coefficient, with the mean and variance derived from results of the USGS SPARROW model (Smith, Schwarz, and Alexander).

Damages. Damages are a quadratic function of delivered loads, calibrated using the means of the parameters described in Horan, Shortle, and Abler (2002).