Opportunities for Reducing Total Maximum Daily Load (TMDL) Compliance Costs: Lessons from the Chesapeake Bay

Lisa A. Wainger*
University of Maryland, Center for Environmental Science, Chesapeake Biological Laboratory, 146 Williams Street, P.O. Box 38 Solomons, Maryland 20688, United States

ABSTRACT: The Chesapeake Bay Total Maximum Daily Load (TMDL) program is an unprecedented opportunity to restore the Chesapeake Bay, yet program costs threaten to undermine its complete implementation. Analyses of Bay TMDL program design and implementation were used to relate program cost-effectiveness to choices in (1) compliance definitions, (2) geographic load allocations, and (3) approaches to engaging unregulated sources. A key finding was that many design choices require choosing an acceptable level of risk of achieving water quality outcomes, and a lack of data can lead to precautionary choices, which increase compliance costs. Furthermore, although some choices managed costs, others decisions may have reduced the potential for cost savings from water quality trading and payment programs. In particular, the choice by some states to distribute the portion of load reductions that improve water quality in the Bay mainstem to many small basins is likely to diminish the potential for market development or reduce funding for the most cost-effective nutrient and sediment reduction practices. Strategies for reducing costs of future TMDLs include considering diminishing marginal returns early in the TMDL design to balance costs and risks in regulatory goal setting and to design rules and incentives that promote innovation and cost-effective compliance strategies.

INTRODUCTION
U.S. EPA regulators have developed limits on the nutrients and sediments allowed to enter the Chesapeake Bay in order to achieve a desirable social outcome: a healthier and more sustainable Bay. Yet, a recent Farm Bureau lawsuit against the EPA is just the most visible evidence of a concern that appears widespread among local governments and businesses—that the costs of achieving the Chesapeake Bay “pollution diet” (or Total Maximum Daily Loads, TMDLs) are too high (e.g., 1). Regulators had aimed to minimize the cost burden placed on polluters through the use of water quality trading markets, user fees, tax incentives, and other approaches. So, is this angst just more of the usual case of polluters not wanting to pay? Or could additional strategies be applied to reduce the costs of compliance?

A regulatory program can reduce the cost burden of regulation through two main approaches: (1) by setting pollution caps where they generate net social benefits and (2) by designing programs that allow regulated parties to seek the most cost-efficient way to achieve goals. Net social benefits (i.e., economic efficiency), in the strictest sense of the term, are when the monetized benefits of a regulation exceed the costs. However, this criterion is not always used because cost–benefit analysis can be an unwieldy tool that often fails to capture important aspects of societal well-being due to limitations of understanding and measurement capabilities.

Cost-efficiency analysis is an alternative to cost–benefit analysis that compares a quantitative, usually nonmonetary, indicator of benefits to judge what level of protection (i.e., pollution reduction) balances costs with desirable outcomes. Cost-efficiency can inform the design of a TMDL by evaluating, for example, the payoff of setting pollution caps to manage for extreme conditions, because costs tend to rise rapidly the more one tries to eliminate risk.

Received: April 9, 2012
Revised: July 15, 2012
Accepted: August 15, 2012
Published: August 15, 2012
The desire to reduce costs of compliance is driving the development of pollution trading in TMDL programs. The TMDLs are pollution caps and the first necessary ingredient to a “cap and trade” approach that harnesses market forces to reduce compliance costs by allowing those with high costs of reducing pollution to purchase credits (offsets) from those with lower costs. However, the caps are only the first of many decisions that must align with market drivers, if markets are to realize their potential for reducing costs. Markets for credits are not the same as markets that evolve to fulfill the needs of motivated buyers and sellers, rather they depend on regulators creating a demand for offsets, typically through regulation.

The question addressed in this paper is: What choices in TMDL program design and implementation promote cost-efficiency? A related question is: How can TMDL implementation decisions be aligned with trading market requirements or other approaches used to reduce costs of compliance? The Chesapeake Bay TMDL provides a useful case study for demonstrating strategies that were or could be used to reduce costs of TMDLs.

TMDL BACKGROUND

Under section 303(d) of the Clean Water Act (CWA), states, territories, and authorized tribes are required to develop lists of impaired waters, establish priority rankings of impaired waters, and develop Total Maximum Daily Loads (TMDLs) for priority waters, which are the maximum amount of one or more pollutants that a waterbody can receive and still meet water quality standards. In the Chesapeake Bay, the U.S. EPA developed the TMDLs for the mainstem and tidal tributaries to promote the coordination required among the six states and the District of Columbia in the watershed. The states and the District have now finalized the detailed (Phase II) watershed implementation plans (WIPs) to achieve the pollutant load limits.

Of note in the Bay TMDL design process is that implementation costs were not estimated until after caps were set and allocated to geographic areas and source sectors (e.g., wastewater, agriculture, stormwater, etc.), an approach that is common among TMDLs. The choice to wait to estimate costs is significant because almost all of the choices made in the design phase influence costs by affecting the relative difficulty of achieving reductions or the ability to take advantage of cost-effective compliance approaches, including trading.

To be clear, programs can be cost-effective even when they create high costs since high benefits may warrant imposing high costs. However, it can also be the case that accepting lower expected benefits may be deemed necessary for managing costs. A reduction in expected benefits does not necessarily translate to setting consistently lower water quality standards, but rather, may require accepting a higher level of risk that standards will be violated in some times and places. For example, engineers are often specific about their risk choices; a levee is said to be designed to withstand a 50-year storm, but not a 100-year storm, within some margin of error. This information about risk reduction is compared to costs to choose an acceptable project design. TMDL design is more complicated than levee design, yet many TMDL design choices are analogous and cost-efficiency can be considered in a similar fashion.

ANALYSIS: HOW CAN TMDL IMPLEMENTATION CHOICES PROMOTE COST-EFFICIENCY?

To consider strategies for reducing costs of TMDLs, some of the decisions made in establishing the Chesapeake Bay TMDL are examined for their potential effects on costs and benefits. A summary of the steps used to establish the TMDLs is shown in Box S1 (Supporting Information) and is referenced throughout. The program implementation decisions made at the federal and state levels are separated into two sections.

FEDERAL DECISIONS

Officials designing the Chesapeake Bay TMDLs made many decisions that managed the costs of compliance. For example, the ability to achieve reductions was considered when allocating loads to basins and jurisdictions so caps were not lower than what was achievable (Box S1—controllable loads). In addition, several technical choices used to set load caps reflected the recognition that perfect compliance with water quality criteria is physically impossible or unnecessary to protect biota. However, other decisions reflected a precautionary approach aimed at ensuring high water quality standards everywhere and most of the time. These decisions are consistent with the language of the CWA, however, risk-averse choices can be counter to approaches typically used to manage costs.

Application of Designated Uses. Among the first choices made that had significant cost implications was the spatially uniform application of designated uses for aquatic habitat that require among the highest water quality standards (Box S1). Designated uses are required by the CWA’s 303(d) provisions in order to establish both the benefits to be achieved and the appropriate water quality to achieve them. Designated use categories are generally determined by states and can require relatively low water quality standards for waters, such as “navigable”, or relatively high standards, such as “swimmable”. This flexibility, which is generally only given when non-attainability of high standards can be demonstrated, means that states can both protect high-quality resources and loosen standards in nonresponsive areas, in order to manage costs.

In the Chesapeake Bay, the highest water quality standards were initially applied uniformly to all areas of the Bay. Regulators would argue, with good reason, that this approach is needed in tidal waters that are connected by complex flows that allow nutrients and sediments to move among connected waterbodies. Therefore, compliance in one water body often relies on compliance in adjacent or nearby waterbodies, and various habitats of the Bay (e.g., deep and shallow water) must be considered simultaneously to protect a variety of aquatic species at different life stages.

Yet, allowing environmental goals to vary spatially has the potential to generate the same or similar benefits at lower cost. Such trade-offs are made routinely. For example, our legal and social systems generally accept the degradation of forests in proximity to suburban development in order to meet our competing need for housing. But, can similar trade-offs be made in waterbodies?

While multiple types of habitat areas are needed to support the food web, estuaries are not without spatial differences in terms of habitat importance. Both physical conditions within portions of the estuary (flushing time, depth, upwelling zones, etc.) and characteristics of the local watershed (e.g., dominance of agricultural or urban land uses) influence whether subareas of the Bay are likely to provide high-quality habitat. These
physical limitations on habitat quality can generate disproportionately high costs of compliance. In recognition of such constraints, the Bay TMDL includes variances for some regions that are unlikely to achieve water quality goals (ref 8, pp 3-16 and 3-17). Variances allow monitored water quality to exceed compliance criteria to manage costs, but, because they do not explicitly set an alternative goal for the waterbody, they may not eliminate ineffective efforts.

Choosing to Manage for Average or Extreme Conditions. To achieve the designated uses, the TMDLs were designed to improve several specific water quality criteria considered necessary for aquatic habitat: dissolved oxygen (DO), algal blooms (chlorophyll-α, ref 8 pp 3-9 to 3-11), and water clarity in areas conducive to seagrass growth (light extinction coefficient, $K_d$, in water ≤2 m deep, ref 8 p 3-12). By using the suite of Chesapeake Bay models and historic monitoring data, the ability to attain the desired environmental targets was evaluated by location as a function of nitrogen, phosphorus, and sediment loads from the watershed (further described in Box S1).

What these water quality standards do not readily convey is how the choice of where to set the cap on nutrients and sediments to achieve these criteria requires choosing how risk-averse to be. Due to natural variation in the physical conditions previously described or in weather patterns, less desirable water quality conditions (i.e., high turbidity) are inevitable, but are not necessarily a problem for living resources. Therefore, policy makers must decide how much flexibility in water quality compliance should be allowed by location or under extremes of weather.

Language in the CWA attempts to provide guidance on acceptable risk by specifying that perturbations to the system should be temporary:

*Maintenance of such integrity requires that any changes in the environment resulting in a physical, chemical or biological change in a pristine water body be of a temporary nature, such that by natural processes, within a few hours, days or weeks, the aquatic ecosystem will return to a state functionally identical to the original.

To implement this concept, EPA uses “critical conditions” to set water quality standards in TMDLs. Defining critical conditions that protect habitat thus requires considering that estuarine organisms are adapted to the naturally high variability of estuaries. Otherwise, if such tolerance is not considered, regulators may require a level of water quality that is not necessary for achieving benefits.

The ecological literature includes substantial research into whether isolated events or extreme stresses degrade habitat. Although the answer is not simple, a general principle has emerged that organisms in highly variable systems such as estuaries are generally adapted to infrequent extreme conditions (within certain bounds) but may be more at risk from chronic stress. Thus, reducing persistence of frequent stresses may be a more appropriate focus of restoration efforts than preventing extreme stresses.

Given this understanding, Bay regulators have made two major adjustments to the TMDL. First, they judge compliance with water quality standards in a way that allows water quality criteria to be exceeded in some portion of area or time. Second, they have used high-flow, but not extreme conditions, to judge whether water quality goals are likely to be achieved. These decisions are, roughly speaking, analogous to the engineering choice of how high to build a levee (e.g., able to withstand the 50-year storm), except, they also include considerations of areal extent and duration of stressful conditions to better capture aquatic habitat needs.

The tool for deciding what areal extent or frequency of water quality exceedances is acceptable is the cumulative frequency diagram (CFD). The CFD is a statistical approach in which water quality measurements are interpolated to estimate the areal extent of exceedences, and multiple sampling periods are used to evaluate frequency of exceedences. Two different versions of the CFD have been developed for different habitats: one for open water and deep channel and another for deep water. The open water and deep channel CFD allows, roughly speaking, 10% of violations in the combination of area and time, but only in prescribed combinations. For example, no instances of the entire area in violation at once are allowed. This acceptable violations rate was essentially a combination of best professional judgment and physical feasibility since data appropriate to judge what level of violation can be tolerated by living organisms are not well-established.

In contrast, the deep water CFD curve (defined as the area between the well-mixed surface waters and the well-mixed bottom waters) is based on “biological criteria”, meaning that it has been adjusted to reflect the extent and duration of stresses that can be tolerated by a healthy benthic community. The area under the curve is approximately 20% of area and time, but it does not mean that 20% of area or time is allowed to exceed the water quality criteria. Rather, a single month’s event covering 56% of the Bay would result in a violation as would areal violations of any size occurring in more than 75% of the months.

The approach for the deep water habitat of using ecological information to set standards provides better information for a cost-efficiency analysis because it is clearer how the TMDL level contributes to a beneficial outcome. While a healthy benthic population is not exactly an outcome that people know they value (and the measurement of the metric used is not without some methodological concerns), it can be related to opportunities for providing food for diverse fish and is widely believed to reflect the overall ecological health and resilience of the system, which has been demonstrated to be valued.

By establishing an acceptable “failure rate” (i.e., allowing some violations of water quality criteria in area and/or time), the regulators have relaxed what might otherwise be an extremely costly requirement to avoid all water quality violations anywhere at any time. However, we do not have information to judge whether the more restrictive 10% curve for open water and deep channel habitats is cost-effective because it has not been empirically connected to a beneficial outcome. If fish were readily able to withstand violations of DO criteria 20% of the time, but we are setting the TMDL to limit violations to 10% of the time, then we are raising the costs without creating commensurate benefits. Under uncertainty, regulators choose to be precautionary. However, the value of being precautionary could be made more clear, and trade-offs made more explicit in future TMDLs, by specifically linking the spatial and temporal aspects of water quality requirements to socially desirable outcomes (e.g., 11, 27).

Another technical choice that set acceptable failure rates was the selection of the hydrologic flow conditions used to project outcomes for chlorophyll-α and DO concentrations under alternative caps. Because high-flow conditions had been demonstrated to generate some of the lowest DO conditions,
TMDL designers chose a historic period of elevated flows to manage, not for the average condition, but for more stressful conditions. The modelers initially used the 3-year period of the highest flow for setting DO standards, which they stated corresponded to a 20-year event, but later deemed this an “extreme condition” and switched to the second-highest 3-year period during the period 1991−2000,8 which was estimated to correspond to a 10-year event frequency. This choice to forego using the highest flow period undoubtedly increased the acceptable loads, thereby reducing costs. Yet, similar to the CFD criterion for open water, this information would need to be translated into some kind of ecological failure rate to be able to judge the cost-effectiveness of this choice.

**Considering Marginal Costs of Abatement.** Economists use economic efficiency analysis to identify the level of pollution reduction at which the cost being imposed on polluters is justified in terms of the gains to society (further discussed in 28). If we had adequate information, the economic efficiency of the specific caps could be reduced to an analysis of whether the marginal cost, or the cost of the last required unit of nutrient or sediment reduction, is equal to the marginal return, or change in social value of that unit of reduction.29 The efficiency analysis would require summarizing all the costs of achieving a given level of nutrient and sediment reduction and all the benefits of cleaner water in terms of increases in the broad array of use and nonuse benefits, including values to future generations. We would also need to understand how small changes in nutrient reductions would generate changes in benefits. For example, we would need to know relationships such as how the risk of a harmful algal bloom decreased for every 2% reduction in nitrogen. Unfortunately, natural systems are not so predictable and the benefits of the TMDL are difficult to unambiguously quantify. The benefits may not arrive incrementally but, instead, may depend on achieving a given threshold of nutrient and sediment reduction.30

As an alternative to this explicit efficiency analysis, simplified methods can be used to suggest whether costs and benefits are balanced. Benefits can be estimated with benefit indicators that can be linked to multiple positive outcomes. Costs can be estimated for proposed nutrient and sediment reduction practices. However, to understand the trade-offs of adjusting caps up or down, both benefits and costs need to be represented in terms of their rate of change per level of reduction required. For cost data, the rate of change depends on the availability of implementation sites and the efficiency and cost by location.

To set the Bay TMDLs, regulators used water quality parameters as proxies for benefits and relied most heavily on the binding constraint of achieving DO standards in the Bay mainstem and tidal Potomac. Figure 1 shows a cumulative benefits curve that reflects how a benefit indicator, the number of waterbodies (referred to as segments) expected to violate the DO requirements (the proxy for benefits) and the curve flattens beyond that reduction effort, meaning that further reductions in nutrients do not create benefits.

![Figure 1. Effort–attainment curve: Number of DO criteria violations as a function of nitrogen loads. (All values in millions of pounds.) Source: ref 8, pp 6–29. The graph shows the correspondence between reduced nitrogen inputs (x axis) and reductions in the number of tributaries exceeding the dissolved oxygen (DO) criterion (y axis). The TMDLs chosen for N and P correspond to virtually no tributary exceedences of the DO requirements (the proxy for benefits) and the curve flattens beyond that reduction effort, meaning that further reductions in nutrients do not create benefits.](image-url)
loads would be expected to have a disproportionate cost savings. The marginal cost curve for N reduction based on partial cost estimates.\textsuperscript{31–33} Theory suggests that the curve should have a convex shape where costs per unit effort initially decline due to economies of scale (dashed line) and then increase with more effort. Data suggest that the TMDL target load for N, which requires roughly 75\% of the E3 (maximum effort) scenario, occurs where marginal costs are high, meaning that the rate of change in cost per pound of nutrient is rising rapidly for every additional pound of nutrient reduction required. Therefore, even a small reduction in the TMDL target loads would be expected to have a disproportionate cost savings.

Figure 2. Expected shape of a marginal cost curve reflecting how choice of TMDLs affects costs. The solid curve represents the general shape of the marginal cost curve for N reduction based on partial cost estimates.\textsuperscript{31–33} Theory suggests that the curve should have a convex shape where costs per unit effort initially decline due to economies of scale (dashed line) and then increase with more effort. Data suggest that the TMDL target load for N, which requires roughly 75\% of the E3 (maximum effort) scenario, occurs where marginal costs are high, meaning that the rate of change in cost per pound of nutrient is rising rapidly for every additional pound of nutrient reduction required. Therefore, even a small reduction in the TMDL target loads would be expected to have a disproportionate cost savings.

The curve is built by assuming that the lowest cost practices are used first and then more expensive technologies are added as necessary to increase effort. Costs of any given practice tend to increase as more implementation is required. For example, some agricultural producers will be willing and able to implement best management practices (BMPs) at low levels of compensation, but some producers will have high opportunity costs or personal preferences that will require high compensation to make BMP implementation acceptable (evidence is in 34).

At the level of effort associated with the current TMDLs, marginal cost is relatively high and the curve has a steep slope (Figure 2) primarily due to the high costs of upgrades at some wastewater treatment plants and the difficulty of finding sufficient opportunities to apply low-cost NPS reduction practices. A steep slope on the marginal cost curve suggests that even a small increase or reduction in the TMDLs can have a large effect on costs of compliance.

Figure 2 suggests that the closer the TMDLs are set to the E3 scenario (E3 = Everyone does Everything, Everywhere, see Box S1), the more benefits per unit of nutrient reduction are needed to justify that choice. Yet, Figure 1 shows that benefits (as judged only by DO) are changing very little in the vicinity of the cap. This simplified analysis, using only one relationship of effect of nitrogen on DO criteria exceedences, suggests that allowing a small increase in DO violations would disproportionately lower costs of TMDL compliance. This analysis is not a complete assessment and is not an explicit marginal cost analysis. However, it indicates how future TMDLs can use marginal costs and marginal benefits to consider the economic efficiency of specific caps.

Even though the marginal cost curve is generalized, evidence from many systems suggests that costs rise rapidly as all available effort is expended—an effect referred to as the law of diminishing marginal returns. What Figure 2 does not show is that in the locations identified by the CBP as the most effective for increasing dissolved oxygen in the mainstem and lower Potomac, wastewater sources were asked to control 90\% of N and 97\% of P (Box S1) indicating very high marginal costs. Whether these high marginal costs create similarly high marginal benefits is unclear. For example, at the Blue Plains wastewater treatment plant (WWTP) in Washington, DC, the upgrades that will take the plant from 5\% to 3\% total nitrogen in effluent have been estimated to reduce less than 5\% of the total nitrogen flux to the Potomac River (evidence in 35) at a cost of at least $1.5 billion.\textsuperscript{36,37}

\section*{STATE LEVEL DECISIONS}

Some states are relying heavily on PS to NPS water quality trading in their WIPs (PA and VA) to achieve TMDL goals while containing costs and all are considering the use of government payments to fund reduction activities. In addition, some type of offsets (in or out of markets) will be needed to offset new emissions due to future growth and land use changes. States’ initial rules for markets promote trading within local watersheds or counties, however, regional and interstate trading is being considered or allowed if the trade would not cause violations of local TMDL caps (MD) or local water quality standards (PA).\textsuperscript{38,39} How such compliance will be verified is yet to be determined.

According to the phase I and II WIPs of MD, VA, and PA, some decisions made by states that affect the potential for achieving cost-effectiveness in trading, offset, or NPS payment programs include the following: (1) allocating load reductions to small geographic areas; (2) establishing baseline requirements for unregulated source sectors based on cost-blind criteria (e.g., caps set per farm acre regardless of ability to reduce nutrient runoff); and (3) setting technology require-
ments rather than performance requirements. Baseline requirements are the minimum activities or conditions that must be met for a landowner to be eligible to sell credits in a water quality trading market and thus primarily affect water quality markets.

**How Basin Size Affects Market Potential and Cost-Effectiveness of Offsets.** In designing the TMDLs, regulators must decide where water quality compliance will be judged which, in turn, affects who can trade with whom and how offsets are used and funded. In the phase I WIPs, the states, under the guidance of the CBP, distributed the Basin-Jurisdiction caps to 92 segment-sheds which are the geographic divisions that are used to report compliance with 303(d) regulations; segment-sheds range in size from roughly 10 to 10,000 square miles (Figure 3). In the phase II WIPs for MD and VA, the segment-shed allocations were intersected with county boundaries to estimate county-level requirements while Pennsylvania maintained allocations to large basin-jurisdictions.

Since 89 of the 92 segment-sheds are listed as impaired for all uses, load reductions are needed locally to comply with water quality requirements. But, a portion of reductions are being required at the sub-basin level that only protect sensitive habitat in the Bay mainstem. Therefore, the caps at the sub-basin scale are typically more stringent than required to bring a local waterway into compliance. A major exception is the James River, in which local conditions necessitated more stringent target loads. As an aside, the reductions needed to comply with local standards are further complicated by the thousands of TMDLs that have been developed to address localized concerns in MD, PA, and VA, including nutrients, sediments, metals, fecal coliform, and pH. Local TMDLs can be more or less stringent than the Bay TMDL and the more stringent TMDL takes precedence.

The reallocation of loads to small segment-sheds affects costs of compliance in two related ways: (1) it limits the spatial extent of initial trading areas, and (2) it potentially compels payment programs to spread money among counties, regardless of the cost-effectiveness of available options. The bigger the watershed in which reductions must occur, the more likely it is to include diverse (and regulated) sources of emitters, which promotes market development and cost-effective targeting of payments. Diverse types of pollution emitters (e.g., different types of wastewater treatment plants or farm operations) are needed for markets because successful trading hinges on having credit buyers and sellers with unequal costs of “producing” pollution reductions. Small trading areas could mean that a WWTP manager who would be motivated to purchase credits because of her high costs of compliance would be unable to use NPS offsets because the number of willing sellers is insufficient to supply enough credits.

Setting caps for small areas can also reduce cost-effectiveness of offset or payment programs by forcing effort to be spread out geographically, when cost-effective practices could be geographically concentrated (e.g., dairy farms). A recent study suggests that the cost savings from expanding geographic areas for offsets could be substantial. The costs associated with the least-cost scenario for meeting the TMDL in the Bay were reduced by 18% when nutrient reductions were allowed to come from anywhere in the watershed vs from each major tributary. The study does not represent the Bay TMDL, but illustrates the general rule that cost-effectiveness can often be improved by maintaining large areas for offsets or trades. The 92 segment-sheds used in the TMDL are largely modeling conventions rather than a geography that has been linked to specific restoration outcomes (e.g., improved fish recruitment) or social benefits, which suggests that distributing caps to larger basins could lower costs while maintaining benefits.

The state WIPs suggest that states are aware of the cost-efficiency loss from allocating to small basins because they allow for interbasin trading—once the reductions necessary to achieve local goals have been met. However, the TMDLs were not provided in a way that separated out the local vs regional allocation of load reductions, creating a barrier to such trading. Cost-effectiveness could be promoted by subdividing each state’s caps into the “local” portion required for each major basin’s water quality compliance and the “regional” portion that is only needed to meet Bay mainstem water quality goals. Further, giving liability for the regional reductions to the states could create savings through economies of scale.

**Using Trading Baselines to Achieve Voluntary Compliance.** A significant hurdle to achieving the TMDLs is that a major source sector—nonpoint source agriculture—has no legal requirement to achieve their allocated load reductions. Because unenforceable caps have been assigned to this sector, a variety of approaches are being used to create incentives for voluntary actions. The agricultural portions of the TMDL reductions are allocated to individual farms either as quantities or trades. This allows farmers to supply enough credits. The reductions needed to comply with nonpoint source agriculture are further complicated by the thousands of TMDLs that have been developed to address localized concerns in MD, PA, and VA, including nutrients, sediments, metals, fecal coliform, and pH. Local TMDLs can be more or less stringent than the Bay TMDL and the more stringent TMDL takes precedence.

The reallocation of loads to small segment-sheds affects costs of compliance in two related ways: (1) it limits the spatial extent of initial trading areas, and (2) it potentially compels payment programs to spread money among counties, regardless of the cost-effectiveness of available options. The bigger the watershed in which reductions must occur, the more likely it is to include diverse (and regulated) sources of emitters, which promotes market development and cost-effective targeting of payments. Diverse types of pollution emitters (e.g., different types of wastewater treatment plants or farm operations) are needed for markets because successful trading hinges on having credit buyers and sellers with unequal costs of “producing” pollution reductions. Small trading areas could mean that a WWTP manager who would be motivated to purchase credits because of her high costs of compliance would be unable to use NPS offsets because the number of willing sellers is insufficient to supply enough credits.

Setting caps for small areas can also reduce cost-effectiveness of offset or payment programs by forcing effort to be spread out geographically, when cost-effective practices could be geographically concentrated (e.g., dairy farms). A recent study suggests that the cost savings from expanding geographic areas for offsets could be substantial. The costs associated with the least-cost scenario for meeting the TMDL in the Bay were reduced by 18% when nutrient reductions were allowed to come from anywhere in the watershed vs from each major tributary. The study does not represent the Bay TMDL, but illustrates the general rule that cost-effectiveness can often be improved by maintaining large areas for offsets or trades. The 92 segment-sheds used in the TMDL are largely modeling conventions rather than a geography that has been linked to specific restoration outcomes (e.g., improved fish recruitment) or social benefits, which suggests that distributing caps to larger basins could lower costs while maintaining benefits.

The state WIPs suggest that states are aware of the cost-efficiency loss from allocating to small basins because they allow for interbasin trading—once the reductions necessary to achieve local goals have been met. However, the TMDLs were not provided in a way that separated out the local vs regional allocation of load reductions, creating a barrier to such trading. Cost-effectiveness could be promoted by subdividing each state’s caps into the “local” portion required for each major basin’s water quality compliance and the “regional” portion that is only needed to meet Bay mainstem water quality goals. Further, giving liability for the regional reductions to the states could create savings through economies of scale.

**Using Trading Baselines to Achieve Voluntary Compliance.** A significant hurdle to achieving the TMDLs is that a major source sector—nonpoint source agriculture—has no legal requirement to achieve their allocated load reductions. Because unenforceable caps have been assigned to this sector, a variety of approaches are being used to create incentives for voluntary actions. The agricultural portions of the TMDL reductions are allocated to individual farms either as quantities or trades. This allows farmers to supply enough credits. The reductions needed to comply with nonpoint source agriculture are further complicated by the thousands of TMDLs that have been developed to address localized concerns in MD, PA, and VA, including nutrients, sediments, metals, fecal coliform, and pH. Local TMDLs can be more or less stringent than the Bay TMDL and the more stringent TMDL takes precedence.

The reallocation of loads to small segment-sheds affects costs of compliance in two related ways: (1) it limits the spatial extent of initial trading areas, and (2) it potentially compels payment programs to spread money among counties, regardless of the cost-effectiveness of available options. The bigger the watershed in which reductions must occur, the more likely it is to include diverse (and regulated) sources of emitters, which promotes market development and cost-effective targeting of payments. Diverse types of pollution emitters (e.g., different types of wastewater treatment plants or farm operations) are needed for markets because successful trading hinges on having credit buyers and sellers with unequal costs of “producing” pollution reductions. Small trading areas could mean that a WWTP manager who would be motivated to purchase credits because of her high costs of compliance would be unable to use NPS offsets because the number of willing sellers is insufficient to supply enough credits.

Setting caps for small areas can also reduce cost-effectiveness of offset or payment programs by forcing effort to be spread out geographically, when cost-effective practices could be geographically concentrated (e.g., dairy farms). A recent study suggests that the cost savings from expanding geographic areas for offsets could be substantial. The costs associated with the least-cost scenario for meeting the TMDL in the Bay were reduced by 18% when nutrient reductions were allowed to come from anywhere in the watershed vs from each major tributary. The study does not represent the Bay TMDL, but illustrates the general rule that cost-effectiveness can often be improved by maintaining large areas for offsets or trades. The 92 segment-sheds used in the TMDL are largely modeling conventions rather than a geography that has been linked to specific restoration outcomes (e.g., improved fish recruitment) or social benefits, which suggests that distributing caps to larger basins could lower costs while maintaining benefits.

The state WIPs suggest that states are aware of the cost-efficiency loss from allocating to small basins because they allow for interbasin trading—once the reductions necessary to achieve local goals have been met. However, the TMDLs were not provided in a way that separated out the local vs regional allocation of load reductions, creating a barrier to such trading. Cost-effectiveness could be promoted by subdividing each state’s caps into the “local” portion required for each major basin’s water quality compliance and the “regional” portion that is only needed to meet Bay mainstem water quality goals. Further, giving liability for the regional reductions to the states could create savings through economies of scale.

**Using Trading Baselines to Achieve Voluntary Compliance.** A significant hurdle to achieving the TMDLs is that a major source sector—nonpoint source agriculture—has no legal requirement to achieve their allocated load reductions. Because unenforceable caps have been assigned to this sector, a variety of approaches are being used to create incentives for voluntary actions. The agricultural portions of the TMDL reductions are allocated to individual farms either as quantities or trades. This allows farmers to supply enough credits. The reductions needed to comply with nonpoint source agriculture are further complicated by the thousands of TMDLs that have been developed to address localized concerns in MD, PA, and VA, including nutrients, sediments, metals, fecal coliform, and pH. Local TMDLs can be more or less stringent than the Bay TMDL and the more stringent TMDL takes precedence.
Nutrient and sediment control activities that exceed the baseline can be sold in markets and the hope is that the incentive to sell credits in functioning water quality markets will be sufficient to generate compliance up to the baseline. Under a system where voluntary reductions by the NPS are necessary to achieve the TMDL, the baselines are intended to ensure that trades between the PS and NPS sectors do not effectively increase the TMDL cap. The concern is that regulated PSs that purchase credits from unregulated sources will be allowed to pollute more (i.e., increase their waste load allocation) without the NPS sector first achieving its reduction targets.

The obvious tension in setting high-effort baselines is that they interfere with a major goal of the trading—encouraging reticent farmers to adopt best management practices. Farms that have done the least may have the least incentive to enter the market, since the ability to recoup costs of implementing multiple practices may be highly uncertain. Further, in Maryland, each farm seeking to trade must meet a percentage reduction in nutrients and/or sediments based on emissions for the enclosing basin, rather than farm characteristics, before selling credits. Because this approach does not consider the cost-effectiveness of implementing particular practices on a farm, some low-cost credits will be unavailable, simply because of where they occur (see Box S2). If baselines cause farms to forego participation in the market, then the TMDL reduction allocated to the agricultural sector will not be met through trading and low-cost credits will not be available to offset high costs at point sources.

Given the low level of trading in other water quality markets, even without tough baseline requirements, the ability of voluntary markets to develop is questionable (discussed in 48, 49). An alternative mechanism for generating compliance in NPS sectors would be for states to establish enforceable pollution limits for a given sector and allow those within the sector to trade among themselves to meet the sector’s baseline. Known as a group cap, this approach has been successfully used to allow groups of emitters to find lower cost options to meeting a cap, rather than requiring specific reductions at every site.

While group caps have most commonly been used with point sources (e.g., S0, S1), a type of group cap is an option for agricultural producers in a different program in North Carolina. As part of the TMDL implementation in the Neuse basin, producers are required to reduce nitrogen runoff by 30% and maintain that level of reduction (Rule .0236). To achieve that goal, producers are required (under threat of civil or criminal penalties) to either individually implement a set of best management practices or join an association that will develop and implement a “collective local strategy” that will, presumably, reduce an individual’s costs of compliance.

Allowing a group to share responsibility for a reduction is cost-efficient because it allows those with lower costs of compliance to do more than those with higher costs. The group can establish mechanisms to compensate those who do more, including the use of credits, thereby creating actual or defacto trading. Without state regulation, the group cap used in North Carolina would not be enforceable suggesting that one of the most straightforward approaches to improving TMDL cost-effectiveness would be to implement laws limiting nutrient emissions from agricultural producers. However, since this may not be practical, an alternative is to use flexible incentives to promote voluntary actions by individuals or groups. For example, allowing a group of farms to be collectively certified for trading by achieving a certain level of pollution reduction could reduce the costs of achieving the baseline and the cost of available credits. Further, groups might be motivated to achieve local benefits since collective action on a small tributary might be the best way to ensure local benefits (e.g., Conewago Creek Initiative).

**Performance Standards Promote Innovation and Cost Savings.** One of the earliest cases of water quality trading—in the Tar-Pamlico estuary of North Carolina—demonstrates the effectiveness of using group performance goals to achieve PS pollution reductions at lower costs. In phase I of the NC Program, the PS sector was given a bubble permit (group compliance requirement), to cap pollution emissions into the estuary, rather than requiring technology or performance standards at individual plants. Emitters were able to achieve new efficiencies when given the additional flexibility offered by the bubble permis53,54 and they reduced costs of compliance through plant-specific innovation and defacto trading.

This case demonstrated that specifying what is to be achieved—i.e., a performance goal—is preferable to specifying how to achieve it—i.e., requiring specific technologies—because the former spurs innovation. A group of farms might decide that jointly creating a wetland is their preferred method of compliance, but with technology requirements, they will not be motivated to try new approaches because they have already been told how to comply. A challenge to using performance-based incentives is the high cost of measuring outcomes. Costs are typically managed by modeling practice performance, rather than monitoring it, but further work is needed to understand whether (1) models can effectively drive innovation by incorporating the details that affect performance, or (2) whether technological innovations might reduce monitoring costs.

### DISCUSSION AND CONCLUSIONS

The TMDLs, if successful, will create a multitude of public benefits. However, few benefits will be realized if regulated parties cannot or will not comply. This analysis considered what choices in TMDL program design promote compliance by enhancing cost-efficiency, including whether accepting more risk of adverse outcomes might be appropriate for reducing costs. The CBP designed the TMDLs by examining benefits and capacity to reduce nutrients and sediments but did not explicitly consider costs until the implementation phase. This sequence of analysis limited the ability to explicitly compare costs and benefits or risks in the design phase, although some program choices, nonetheless, managed costs. The Chesapeake Bay TMDL is typical of most TMDLs, because the design process promotes the establishment of the maximum pollution load and source sector allocations independently of costs. Costs typically enter the process in an implicit manner through decisions about how compliance will be calculated or as part of the implementation strategy, rather than used directly to refine the caps or allocations.

The CFD method for assessing compliance was used here to illustrate how strategies for reducing regulatory costs often translate into accepting more uncertainty of beneficial outcomes. The CFD is used to set caps in a manner that incorporates natural system variability and, in some cases, tolerance of aquatic organisms for that variability. However, uncertainty about effects of temporary water quality exceedences in some habitats led to uneven use of this cost-saving approach. In the portion of the Bay where TMDLs were set...
based on data (the deep water criteria are linked to benthic community condition and system health), the empirically derived standards are less stringent than the standards where best professional judgment was used (i.e., the 10% curve used for open water and deep channel). The data-derived standards are certain to be cheaper to meet than the more stringent ones based on scientific judgment, but lowering costs in data-poor areas means accepting a higher risk of adverse outcomes. Because being precautionary is expensive, such decision-making will be served by developing measures of beneficial outcomes for different TMDLs levels (e.g., reasonably proxies of benefits based on biotic responses) to weigh whether extra precaution warrants the cost.

Some of the largest cost savings for achieving the TMDL could come from ensuring more participation by agricultural producers. Yet, expecting markets to create incentives for unregulated entities to voluntarily reduce pollution is an untested approach and will, at a minimum, require aligning regulatory requirements with market needs. In addition, payment or offset programs will be more cost-efficient if they can choose from among the most cost-effective options that produce equivalent benefits. However, the allocation of loads reductions that are only needed to protect the Bay mainstem to many (sometimes small) geographic areas is at odds with both efficient use of markets and payments. Separating the local reduction need from the mainstem need and allocating reductions to the largest geographic boundaries that are effective for ensuring benefits could provide the flexibility necessary to achieve cost savings across a range of programs.

Further, some of the approaches used to set trading baselines are likely to leave low-cost nutrient and sediment reductions unimplemented. Baselines calculated from basin-wide rather than farm-specific characteristics (e.g., in MD) will prevent some producers from creating competitively priced credits and, in general, baselines may create barriers to entry for farms that have done the least. Because market development is uncertain, alternative approaches to allocating agricultural reductions, legally limiting emissions, or using new institutional arrangements may be needed to successfully engage the agricultural sector.

Overall, the analysis suggests that cost-effectiveness may be promoted by evaluating how decisions change costs and benefits during the TMDL design and implementation. Managing the costs of TMDLs is likely to require accepting a somewhat higher risk of adverse ecological outcomes (i.e., building the 50-year levee and not the 100-year levee). However, given the difficulty of paying for and enforcing the TMDLs, the risk may not be higher than the expected outcomes of more precautionary TMDLs.

**ASSOCIATED CONTENT**

[1] Supporting Information

Text Box S1 summarizes the steps used by the CBP to establish the TMDLs; Text Box S2 provides an example of how trading baseline requirements may prevent low-cost credits from reaching the marketplace; Table S1 shows current nutrient and sediment control requirements for farms by state. This material is available free of charge via the Internet at http://pubs.acs.org.

**AUTHOR INFORMATION**

*Corresponding Author*

E-mail: wainger@umces.edu.

**Notes**

The authors declare no competing financial interest.

**ACKNOWLEDGMENTS**

I thank those who provided helpful comments on this paper, including Andrew Almeter, Jim Boyd, Walter Boynton, Dennis M. King, Ben Koch, Jay Messer, J.B. Ruhl, Rob Wolcott, and two anonymous reviewers. Partial funding and support for this research was provided by the U.S. Environmental Protection Agency’s Ecosystems Research Program. The opinions and ideas expressed in this paper are the author’s and do not necessarily reflect the views of EPA.

**REFERENCES**

(1) Blankenship, K. Local officials worry that TMDL actions are much too costly. Chesapeake Bay J. 2011, 21 (5), 1.


(13) Funderburk, S. L.; Jordon, S. J.; Mihursky, J. A.; Riley, D. Habitat Requirements for Chesapeake Bay Living Resources; Habitat Objectives Workshop, Living Resources Subcommittee, Chesapeake Bay Program; Chesapeake Research Consortium, Inc.: Solomons, MD, 1991.

(19) U.S. EPA. Ambient Water Quality Criteria for Dissolved Oxygen, Water Clarity and Chlorophyll a for the Chesapeake Bay and Its Tidal Tributaries; EPA 903-R-03-002; U.S. Environmental Protection Agency Region 3 Chesapeake Bay Program Office: Annapolis, MD, 2003.
(21) Sherk, G. Integrated Analysis Coordinator, EPA/Chesapeake Bay Program Office, Annapolis MD, Personal communication 2011.
(37) Maryland Governor’s Office. Governor O’Malley seeks federal funding for clean water, Chesapeake Bay. http://www.gov.state.md.us/pressrelease/100311f.asp.
(38) PA DEP. Final Trading of Nutrient and Sediment Reduction Credits - Policy and Guidelines (Revisions to the Interim Final Trading of Nutrient and Sediment Reduction Credits - Policy and Guidelines); 392-0900-001; Pennsylvania Department of Environmental Protection: Harrisburg, PA, 2007.
(45) Ruth, M.; Gasper, R.; Bagley, G.; Selman, M.; Marshall, L. Multiple Ecosystem Markets in Maryland: Quantifying the Carbon Benefits Associated with Nutrient Trading; Center for Integrative Environmental Research (CIER), University of Maryland, 2010.
(46) Chesapeake Bay Environmental Markets Team. Considerations for Baseline Eligibility Associated to Water Quality Credit Programs Implemented under a Chesapeake Bay Total Maximum Daily Load (TMDL); USDA Office of Environmental Markets, 2010.