JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

USING MULTIPLE WATERSHED MODELS TO PREDICT WATER, NITROGEN, AND PHOSPHORUS DISCHARGES TO THE PATUXENT ESTUARY[1]

Kathleen M.B. Boomer, Donald E. Weller, Thomas E. Jordan, Lewis Linker, Zhi-Jun Liu, James Reilly, Gary Shenk, and Alexey A. Voinov[2]

ABSTRACT: We analyzed an ensemble of watershed models that predict flow, nitrogen, and phosphorus discharges. The models differed in scope and complexity and used different input data, but all had been applied to evaluate human impacts on discharges to the Patuxent River or to the Chesapeake Bay. We compared predictions to observations of average annual, annual time series, and monthly discharge leaving three basins. No model consistently matched observed discharges better than the others, and predictions differed as much as 150% for every basin. Models that agreed best with the observations in one basin often were among the worst models for another material or basin. Combining model predictions into a model average improved overall reliability in matching observations, and the range of predictions helped describe uncertainty. The model average was not the closest to the observed discharge for every material, basin, and time frame, but the model average had the highest Nash–Sutcliffe performance across all combinations. Consistently poor performance in predicting phosphorus loads suggests that none of the models capture major controls. Differences among model predictions came from differences in model structures, input data, and the time period considered, and also to errors in the observed discharge. Ensemble watershed modeling helped identify research needs and quantify the uncertainties that should be considered when using the models in management decisions.

(KEY TERMS: watersheds; watershed management; nonpoint source pollution; simulation; hydrological modeling; ensemble modeling; model comparison; model average; land use; model structure; model performance.)

Boomer, Kathleen M.B., Donald E. Weller, Thomas E. Jordan, Lewis Linker, Zhi-Jun Liu, James Reilly, Gary Shenk, and Alexey A. Voinov, 2012. Using Multiple Watershed Models to Predict Water, Nitrogen, and Phosphorus Discharges to the Patuxent Estuary. *Journal of the American Water Resources Association* (JAWRA) 1-25. DOI: 10.1111/j.1752-1688.2012.00689.x

[1]Paper No. JAWRA-11-0062-P of the *Journal of the American Water Resources Association* (JAWRA). Received May 9, 2011; accepted July 31, 2012. **Discussions are open until six months from print publication.**

[2]Respectively, Ecologist (Boomer), Senior Ecologists (Weller, Jordan), Smithsonian Environmental Research Center, 647 Contees Wharf Road, Edgewater, Maryland 21037-0028; Modeling Coordinator (Linker) and Integrated Analysis Coordinator (Shenk), U.S. Environmental Protection Agency Chesapeake Bay Program, Annapolis, Maryland 21403; Associate Professor (Liu), Department of Geography, University of North Carolina, Greensboro, North Carolina 27402-6170; Planner (Reilly), Maryland Department of Planning, Baltimore, Maryland 21201 [Reilly now at Reilly Consulting, Lafayette Hill, Pennsylvania 19444]; and Associate Professor (Voinov), The Gund Institute for Ecological Economics, University of Vermont, Burlington, Vermont 05405 [Voinov now at International Institute for Geo-information Science and Earth Observation, Enschede, The Netherlands] (E-Mail/Boomer: boomerk@si.edu).

INTRODUCTION

We analyzed an ensemble of watershed models that can all predict water, nitrogen, and phosphorus discharges to the Patuxent River, which is a subestuary of the Chesapeake Bay and the sixth largest tributary to the Bay. Watershed models are essential tools for linking nonpoint sources with surface water pollution and for predicting the effects of management efforts on water quality (Miller *et al.*, 2004). In the Patuxent, the larger Chesapeake Bay, and many other estuaries worldwide, efforts to reduce watershed nutrient discharges and restore degraded estuaries rely on watershed models to plan management actions and to help enforce water quality regulations (Boesch, 2002; Smith *et al.*, 2006; USEPA, 2010a; NRC, 2011).

Despite their critical role in management programs, watershed models often perform poorly because of imperfect knowledge of hydrological processes, biogeochemical processes, and human activities (Radcliffe *et al.*, 2009). The resulting uncertainty in model predictions is difficult to quantify and interpret, and that can undermine scientific and public confidence in model predictions (Radcliffe *et al.*, 2009). In addition, most applications rely on a single model, and this provides no opportunity to evaluate the structural uncertainty inherent in choosing the conceptual and mathematical underpinnings of the model.

Analyzing a set of two or more models of a watershed (ensemble modeling) can help objectively evaluate model skill and the uncertainty in model predictions (Beven, 2007). Ensemble modeling is especially appropriate for environmental systems in which dynamic processes operate over a range of temporal and spatial scales (Clark, 2007). Each model represents a different set of hypotheses describing the dominant landscape processes affecting watershed discharge. Comparing models contrasts different hypotheses about system drivers (Bloschl, 2006) and helps to identify how models can be refined and improved (e.g., McIntyre *et al.*, 2005; Dezetter *et al.*, 2008). The multi-model approach has gained widespread acceptance in other disciplines, including financial forecasting, socioeconomics, weather and climate, and wildlife management (e.g., Givens, 1999; Koop and Tole, 2004; Gneiting and Raftery, 2005; Phillips and Gleckler, 2006). Initial applications in watershed modeling reported that multi-model syntheses provide better estimates and a stronger basis for informing watershed management decisions than a single model (Vrugt and Robinson, 2007; Hsu *et al.*, 2009; Huisman *et al.*, 2009), but ensemble watershed modeling has not been used much outside of the European Union, possibly because of the higher cost of implementing multiple models and the reluctance of the watershed modeling community to embrace uncertainty analyses (Pappenberger and Beven, 2006).

The predictions from a set of models can also be combined into a model average, which can work better than relying on a single "best" model for supporting management decisions, especially when there is not enough information to identify the best model or when the data do not favor a particular model (Kadane and Lazar, 2004). Because of the reluctance to examine multiple models, the applications of model averaging in watershed analysis have also been limited (Vrugt and Robinson, 2007), but would likely advance knowledge of terrestrial hydrologic processes (Sivakumar, 2008) and improve model accuracy (Duan *et al.*, 2007). The range of predictions for a defined endpoint provides an initial quantitative estimate of the overall uncertainty in the system processes.

In our study of Patuxent watershed models, we evaluated the abilities of the models to predict observed water and nutrient discharge data (model skill) and to estimate quantities that are important for management decisions, but not measured. Instead of seeking a best model, we focused on how an ensemble of models and model averaging can improve predictions of watershed discharges, help quantify model uncertainty, and increase understanding of terrestrial-aquatic linkages and the impacts of human activities on aquatic ecosystems.

We analyzed the watershed models as they were published because the models were not amenable to further standardization and because the published results have already been used to draw scientific inferences and to guide management decisions. Our analysis differed from a common approach to rigorous model comparison, which reruns a set of similar models with standardized inputs and calibration data, and then compares results for the same outputs and time periods (Breuer *et al.*, 2009). Such an effort focuses on quantifying how model structure affects model output with everything else controlled. This approach is very discerning from a modeling perspective. It is not always feasible, it ignores important differences arising from user choices and constraints during model implementation, and it may not reveal the full contrast among models that is needed to understand their management implications. As we analyzed the models as published, some of the differences among models that we report are due to differences in input data, calibration data, or time period considered rather than to differences in model structure.

METHODS

Overview

We identified six watershed models that had been applied to the Patuxent River watershed or that predicted loads from the Patuxent watershed as a part of modeling the larger Chesapeake Bay basin. The models were published in peer reviewed literature or actively used in land use planning. Three models had more than one published version, so altogether there were 10 implementations of the six models (Table 1). The scope and complexity of the models vary widely, but all the models are intended to quantify how natural factors and anthropogenic stressors influence total nitrogen (TN) and total phosphorus (TP) discharges from the watershed.

We compared model predicted outputs for selected endpoints, where "endpoint" refers to the estimated discharge for a combination of material, basin, and time frame. There were three output materials (water, TN, and TP), four prediction basins (Laurel, Western Branch, Bowie, and the entire Patuxent watershed), and three time frames (average annual, annual time series, and monthly time series). We examined the average annual predictions because some of the models predict only average annual loads and because management decisions are often based on annual average loads to factor out the effects of extreme weather or other unusual events that may affect a single year. We also examined predictions of annual and monthly time series loads to quantify how the models perform in representing temporal variability in water and nutrient discharges. The analysis of monthly time series generally confirmed to the lessons learned from the annual time series, so the descriptions of the monthly analysis is reported only in the Supporting Information.

Two sets of endpoints had no measurements available to evaluate model performance: predictions of average annual and annual time series discharges from the entire Patuxent watershed and the predicted proportions of nonpoint TN and TP discharges allocated to agriculture and to developed land for all four watersheds. Making predictions for unmeasured endpoints is a major objective of watershed modeling, and these two endpoints are good examples of unmeasured endpoints that environmental decision makers need to estimate. Understanding the impact of the Patuxent watershed on the Patuxent and Chesapeake estuaries demands estimates of the total nutrient

TABLE 1. Ten Implementations of Six Watershed Models.

Model	Abbreviation	Year*	General Description	References
Models predicting average annual TN and TP				
Maryland Department of Planning Assessment and Accounting System	MDP90 MDP97	1990 1997	Export coefficient model	MOP (1993, 1995); Maryland Department of Planning (MDP), Maryland Department of the Environment (MDE), and Maryland Department of Natural Resources (DNR) (2007); Tassone *et al.* (1998)
USGS SPAtially Referenced Regressions On Watershed attributes	SPARROW87 SPARROW92 SPARROW97	1987 1992 1997	Nonlinear statistical model	Smith *et al.* (1997), Preston and Brakebill (1999)
Models producing time series predictions of flow and nutrients				
Smithsonian Environmental Research Center statistical model	SERC	1997-1999	Linear statistical model	Jordan *et al.* (2003), Weller *et al.* (2003)
Smithsonian Environmental Research Center Integrated Landscape Model	SERCLM	1984-1999	Spatially lumped simulation model	Liu and Weller (2008), Liu *et al.* (2008)
Chesapeake Bay Program Hydrologic Simulation Program	CBP4 CBP5	1984-2000 1984-2000	Spatially lumped simulation model	Donigian *et al.* (1994), Linker *et al.* (2000)
Patuxent Landscape Model	PLM	1986-1993	Spatially distributed simulation model	Costanza *et al.* (2002)

*For MDP and SPARROW, year refers to the date of the geographic and load data used to predict average annual loads. For the other models, the range of years represents the period over which the time series of flow and nutrient discharges were simulated.

loads from the watershed, and developing strategies for reducing nutrient loads demands information on where the loads originate within the watershed. For every endpoint considered, we present the model average prediction, calculated as the simple average of estimates from the models capable of predicting that endpoint. We estimated the skill of the model average as if it were another model in the ensemble.

The following provides more detailed descriptions of the study area, the streamflow and nutrient data, the models themselves, and the methods of analysis.

Study Area

The 2,300 km^2 Patuxent River watershed is entirely in Maryland, U.S., mainly in the Coastal Plain physiographic province (72%) with the reminder in the Piedmont (Figure 1). The upper watershed is located between the cities of Washington, D.C. and Baltimore, Maryland. The 2001 National Land Cover Database (Homer *et al.*, 2004, 2007) indicated that 17% of the watershed area was developed, 11% was cropland, 24% was grassland, and 46% was forest. Two reservoirs in the northern part of the watershed are managed for flood control and drinking water supply. From 1997 to 1999, 150 million cubic meters (Mm^3) of water (about 9% of the 1,664 Mm^3 freshwater flow to the estuary) was withdrawn for water supply purposes (Jordan *et al.*, 2003), but a roughly equal amount (142 Mm^3) was returned to the river network in the discharges from 8 major (>500,000 gallons/day) and 17 minor wastewater treatment facilities (Jordan *et al.*, 2003; Weller *et al.*, 2003).

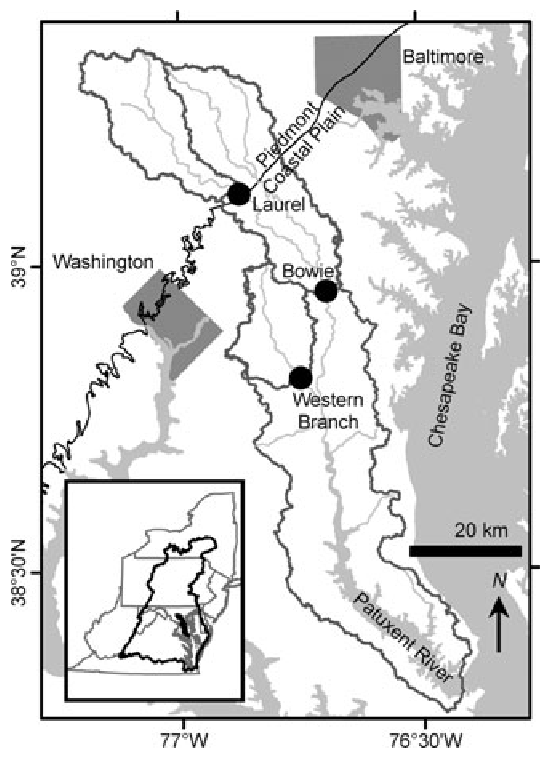

FIGURE 1. The Patuxent River Watershed near the Cities of Washington, D.C. and Baltimore, Maryland. Three U.S. Geological Survey monitoring stations (dots) provided measured water, TN, and TP loads for testing model predictions for the associated watersheds (thick lines). The fall line separates the region into the Piedmont and Coastal Plain physiographic provinces. The inset shows the Patuxent watershed (black) within the Chesapeake Bay watershed (thicker line) and the mid-Atlantic states (thinner lines).

Streamflow and Nutrient Measurements

To test the models, we used measurements of water flow, nitrogen loads, and phosphorus loads for three U.S. Geological Survey (USGS) gauging sites (Figure 1) with continuous flow data and routinely measured TN and TP concentrations (Langland *et al.*, 1995; Darrell *et al.*, 1998; Michael Langland, USGS, August 24, 2009, personal communication; Langland *et al.*, 1999; USGS, 2011a). We used the available measurements for 1984-2000 from three basins: Laurel (gauge 01592500, 342 km^2 basin area), Western Branch (01594526, 232 km^2), and Bowie (01594440, 907 km^2). The Laurel basin is in the Piedmont physiographic province and contains the two water supply reservoirs, and the Western branch basin is in the Coastal Plain. The Bowie basin is 70% Piedmont and includes the Laurel basin (Weller *et al.*, 2003). The USGS applied the ESTIMATOR model (Cohn *et al.*, 1989) to estimate monthly annual, and average annual nutrient loads from continuous discharge data and nutrient concentrations in water quality samples. For the Laurel and Bowie basins, the USGS also provided 95% prediction intervals for the annual and monthly nutrient load estimates produced using ESTIMATOR.

Watershed Models

This section provides brief descriptions of the six watershed models in our ensemble, along with references to more detailed information. The models are presented in rough order of increasing complexity: an uncalibrated export coefficient model, two statistical models, two spatially lumped simulation models, and a spatially distributed simulation model. All the models estimate point source contributions from wastewater treatment plants from discharge monitoring reports or from permitted discharges when monitoring reports are not available.

Maryland Department of Planning Assessment and Accounting. The Maryland Department of Planning (MDP) model is an export-coefficient model for the state of Maryland (MOP, 1993, 1995; MDP, 2007; Tassone *et al.*, 1998). Export coefficient models assume a constant TN or TP yield (kg per ha per year, the "export coefficient") for each land use class and estimate watershed load by summing the products of land use area and export coefficient across all land use classes. The MDP model has been used throughout Maryland to assist local town and county planners in developing growth strategies that minimize impacts to surface water bodies. The export coefficients were derived from the Chesapeake Bay Program's HSPF Model, Version 4 (CBP4) (described below) for each land use in the Department of Planning multi-year land use database and adjusted by the average amount of impervious surface area associated with different developed land densities. The model does not predict water discharge, and the model was not calibrated to match observed TN or TP loads. There are two implementations, one for 1990 and one for 1997. Land use areas for 1997 were estimated using linear interpolation over time between the 1990 and 2002 land use maps (MDP, 2003a,b; U.S. Department of Commerce and U.S. Census Bureau, 2005).

Smithsonian Environmental Research Center. The core of the Smithsonian Environmental Research Center (SERC) model of the Patuxent watershed is a set of statistical models fit to measured water discharge and N and P concentrations collected weekly for 2 years from 22 study watersheds in the Patuxent River and adjacent Rhode River basins between July 1997 and August 1999 (Jordan *et al.*, 2003; Weller *et al.*, 2003). Annual rainfall was below average in the first year and above average in the second (Jordan *et al.*, 2003). The statistical models predicted discharge and nonpoint source nutrient concentrations from proportions of cropland and developed land, physiographic province, and time (Jordan *et al.*, 2003; Weller *et al.*, 2003). Landsat-derived land cover estimates (EPA-EMAP, 1994) were lumped to three categories (cropland, developed land, and other land) for use in the models. The Patuxent watershed was divided into 23 sections, and the fitted models were applied to the land cover and physiographic province data to predict weekly water discharge and weekly average nutrient concentrations leaving each section. Weekly nonpoint source material discharges were calculated by multiplying the weekly flow and average weekly concentration predictions. The effects of point sources and reservoir management were included by assimilating data on monitored discharges from two reservoirs and from wastewater treatment plants. The SERC model was applied to explore the effects of land use change and future development on watershed discharges (Weller *et al.*, 2003). The SERC model has also been linked to an estuarine water quality model (CE-QUAL-W2) to explore the effects of weather, watershed characteristics, and alternate land use scenarios on estuarine water quality and biological responses (Breitburg *et al.*, 2003; Lung and Bai, 2003; Lung and Nice, 2007).

SPAtially Referenced Regressions on Watershed Attributes. The USGS developed a set of nonlinear regressions called SPAtially Referenced Regressions On Watershed attributes (SPARROW) to relate observed TN and TP loads to spatially explicit nutrient sources reduced by losses to land-surface and in stream processes (Smith *et al.*, 1997; Preston and Brakebill, 1999). The nutrient sources include atmospheric deposition, urban land area, fertilizer application, livestock production, and point sources. The model statistically fit directly to TN and TP loads observed at points throughout the stream network, and does not estimate water discharge. The nonlinear regression procedure fits source coefficients for each nutrient source and delivery coefficients that relate nutrient losses to watershed characteristics, such as slope, soil permeability, stream density, and wetland area. Stream nutrient removal is represented as an exponential decay function of stream length and discharge volume. SPARROW models have been developed for several U.S. basins and analyzed to quantify nutrient sources, to estimate nutrients lost in river transport, to estimate nutrient delivery to estuaries, and to develop regulatory limits for implementing total maximum daily load (TMDL) regulations (see http://water.usgs.gov/nawqa/sparrow/).

We analyzed results from three versions of SPARROW models developed for the Chesapeake Bay watershed. Nutrient loads for the first version (SPARROW87) were estimated from 1950 to 1995 concentration and daily flow measurements from 109 Chesapeake watershed sites (79 for TN and 84 for TP), including 6 sites in the Patuxent watershed (Brakebill and Preston, 2003). The loads were normalized (Smith *et al.*, 1997) for 1987, the year for which input data were assembled (Preston and Brakebill, 1999). Land cover in 1 km pixels was mapped by integrating three data sets (U.S. Environmental Protection Agency Environmental Monitoring and Assessment Program [EPA-EMAP] [1994], National Oceanographic and Atmospheric Administration Coastal Change and Analysis Program [NOAA-CCAP] [2006], and USGS Geographic Information Retrieval and Analysis System [GIRAS] [Gutierrez-Magness *et al.*, 1997]). Stream networks were modified from the River Reach File 1 to derive hydrologic units (RF1, 1:500,000 scale) (Alexander *et al.*, 1999).

For the second version (SPARROW92), water quality data for 1950 to 1995 came from 132 sites (103 for TN and 121 for TP), including 6 sites, in the Patuxent watershed. Data were normalized (Smith *et al.*, 1997) for 1992 (SPARROW Version 2.0) (Brakebill *et al.*, 2001; Brakebill and Preston, 2003). Land cover came from integrating two data sets (EPA-EMAP [1994] and the 1990 National Land Cover Data [NLCD] [Vogelmann *et al.*, 2001]), and the watershed network came from the National Hydrography Data (NHD), 1:100,000 scale (USGS, 1999).

For the third version (SPARROW97), 1950 to 2000 load estimates from 125 sites (87 for TN and 103 for TP), including 6 Patuxent sites, were normalized (Smith *et al.*, 1997) for 1997 inputs (SPARROW Version 3.0) (Brakebill and Preston, 1999, 2004). Land cover data from the circa 1990 NLCD (Vogelmann *et al.*, 2001) were updated to 1997 using a change detection process based on spectral change between individual Landsat images. The watershed network was based on the stream network used in SPARROW92, with minor modifications, such as the addition of major reservoirs.

Smithsonian Environmental Research Center Landscape. Smithsonian Environmental Research Center Landscape (SERCLM) is a modular simulation model of the Patuxent watershed that was developed to generalize the analysis and application of the SERC model (described above) beyond the 2-year time domain and Patuxent only spatial domain of the SERC model. The SERCLM model includes three sub-components (Liu and Weller, 2008; Liu *et al.*, 2008). First, the TOPMODEL rainfall-runoff model (Beven and Kirby, 1979) is applied to estimate daily water discharge from 210 watersheds composing the Patuxent basin. TOPMODEL was manually calibrated to match observed flow at SERC and USGS monitoring stations (described previously). Second, two statistical models predict TN and TP concentrations from the proportions of cropland and developed land, physiographic province, time of the year, and water discharge estimated using TOPMODEL. The TN and TP models were fit to the water quality data set (Jordan *et al.*, 2003) described above. Finally, a stream routing model (Liu and Weller, 2008) combines the predicted discharges from the 210 watersheds with monitored data on reservoir and point source discharges, and then routes water and nutrients to the estuary while also accounting for nutrient uptake during transport. The stream routing parameters were calibrated manually (Liu and Weller, 2008; Liu *et al.*, 2008) to achieve the best match of streamflow, TN, and TP concentrations from SERC and USGS monitoring stations (described earlier).

Chesapeake Bay Program Model. The Chesapeake Bay Program (CBP) model of the Chesapeake Bay watershed is an adaptation of the Hydrologic Simulation Program — Fortran (HSPF) (Bicknell *et al.*, 2001), which was derived from the Stanford Watershed Model (Crawford and Linsley, 1966). HSPF uses a mass-balance approach to solve a linked set of equations representing natural and anthropogenic mechanisms that control nutrient transport and delivery to streams (USGS, 2011b). For the Chesapeake Bay application, the hydrologic simulation model is linked to a regional atmospheric deposition model (Linker *et al.*, 1996, 2000, 2008). Four increasingly refined versions have been released since 1994. Each version offered a more detailed segmentation, longer simulation period, and increasingly detailed representation of land use and best management practices (BMPs). The model was developed to quantify nutrient loads and their sources and to estimate load reductions from improved management practices. Model analyses of the impacts from alternative land management scenarios have helped to guide federal and state policy development, and loading estimates from CBP4 are used by regional, county, and municipal land managers to estimate loading rates associated with different land use types (e.g., the MDP model).

CBP4 simulates hourly sediment and nutrient discharge for a period of 17 years (1984-2000) in 94 model segments with an average size of 1,900 km^2. Land use and land cover data were derived from satellite imagery (EPA-EMAP, 1994) and ancillary data and consolidated into nine land use classes: pervious and impervious urban areas; mixed developed land; high till and low till croplands; livestock feeding areas; hay fields; pasture; and forest. Additional input data included fertilizer and manure applications, point source discharges, septic system densities, atmospheric deposition, and BMP reduction factors. The model was calibrated to 1984-1997 flow and TN and TP concentration data from 20 monitoring stations in the Chesapeake watershed, including the Bowie station in the Patuxent watershed (described above). Data from the Bowie station, dominated the calibration of model parameters for the Patuxent watershed.

The Chesapeake Bay Program's HSPF Model, Version 5 (CBP5) is the current model version (USEPA, 2010a). It divides the Chesapeake Bay watershed into nearly 1,000 sub-units with an average size of 171 km^2. The model simulates discharges over a 21-year period (1984-2005), but we analyzed outputs for 17 years (1984-2000). The CBP5 expands the land use classification to 24 categories and incorporates annual land use change. The base land use data was derived from a combination of 2000 land cover data developed by the University of Maryland's Regional Earth Science Applications Center (RESAC) (Goetz *et al.*, 2004), the 1992 NLCD (Vogelmann *et al.*, 2001), agricultural census data (http://www.agcensus.usda.gov/), and road network overlays (Tele Atlas, 2004). The model was calibrated to 1984-1995 flow and TN and TP concentration data from 296 stream monitoring stations in the Chesapeake watershed and nearby areas in Maryland and Virginia. Data from the same stations, but for the years 1995-2005 were used for model validation. There were seven stations in the Patuxent (including the three USGS sites described above) that dominated parameter estimates for the Patuxent watershed. The model has been applied to guide regulations to implement the Chesapeake Bay TMDL (USEPA, 2010c).

Patuxent Landscape. The Patuxent Landscape Model (PLM) is a spatially distributed simulation model for the Patuxent watershed that calculates daily hydrologic discharge and nutrient loads to streams (Costanza *et al.*, 2002). PLM was developed to evaluate how human settlements and agricultural practices affect hydrology, plant productivity, and nutrient cycling; and the model was applied to different scenarios of land use change to help guide regional management decisions. Hydrological, ecological, and biogeochemical processes are simulated in each grid cell using a set of Structural Thinking Experimental Learning Laboratory with Animation (STELLA) (http://www.iseesystems.com) modules from the Library of Hydroecological Modules (LHEM) (http://www.likbez.com/LHEM/) (Voinov *et al.*, 2004). The grid cells were linked within a Spatial Modeling Environment (Maxwell and Costanza, 1997) that used spatial data (land use, soil properties, climate, and nutrient input) to integrate the grid cell dynamics into regional-surface and groundwater hydrology and nutrient transport simulations (Voinov *et al.*, 1999, 2007). For the Patuxent River application, the watershed was segmented into 2,352 one km^2 grid cells. Land use and land use change were derived from MDP data for the years 1985 and 1990, modified using agricultural census data (http://www.agcensus.usda.gov/), and consolidated into five classes: forest, agriculture, rural, residential, and urban. Each land use type was modeled using equations and parameters describing the local biogeochemical dynamics. The parameters were calibrated to flow and nitrogen concentrations collected at the USGS Bowie gauge station between January 1986, and December 1993 (Voinov *et al.*, 2004).

Land Type Inputs

Nine different land use or land cover data sets were used in the 10 model implementations. We summarized the differences in land types among the models to provide background information needed to interpret differences among model load predictions. To describe the dominant land uses in the four study watersheds, we tabulated the average and ranges across the nine data sets for the percentages of four land types (cropland, grassland, developed land, and forest land) in each watershed. To document the differences in land cover percentages among models, we tabulated the percentages of the same land types used in the 10 model implementations for the entire Patuxent watershed.

Comparing Model Estimates to Observed Loads

Average Annual Loads. Every model implementation could predict some or all the average annual endpoints for flow, TN, and TP discharges from the three monitored watersheds (Table 2). The MDP and SPARROW implementations predict annual average TN and TP loads directly. The other models predict loads through time, but over different ranges of years (Table 1). Averaging across all the available years from each model produced annual average predictions for the time series models. The number of years available ranged from 2 (SERC) to 17 (CBP). The average annual predictions were tabulated and plotted against the observed annual averages from the 1984 to 2000 USGS monitoring data (above), and the range of predictions among models was reported as an initial characterization of uncertainty. We also tabulated the difference between each average annual prediction and the annual average observed values. We calculated the percent difference:

$$\% \text{ difference} = 100\left(\frac{P-O}{O}\right), \tag{1}$$

where P is a predicted flux and O is the corresponding observed flux. For each endpoint, the models were also ranked in order of increasing absolute value of the percent difference from the data with model rank 1 assigned to the model with the lowest such difference.

TABLE 2. Numbers of Models Predicting Average Annual Material Fluxes Leaving Three Basins Monitored by USGS Sampling Stations.

	Basin		
Material	**L**	**B**	**W**
Flow	4	5	3
TN	9	10	8
TP	9	9	8

Notes: Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). Tables 8, 9, and 10 indicate the specific models associated with any number in the table.

TABLE 3. Numbers of Models Predicting Annual Time Series Material Fluxes Leaving Three Basins Monitored by USGS Sampling Stations.

Material	Basin L	Basin B	Basin W
Flow	4	4	3
TN	5	5	–
TP	5	4	–

Notes: Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). Tables 11, 12, and 13 indicate the specific models associated with any number in the table. Annual time series nutrient data were not available for Western Branch (–).

Annual Time Series. The two SERC models, the two CBP models, and PLM could provide annual time series predictions (Table 3). The annual predictions from each implementation were plotted against annual observed values for the same years. For the nitrogen and phosphorus plots, we included the 95% confidence limits for the observed load. We plotted the confidence limits along the model prediction axis to reveal how the differences between model predictions and observations compare to the confidence limits on the observed loads. We summarized the difference from the observed value for each model and endpoint by applying Equation (1) in each year, then recording the average and range of percentage difference across years. We also calculated the Nash–Sutcliffe (NS) coefficient of model performance (Nash and Sutcliffe, 1970):

$$NS = 1 - \frac{\sum_{i=1}^{n}(O_i - P_i)^2}{\sum_{i=1}^{n}(O_i - \overline{O})^2}, \qquad (2)$$

where O_i is the observed discharge (of water, TN, or TP) in year i, P_i is the modeled discharge, $\overline{O}$ is the mean observed discharge, and n is the number years. A NS value of 1 indicates a perfect fit, a value of 0 indicates that the model is predicting no better than the average of the observations (Nash and Sutcliffe, 1970). Negative values indicate the model is performing worse than using the average of the observations. We used the NS values to rank the model performance for each endpoint, assigning rank 1 to the model with the NS closest to one (e.g., Gordon *et al.* 2004).

Estimating Total Loads to the Estuary

We tabulated average annual predictions of flow, TN, and TP discharge from the entire Patuxent watershed (Table 4), and we plotted the annual time series of predicted flow, TN, and TP discharges together with annual precipitation amounts (http://ches.communitymodeling.org/models/CBPhase5/datalibrary/meteorological-data.php). We also tabulated the molar ratios of N and P discharged to the estuary. This ratio is often used to draw inferences about which nutrient is more limiting to aquatic production (Redfield, 1958; Glibert *et al.*, 2006).

TABLE 4. Numbers of Models Predicting Average Annual or Annual Time Series Material Fluxes from the Entire Patuxent Basin.

Material	Average Annual	Annual Time Series
Flow	4	4
TN	9	4
TP	9	4

Note: Table 14 indicates the specific models associated with any number in the table.

Allocating Loads to Land Types

We compared how the models attributed the nonpoint source nutrient loads to developed land, agriculture, and other land types (Table 5). For the MDP model, we estimated the total nutrient loads from developed land by adding the loads from residential, urban, commercial, industrial, transportation, and utility lands, and we estimated the total nutrient loads from agriculture by summing loads from row crops, pasture, orchards, and confined feeding lots. For the SERC models, statistical model coefficients for cropland and developed land and the areas of the two land covers were used to isolate crop land and developed loads, with the remaining nonpoint source load attributed to other land. For the CBP models, we summed the nutrient loads from agricultural areas (including conventional-till, conservation-till, hay fields, pasture, and confined animal operations) to estimate the agricultural contributions; and we summed the loads from low-, medium-, and high-intensity developed land and developed open space to estimate the total contributions from developed lands. The SPARROW models attributed nonpoint loads to

TABLE 5. Numbers of Models that Could Allocate Nitrogen and Phosphorus Loads to Land Cover Types for Four Basins.

Material	Basin L	Basin B	Basin W	Basin P
TN	9	9	8	8
TP	9	9	8	8

Notes: Letters indicate basins: L (Laurel), B (Bowie), W (Western Branch), and P (entire Patuxent).

fertilizer inputs, manure inputs, developed land area, and atmospheric deposition inputs; however, atmospheric deposition was integrated into the nutrient contributions from the different land types in the MDP, SERC, and CBP models. To make the SPARROW estimates more consistent with the other models, we apportioned its estimated loads from atmospheric deposition to agriculture, developed land, and other land using the proportions of these three categories in the land cover data. The agricultural load from SPARROW then included part of the atmospheric deposition load plus the loads attributed to fertilizer and manure, whereas the total nonpoint source load from developed land included the load attributed to developed land area plus the portion of the atmospheric deposition load that could be attributed to atmospheric deposition on developed land. It was not possible to attribute loads to agriculture and developed land with the PLM model because the model does not track the origins of the delivered nutrient loads.

RESULTS

Land Type Inputs

Land type proportions for the entire Patuxent basin differed considerably among the sources of land data used by the models (Tables 6 and 7). The proportion of developed land ranged from 11 to 37%, and the cropland proportion ranged from 9 to 31%. We summed row crop, grassland (called pasture in some data sets), and animal feeding areas (represented only in some data sets) to estimate agricultural area. The resulting agricultural areas ranged from 15 to 36% of the basin. Forest (43-62% of the basin) and all other land (barren areas and wetlands, 0-2% of the basin) were more similar among the land data sets.

The estimated proportions of cropland and developed land also differed widely among the land data sets for each of the three monitored watersheds (Table 7). Despite those differences, all the data sets consistently identify the Laurel basin as the most agricultural of the four watersheds (average of 26% cropland and 14% developed) and the Western Branch as the most developed (average 13% cropland and 33% developed, Table 7).

Comparing Model Estimates to Observed Loads

Average Annual Loads. Predictions differed widely among models for all the average annual endpoints (Figure 2; Table 8). Streamflow predictions from the SERC and CBP models were more accurate for Western Branch (maximum absolute difference from observed flow of 11% or less, Table 9) than at Bowie and Laurel (up to 26–61% absolute difference from observed, respectively). The models tended to overestimate observed annual average flows for the Western Branch watershed, where there was a higher proportion of developed land and no reservoirs

TABLE 6. Land Cover or Land Use Percentages for the Entire Patuxent Basin as Used in 10 Model Implementations.

Implementation	Crop	Grass	Developed	Forest	Year	Data Source
MDP90	25	4	20	46	1990	MDP (2003a,b)
MDP97	24	3	27	43	1997	MDP (2003a,b)
SPARROW87	12	7	37	44	1990	Geographic Information Retrieval and Analysis System (GIRAS) (Gutierrez-Magness *et al.*, 1997); EPA-EMAP (1994); 1992 NLCD (Vogelman *et al.*, 2001)
SPARROW92	–	–	–	–	1992	EPA-EMAP (1994), 1992 NLCD (Vogelmann *et al.*, 2001)
SPARROW 97	–	–	–	–	1997	1990 NLCD (Vogelmann *et al.*, 2001); Landsat image change detection (Brakebill and Preston, 2004)
CBP4	12	7	37	44	1990	EPA-EMAP (1994); NOAA-CCAP (2006); USGS GIRAS (Gutierrez-Magness *et al.*, 1997)
SERC	10	28	12	49	1990	EPA-EMAP (1994)
SERCLM	10	28	12	49	1990	EPA-EMAP (1994)
PLM	31	5	11	51	1973	MDP (2003a)
	25	4	20	46	1990	MDP (2003a,b)
	24	3	27	43	1997	MDP (2003a,b)
CBP5	13	11	18	57	1984	1990 NLCD (Vogelmann *et al.*, 2001); 2000 RESAC
	9	7	22	62	2000	Land cover (Goetz *et al.*, 2004); Agricultural Census Data (http://www.agcensus.usda.gov)

Notes: Land cover numbers for the SPARROW 1992 and 1997 implementations were not included in the SPARROW publications (–). For all the data sets, the land not occupied by cropland, grassland, developed land, or forest was less than 3% of basin area.

TABLE 7. Means and Ranges (in parentheses) of Land Type Percentages Across Nine Land Use or Land Cover Datasets Used in 10 Model Implementations.

Basin	Crop	Grass	Developed	Forest
Laurel	26 (11-47)	15 (7-40)	14 (3-24)	44 (34-62)
Bowie	17 (8-32)	10 (5-30)	28 (16-43)	43 (36-56)
Western Branch	13 (3-23)	10 (2-33)	33 (22-45)	42 (36-50)
Entire Patuxent	18 (9-31)	9 (3-28)	22 (11-37)	50 (43-62)

Notes: Table 6 provides citations and details on how the 10 model implementations represented the entire Patuxent basin. The MDP analysis did not separate cropland and grassland, so the cropland + grassland sum for each model is shown for comparison to MDP. Land not occupied by cropland, grassland, developed land, or forest was always <3% of every basin. The cropland means and ranges summarize three land data sets for Western Branch and four data sets for the other basins. For the other land types, seven land data sets are summarized for Western Branch and eight for the other basins.

or point source contributions. For Laurel and Bowie, the models tended to underestimate streamflow. The SERCLM model best predicted flow at Laurel and Bowie, whereas CBP5 best predicted flow at Western Branch (Table 10), but both of these models also made relatively poor predictions for other basins. The SERC statistical model did not provide the best estimates for any endpoint, but consistently provided reasonable flow estimates (5–26% absolute difference from observed flow) for all three watersheds. The CBP4 model more significantly overestimated flow for Laurel and Bowie than did the other models, possibly because it did not account for water removed from the two reservoirs.

Predicted average annual TN loads were less accurate than flow predictions, with differences from observed TN ranging between −83 and 42%. Underestimates were more common for Laurel and Western Branch, the two smaller basins, than for Bowie (Table 9). With the exception of the SERC model, the export coefficient and statistical models predicted TN loads more accurately than the simulation models. For example, the MDP90 and MDP97 models best predicted TN loads at Laurel and Western Branch, and the SPARROW97 version best predicted average annual loads at Bowie. Although the SERC model provided accurate flow predictions, it underestimated TN loads by more than the other models, possibly because its 2-year time frame (August 1997-July 1999) provides a poor representation of average annual conditions (see Discussion).

As in the flow comparison, models that predicted well for one watershed often predicted poorly for others. For example, the MDP90 model predictions were closest to the observed values at Western Branch, but were poor at Bowie (rank 7.5 out of 11). The SPARROW97 model predictions most closely matched

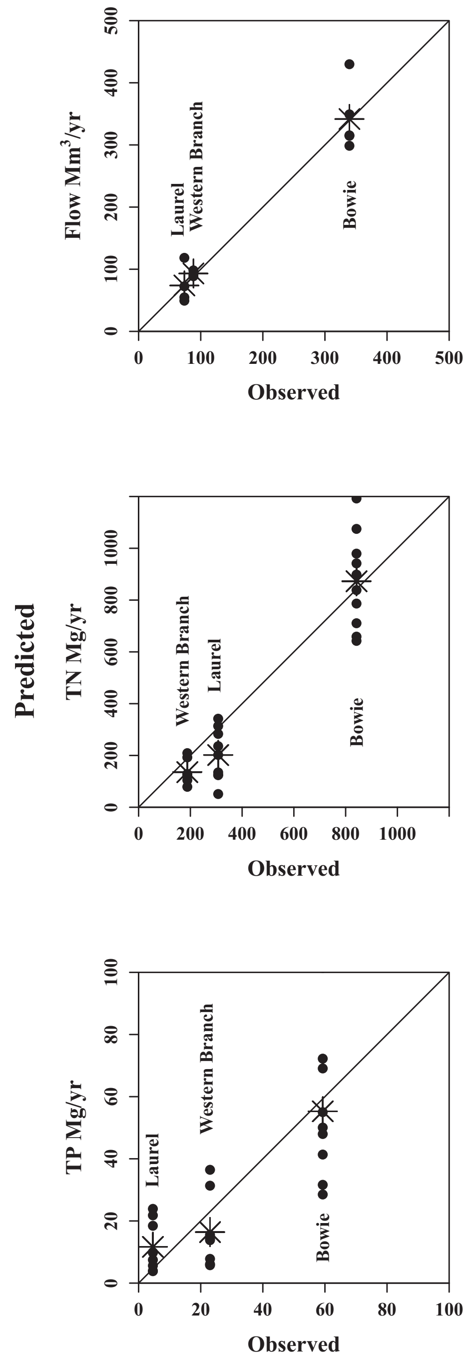

FIGURE 2. Annual Average Predictions from 10 Model Implementations *vs.* Measured Long-Term Average Annual Discharges of Water, TN, and TP. Model predictions (Table 8) are shown with small dots, and the model average prediction is shown with an asterisk. Predictions would equal observations along the solid diagonal line.

the observed annual average TN load at Bowie, but provided the poorest estimate at Western Branch.

Model performance in predicting observed TP loads was generally poorer than for TN, and the range of TP predictions among models was wider. Predictions for Laurel generally were more than 100% greater than

TABLE 8. Average Annual TN and TP Discharges from Three Monitored Watersheds Predicted by 10 Model Implementations and Measured.

	Flow Mm^3/yr			TN Mg N/yr			TP Mg P/yr		
Implementation	L	W	B	L	W	B	L	W	B
MDP90	–	–	–	284	193	711	21.8	13.8	50.1
MDP97	–	–	–	314	209	787	23.9	14.3	55.0
SPARROW87	–	–	–	342	126	1,075	9.8	5.8	41.4
SPARROW92	–	–	–	236	105	1,193	7.5	7.8	31.7
SPARROW97	–	–	–	202	79	839	3.9	6.1	28.6
SERC	54.8	92.4	315	124	125	659	5.6	31.4	69.1
SERCLM	72.7	98.3	349	135	119	980	10.0	36.5	101.4
CBP4	119	–	430	51.4	–	899	18.5	–	72.3
CBP5	49.5	88.5	316	125	129	643	4.0	15.6	48.0
PLM	–	–	299	–	–	942	–	–	–
Model average	73.9	93.2	342	202	136	873	11.7	16.4	55.3
Observed	73.7	88.3	339	308	188	842	4.6	23.0	59.3

Notes: Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). The MDP and SPARROW models did not predict flow, and some models did not predict TN and TP for all basins (–). Observed data are means for the years 1984-2000.

TABLE 9. Percent Difference from USGS Observed Discharges (Equation 1) for Average Annual Model Predictions.

	Flow			TN			TP		
Model	L	W	B	L	W	B	L	W	B
MDP90	–	–	–	–8	2	–16	374	–40	–16
MDP97	–	–	–	2	11	–7	420	–38	–7
SPARROW87	–	–	–	11	–33	28	114	–75	–30
SPARROW92	–	–	–	–23	–44	42	63	–66	–47
SPARROW97	–	–	–	–34	–58	0	–16	–73	–52
SERC	–26	5	–7	–60	–33	–22	22	36	17
SERCLM	–1	11	3	–56	–37	16	117	59	71
CBP4	61	–	27	–83	–	7	302	–	22
CBP5	–33	1	–7	–59	–31	–24	–14	–32	–19
PLM	–	–	–12	–	–	12	–	–	–
Model average	0	6	1	–34	–28	4	154	–29	–7

Notes: Negative values indicate underpredictions; positive values indicate overprediction. Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). Some models did not predict some endpoints (–).

the USGS observed loads, possibly because the models did not adequately account for P retention by the reservoirs. Only SPARROW97 and CBP5 were within 15% of the observed load. In contrast, the models tended to underestimate TP loads to Western Branch and Bowie. As with flow and TN, the highest ranked models differed among basins. At Laurel and Western Branch, the CBP5 model best matched the observed annual average TP load, whereas the MDP97 model best predicted TP loads observed at Western Branch and Bowie. The best performing TP models for a basin were not the models that were best for flow or TN.

In summary, the average annual analysis revealed several key patterns. No model consistently excelled across the materials and watersheds considered, and the models that best matched the observations at one endpoint were often among the worst models for another endpoint. Model skill in predicting the observed data was best for flow, intermediate for TN, and poor for TP. There was no relationship between how well the simulations predicted flow and how well they predicted TN or TP loads. The export coefficient and statistical models (MDP, SPARROW, and SERC) were generally better predictors of TN loads than the simulation models (SERCLM, CBP, and PLM). All the models were poor predictors of TP loads. For all materials, models were closest to the observations at Bowie (the largest basin) and most different from the observations at Laurel (the smallest basin), suggesting that model performance improved with watershed size.

Annual Time Series. The analysis of annual time series endpoints (Table 3) supported the findings from the average annual analysis. Model performance again differed among materials and locations (Figure 3; Tables 11-13). For example, SERCLM best predicted flow at the Laurel outlet (NS = 0.9), but was the least accurate at Western Branch (NS = –8.2). The

TABLE 10. Ranked Performance of Models for Average Annual Predictions Based on the Absolute Value of Percent Difference from USGS Observed Discharges (Table 9).

	Flow				TN				TP				
Model	**L**	**W**	**B**	**Mean Rank**	**L**	**W**	**B**	**Mean Rank**	**L**	**W**	**B**	**Mean Rank**	**Overall Rank**
MDP90	–	–	–	–	2	1	6.5	3.2	9	5	3	5.7	4.4
MDP97	–	–	–	–	1	2	3.5	2.2	10	4	1.5	5.2	3.7
SPARROW87	–	–	–	–	3	5.5	10	6.2	5	9	7	7	6.6
SPARROW92	–	–	–	–	4	8	11	7.7	4	7	8	6.3	7.0
SPARROW97	–	–	–	–	5.5	9	1	5.2	2	8	9	6.3	5.8
SERC	3	2	3.5	2.8	9	5.5	8	7.5	3	3	4	3.3	4.6
SERCLM	2	4	2	2.7	7	7	6.5	6.8	6	6	10	7.3	5.6
CBP4	5	–	6	5.5	10	–	3.5	6.8	8	–	6	7	6.4
CBP5	4	1	3.5	2.8	8	4	9	7	1	2	5	2.7	4.2
PLM	–	–	5	5	–	–	5	5	–	–	–	–	5.0
Model average	1	3	1	1.7	5.5	3	2	3.5	7	1	1.5	3.2	2.8

Notes: Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). Some models did not predict some endpoints (–).

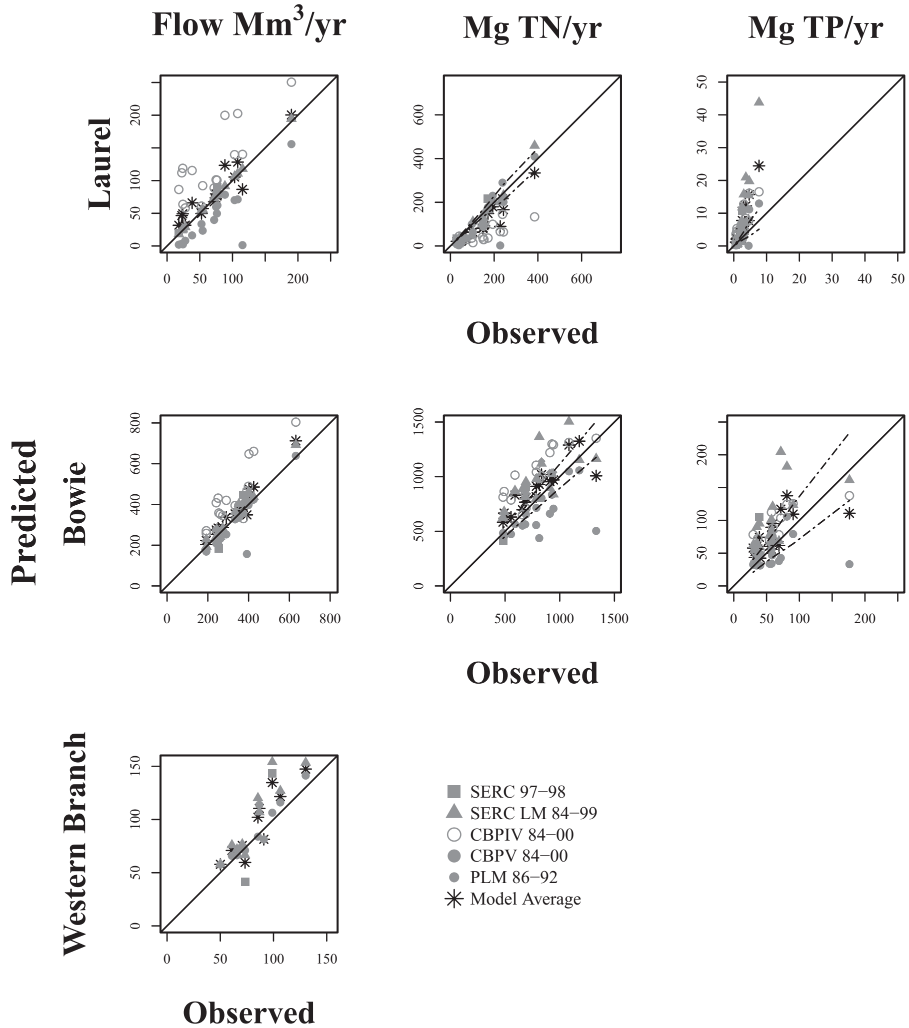

FIGURE 3. Annual Time Series Predictions from Five Model Implementations *vs.* Measured Annual Discharges of Water, TN, and TP. Predictions would equal observations along the solid diagonal line. For the Laurel and Bowie watersheds, the dotted lines on the TN and TP plots are the 95% confidence limits for the USGS estimates of observed loads. The confidence limits are plotted against the vertical axis to show how differences between model predictions and observed loads compare to the 95% confidence limit of the observed loads.

TABLE 11. Percent Differences from USGS Observed Discharges (Equation 1) for Annual Time Series Predictions.

Model	Years	Flow L	Flow W	Flow B	TN L	TN B	TP L	TP B
SERC	2	15 (14 to 16)	17 (16 to 19)	8 (7 to 10)	20 (15 to 25)	20 (19 to 21)	141 (−33 to 315)	126 (116 to 135)
SERCLM	16	3 (2 to 6)	21 (−9 to 56)	6 (−8 to 19)	−15 (−49 to 19)	22 (−13 to 68)	177 (−45 to 468)	79 (−8 to 186)
CBP4	17	126 (6 to 397)	–	34 (−12 to 72)	−54 (−85 to 12)	34 (−1 to 76)	166 (37 to 455)	46 (−22 to 165)
CBP5	17	−52 (−99 to −11)	6 (−10 to 31)	−10 (−60 to 2)	−39 (−99 to 22)	−17 (−62 to 1)	9 (−98 to 137)	−13 (−81 to 31)
PLM	7	–	–	−1 (−1 to −1)	–	18 (−9 to 48)	–	–
Model average	17	26 (−25 to 106)	14 (−19 to 36)	8 (−11 to 19)	−19 (−61 to 96)	13 (−24 to 53)	146 (−11 to 287)	32 (−37 to 95)

Notes: The percent difference from observed discharges was calculated for every model in every year. For each model and material, the table presents the mean and range (in parentheses) of those differences across years. Some models did not predict some endpoints (–).

TABLE 12. Nash-Sutcliffe Coefficients for Annual Time Series Predictions.

Model	Years	Flow L	Flow W	Flow B	TN L	TN B	TP L	TP B
SERC	2	0.90	−8.22	−0.36	0.81	−1.32	−35.70	−185
SERCLM	16	1.00	−0.17	0.90	0.91	−0.11	−44.00	−1.39
CBP4	17	−0.89	–	−0.47	−0.40	−0.65	−6.61	0.25
CBP5	17	0.30	0.77	0.67	0.51	−0.17	−1.43	−0.26
PLM	7	–	–	1.00	–	−1.19	–	–
Model average	17	0.85	0.40	0.86	0.69	0.55	−10.68	0.23

Notes: Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch). TN and TP flux measurements were not available for Western Branch; PLM predicted only flow and TN at Bowie; and CBP4 did not predict fluxes for Western Branch (–).

TABLE 13. Ranked Model Performance for Annual Time Series Predictions Based on Nash-Sutcliffe Coefficients (Table 12).

Model	Years	Flow L	Flow W	Flow B	Flow Mean Rank	TN L	TN B	TN Mean Rank	TP L	TP B	TP Mean Rank	Overall Rank
SERC	2	2	3	2	2.3	2	2	2.0	4	5	4.5	2.9
SERCLM	16	1	4	3	2.7	1	3	2.0	5	4	4.5	3.0
CBP4	17	5	–	6	5.5	5	6	5.5	2	1	1.5	4.2
CBP5	17	4	1	5	3.3	4	4	4.0	1	3	2.0	3.1
PLM	7	–	–	1	1.0	–	5	5.0	–	–	–	3.0
Model average	17	3	2	4	3.0	3	1	2.0	3	2	2.5	2.6

Notes: Models with NS coefficients closest to 1 were ranked highest (see text). Letters indicate basins: L (Laurel), B (Bowie), and W (Western Branch).

CBP5 best predicted flow at Western Branch (NS = 0.77), but provided relatively poor predictions at Laurel (NS = 0.3). TN loads predictions were again less accurate and more different among models than flow predictions. The SERC and SERCLM models performed well in predicting annual TN loads (NS > 0.8) at Laurel, but all the models performed poorly in predicting annual TN loads at Bowie (NS < 0).

The ranges of differences between the TP load predictions and observations were again higher than those for TN. At Laurel, TP loads were generally overestimated, and all the models had years with predictions >100% of the observed annual loads (Table 11). Differences between observations and the model predictions were smaller for the larger Bowie watershed. The NS coefficients (Table 12) indicate that none of the models are good predictors of TP loads. NS values were below 0 for all but one TP endpoint, indicating the models were less reliable than using the average observed loads as a predictor. The CBP4 model had a positive coefficient (NS = 0.25) at Bowie, but the value (NS = 0.25) was still below the threshold of good performance (i.e., NS > 0.5) (Moriasi *et al.*, 2007). The SERC and SERCLM models tended to overestimate observed TP loads by more than 100% because of the measured TP loads used to fit the SERC models were higher than the USGS TP observations (see Discussion).

For both TN and TP, most model predictions fell outside the 95% confidence limits for the observed loads (Figure 3) produced using the ESTIMATOR model (Cohn *et al.*, 1989), suggesting that the differences between model predictions and observations are statistically significant. A few model predictions fell within the 95% confidence limits of the observations, but such agreement was again not consistent among materials and locations for any model.

In summary, the analysis of annual time series predictions supported the key patterns identified in the analysis of average annual endpoints: no model consistently excelled across the endpoints, model skill was highest for flow, intermediate for TN, and poor for TP; and skill in predicting one material seemed unrelated to the skill in predicting the others. In addition, models that performed well in predicting an average annual endpoint were not necessarily among the best models for predicting the annual time series for the same material and basin. In some cases, this may partly reflect the limited number of years available to estimate the annual average (see Discussion).

Performance of the Model Average. We compared the performance of the model average to the individual models for every endpoint. For many of endpoints considered above, at least one model provided a better estimate than the model average, but across all the endpoints, the model average performed more consistently and more reliably than any single model (Figures 2 and 3, Tables 8-13). The model average had the best overall rank across all the average annual and annual time series endpoints (Tables 10 and 13). The model average worked well partly because each individual model performed poorly for some endpoints. For example, the annual average TN load predicted by SPARROW97 was most similar to the observed long-term average at Bowie, but least similar at Western Branch. The MDP90 TN predictions agreed closely with the long-term data at Laurel and Western Branch, but not at Bowie. Model performances also varied among response variables. For example, the MDP97 model effectively predicted long-term TN loads but not TP loads; and the reverse was true for CBP5. The model average did perform poorly for particular endpoints where a single, very poorly performing model dominated the average (as when the CBP4 model greatly overestimated annual average flow at Laurel) or where all the models either over- or under-predicted the observations (e.g., annual TP loads at Laurel). The model average performed better for the annual time series endpoints (and monthly time series endpoints, Supporting Information) than for the average annual endpoints.

Total Loads to the Estuary

Across the model set, predicted average annual flow from the entire watershed to the Patuxent estuary ranged between 840 and 9,100 Mm^3/yr, and the model average flow was 2,900 Mm^3/yr (Table 14). Predicted average annual TN loads to the subestuary ranged between 1,400 and 2,900 Mg N/yr. Estimates from the SERCLM and CBP5 models were remarkably similar (1,750 Mg N/yr), whereas estimates from MDP, SPARROW92, and CBP4 were consistently higher than the other models. The model average was 2,115 Mg N/yr. Predicted average annual TP loads ranged between 60 and 340 Mg P/yr, and the model average was 191 Mg P/yr.

The molar ratio of TN to TP in the average annual discharge ranged from 11 to 78, with a model average of 24 (Table 14). The ratio for the two SERC models was 11, which is less than the Redfield ratio of 16 (Redfield, 1958; Glibert *et al.*, 2006), suggesting a possible excess of phosphorus over nitrogen relative to the needs of phytoplankton. The other seven model implementations had TN:TP ratios above 16, suggesting a relative excess of nitrogen in watershed discharges to the estuary. The SPARROW92 implementation had the most extreme TN to TP ratio (78, Table 14).

Annual time series predictions of flow, TN, and TP increased directly with annual precipitation, but the response to precipitation differed among models (Figure 4). Absolute and relative differences among the model predictions were greater during wetter years. CBP4 estimates of flow were consistently higher and increased more with additional precipitation than did flow estimates from the SERCLM or CBP5 models, which were consistently similar throughout the study period. Between 1984 and 2000, flow estimates from the SERCLM, CBP4, and CBP5 models ranged between 500 to 1,500 Mm^3/yr, 5,000 to 16,000 Mm^3/yr, and 850 to 2,700 Mm^3/yr, respectively. All three models predicted their highest annual discharge during the wettest year (1996), but they did not all predict the lowest annual flow during

TABLE 14. Predicted Average Annual Flow, Total Nitrogen Load, and Phosphorus Load to the Patuxent Estuary and Atomic N to P Ratios.

Model	Years Averaged	Flow (Mm^3/yr)	Mg N/yr	Mg P/yr	N:P
MDP90	–	–	2,624	186	31
MDP97	–	–	2,860	198	32
SPARROW92	–	–	2,755	78	78
SPARROW97	–	–	1,428	64	49
SERC	2	858	1,568	311	11
SERCLM	16	848	1,749	338	11
CBP4	17	9,076	2,185	189	26
CBP5	17	867	1,751	167	23
Model Average	–	2,912	2,115	191	24

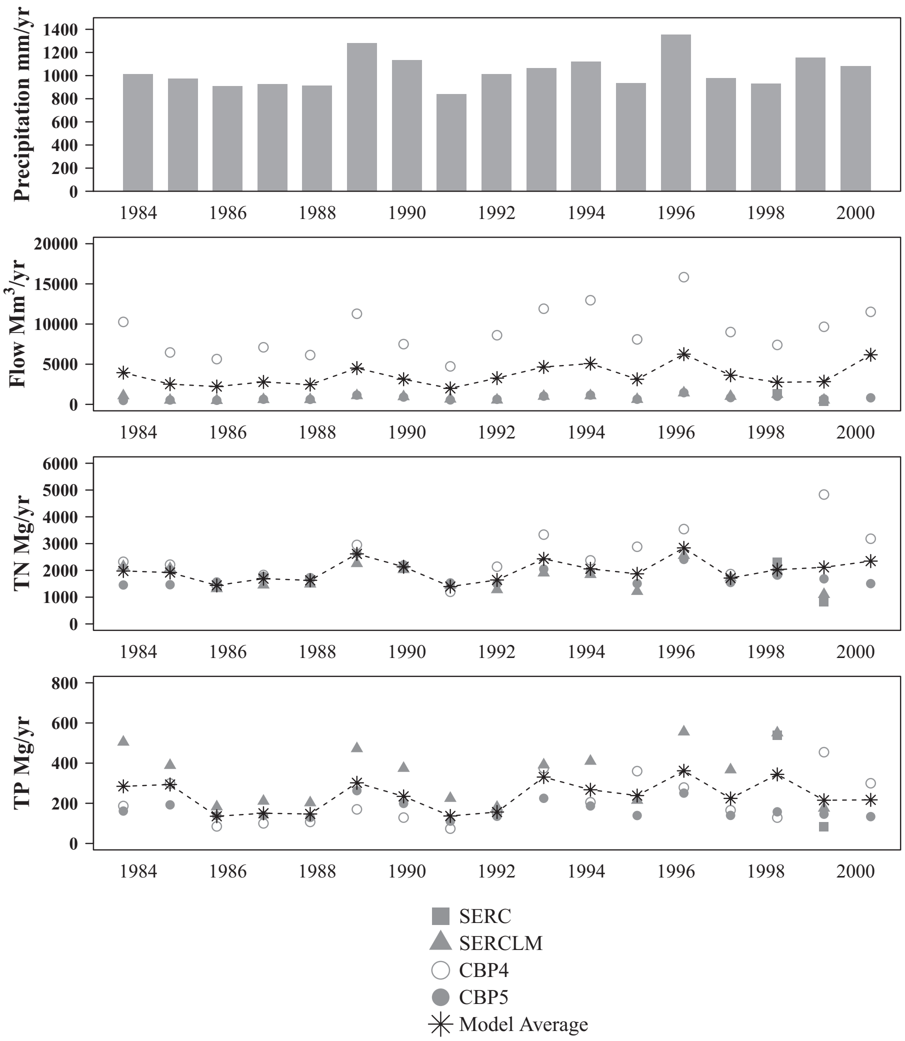

FIGURE 4. Annual Precipitation and Annual Time Series Predictions of Water, TN, and TP from the Entire Patuxent Watershed. The dashed line connects the model average predictions.

the driest year (1991). Among years, the model average annual P load ranged between 2,400 and 6,700 Mm^3/yr.

Patterns in the predicted annual time series nutrient loads were similar to the patterns in the predicted annual flow. For TN, the SERCLM and CBP5 models consistently predicted lower loads (range: 1,100-2,600 and 1,400-2,600 Mg N/yr, respectively) than the CBP4 (range: 1,200-4,800 Mg N/yr). The model average ranged between 1,400 and 2,800 Mg N/yr. For TP, the SERC and SERCLM models predicted higher annual loads (range: 85-540 and 175-560 Mg P/yr, respectively) than the CBP4 and CBP5 models (75-460 and 110-270 Mg P/yr, respectively), probably because of differences in the underlying measurements used for model calibration (see Discussion). The model average ranged between 135 and 365 Mg P/yr.

Allocating Loads to Land Types

The attribution of nonpoint source (NPS) annual average TN loads to agriculture and to developed land differed among models (Tables 5, 15, and 16). In the Laurel watershed, all models identified agriculture as the majority TN source (range 51-95% of NPS load). The SERCLM model predicted the highest agricultural contribution despite using the lowest estimated proportion of agricultural land area, indicating that the differences among models in source allocation reflect more than just the differences in land type proportions. In the other three watersheds, some models attributed the majority of the TN load to agriculture, whereas other models attributed the majority to developed land. For Bowie, the SPARROW and SERC models attributed more nitrogen to agriculture, whereas the MDP and CBP models attributed more

TABLE 15. Model Average and Range Among Models (in parentheses) for the Percentages of Predicted Average Annual TN and TP Discharges Allocated to Cropland and Developed Land in Four Watersheds.

	TN		TP	
Basin	**Cropland**	**Developed Land**	**Cropland**	**Developed Land**
Laurel	72 (51 to 95)	12 (1 to 24)	64 (40 to 91)	12 (3 to 17)
Bowie	46 (23 to 69)	39 (17 to 59)	35 (21 to 51)	48 (34 to 59)
Western Branch	47 (18 to 62)	37 (17 to 56)	47 (19 to 81)	38 (17 to 67)
Entire Patuxent	49 (27 to 72)	35 (18 to 54)	47 (23 to 75)	35 (12 to 48)

TABLE 16. Percentages of Predicted Average Annual TN and TP Discharges from the Bowie Basin Allocated to Cropland and Developed Land by Nine Model Implementations.

	TN		TP	
Model	**Agriculture**	**Developed Land**	**Agriculture**	**Developed Land**
MDP90	37	50	51	46
MDP97	31	58	44	54
SPARROW87	63	27	30	37
SPARROW92	56	40	27	55
SPARROW97	71	25	26	52
SERC	49	36	46	44
SERCLM	66	27	50	34
CBP4	25	54	21	52
CBP5	32	50	30	56
Model average	48	41	36	48

to developed land. For all four watersheds, the model average contribution from agriculture was larger than the model average contribution from developed land.

The allocation of NPS TP loads differed even more among the models than did the TN allocations. For the Laurel basin, the agricultural contribution to NPS TP loads ranged from 40 to 91% among models, and all models estimated that developed land contributed a much smaller fraction (<17%) of the NPS TP load. For Bowie, the MDP90, SERC, and SERCLM models predicted that more TP came from agriculture than developed land, whereas the MDP97, SPARROW, and CBP models predicted the reverse. For three basins, the model average indicated that agriculture contributed more TP than developed land, but the developed land contribution at Bowie was higher than the agricultural contribution.

DISCUSSION

Our ensemble analysis of watershed models yielded insights that could not have been revealed by examining a single model. We found that none of the individual models was consistently best in matching observed loads across the set of endpoints that we examined. Instead, the model average prediction was the most consistently reliable predictor, and the range of predictions among models provided a first order estimate of uncertainty. For nutrients, the range among model predictions was much larger than the confidence limits for the observed loads, suggesting that the uncertainties in modeling are much larger than the uncertainties in load measurement. In some cases, consensus among the models would justify confidence in the model predictions and the underlying knowledge, but other analyses revealed large differences among the models. Those differences suggest areas where more research is needed to provide better data or better understanding of watershed processes. These insights from ensemble modeling support the importance of considering multiple models and the need to adaptively manage land use practices to protect water resources. The following sections provide more details on each of these findings.

There Is No "Best" Model

No single model consistently outperformed the other models across the endpoints for which we compared predictions to observations. The models that most closely matched the data for one endpoint often were among the worst models for another endpoint (Tables 10 and 13, Figures 2, 3, and 4). Such an outcome is not inevitable, and some ensemble analyses could identify a model that is generally superior to the alternatives across particular sets of predictions. However, our finding of "no best model" matches the experience of a more intensive watershed model comparison implemented across the European Union (Bormann *et al.*, 2007, 2009; Breuer and Huisman, 2009; Breuer *et al.*, 2009; Huisman *et al.*, 2009; Viney *et al.*, 2009).

The Range of Predictions Helps Quantify Model Uncertainty

The range among models in an ensemble provides a first order estimate of model uncertainty for predicting an endpoint, thus helping to objectively assess confidence in model predictions (Beven, 2007; Clark, 2007). As the range accounts for differences among models, it reflects uncertainty in the model structure (Tobias and Li, 2004). In our model set, the SERC and SPARROW models were statistical models that could provide predictions with confidence limits, but the loading coefficient (MDP) and simulation models (SERCLM, CBP, and PLM) provided no estimates of prediction uncertainty. Indeed, most watershed simulation models do not provide uncertainty estimates (Pappenberger and Beven, 2006). The uncertainty range from ensemble modeling can help describe prediction uncertainty while research continues on better methods to quantify uncertainty in complex simulation models. There has been some success in quantifying prediction uncertainty in simple watershed models (e.g., Alexander *et al.*, 2002). Increases in computer power and ongoing research in uncertainty analysis methods may eventually enable more complete uncertainty analyses of more complex simulation models, like the CBP or PLM models in our ensemble, but such analyses of individual models cannot capture uncertainty in the underlying model structure.

The Model Average Is the Most Consistent Predictor

Averaging across the model ensemble provided more reliable predictions across the set of endpoints than any single model. Unlike any of the individual models, the model average estimates generally were within the top 10% across all endpoints and always within the top 50% of the ranked performances of individual models (Tables 10 and 13). The model average was not always the closest to the observed data for every endpoint, so it should not be enshrined as a "best" model (see previous section). However, when data are not available to test model predictions, the model average is likely to be a better estimate than any single model. Other studies have also reported that the model average is more consistently reliable than a single model (Breuer and Huisman, 2009; Viney *et al.*, 2009).

We used the simple average of model predictions in our investigations of model averaging, but model averaging is most successful when it incorporates penalties for model complexity, considers uncertainties of model predictions, and weights models by past performance (i.e., Bayesian model averaging) (Kadane and Lazar, 2004; Tobias and Li, 2004). These measures minimize the influence of poor or overly complicated models. Bayesian model averaging has gained widespread application in financial forecasting and socioeconomics (e.g., Wright, 2008; Tobias and Li, 2004), weather (e.g., Koop and Tole, 2004; Gneiting and Raftery, 2005), and more recently, in hydrology (e.g., Gourley and Vieux, 2005). We could not apply penalties for model complexity, because we did not have measures of model complexity and uncertainty (see above) for most of the models in our ensemble. Better methods quantifying model complexity and uncertainty for large simulation models (see above) would enhance the interpretation of individual models and the power of multi-model approaches like model averaging.

Consistencies and Differences Among Models

Nutrient Source Areas. Our ensemble analysis found some patterns of agreement among the models. Despite the differences in land data and nutrient generating activities considered, all the models indicate that agricultural lands release more N and P per unit area than do developed lands. Empirical watershed studies have also reported higher rates of nutrient release from croplands than from developed lands (Beaulac and Reckhow, 1982; Jordan *et al.*, 1997a,b, 2003; Liu *et al.*, 2000).

Watershed models often are analyzed to identify the source areas for nutrient discharges. In our comparison, the models allocated different fractions of the nutrient loads to agriculture and to developed land. In some watersheds, some models identified agriculture as the dominant source area, whereas other models identified developed land as the more important source (Tables 15 and 16). Other model comparisons also have reported widely different partitioning among nutrient sources for different models (Valiela *et al.*, 2002; Jordan *et al.*, 2003; Weller *et al.*, 2003). The differences are important because the estimates could affect how management efforts are targeted to the two land types (Valiela *et al.*, 2002). Together, the models we studied indicate that agriculture and developed land both contribute substantially to nutrient loads, and both land types should be managed to improve water quality.

Watershed Estuary Linkages. Watershed models have been linked to estuarine models to help understand and manage the impacts of human activity and management efforts on estuarine water quality and living resources (e.g., Cerco, 1995; Brandt and Mason, 2003; Lung and Nice, 2007). These efforts often estimate the atomic ratio of nitrogen and phosphorus in estuarine inputs and compare that estimate to the Redfield ratio 16, which represents the average atomic ratio of N to P in phytoplankton (Redfield, 1958; Glibert *et al.*, 2006). TN to TP ratios greater than 16 may mean that N is abundant relative to P, so that P is the nutrient more limiting to phytoplankton production. TN to TP ratios less than 16 imply the reverse, that N supply is more limiting to primary production.

The predicted TN to TP ratio varies widely among the models (Table 14), ranging from 78 (SPARROW92) to 11 (SERC and SERCLM). Only the two SERC models suggest N limitation by predicting ratios below 16. The other six model implementations, all predict TN to TP ratios greater than 16, suggesting that P is the more limiting nutrient. Inferences based on TN to TP ratios are not conclusive because phytoplankton respond to dissolved inorganic nutrient concentrations (not total concentrations), and because much of the TP in watershed discharges is attached to particles that may become buried in sediment and remain unavailable to phytoplankton (Hartzell *et al.*, 2010). The wide range of predicted TN to TP ratios does indicate that knowledge of the linkage between watershed discharges and estuarine nutrient limitation remains uncertain. Relying on a single watershed model would limit our understanding of that uncertainty, whereas using multiple models can improve our confidence in the overall model predictions.

Relative Predictably of Flow, TN Load, and TP Load. There were consistent patterns across models in the relative uncertainties of predictions for different materials. For all the models that had published calibration results, the performance of the calibrated model was better for water discharge than for TN load and worst for TP load (Linker *et al.*, 2000; Costanza *et al.*, 2002; Weller *et al.*, 2003; Liu *et al.*, 2008). Our ensemble analysis confirmed those patterns. For all three time frames that we considered (average annual, annual time series, and monthly), performance metrics were best for flow, intermediate for TN, and worst for TP; and the ranges of estimates among models were narrowest for flow, intermediate for TN, and widest for TP (Tables 8 and 11, Figures 2, 3, and 4, Supporting Information). Nutrients are harder to predict than flow partly because nutrient release, transport, and removal are all strongly driven by factors that are temporally episodic and spatially heterogeneous, making them hard to represent with deterministic models.

Controls of Phosphorus Delivery. The poor performance and high uncertainty in predicting TP loads suggests that the models do not capture the dominant watershed processes controlling the transport and delivery of phosphorus. The P in streams is mostly associated with sediments (Jordan *et al.*, 1997a,b), so effectively modeling P loads requires a good representation of sediment generation and transport. Most watershed models assume that hillslope erosion is the proximal source of sediment (and associated particulate P) in stream loads, although empirical tests have demonstrated that hillslope erosion models are poor predictors of suspended sediment loads (Boomer *et al.*, 2008; Wilkinson *et al.*, 2009) and that other processes likely control sediment generation and transport (de Vente and Poesen, 2005). Important processes identified in field studies include: gully erosion (Wells *et al.*, 2009), seepage erosion (Fox and Wilson, 2010), stream bank erosion (e.g., Walter and Merritts, 2008; Devereux *et al.*, 2010; Mukundan *et al.*, 2010), in-stream erosion and deposition (Dearing and Jones, 2003), and floodplain deposition (e.g., Noe and Hupp, 2009). Models that move beyond hillslope erosion to account for more of these processes (Prosser *et al.*, 2001; Wilkinson *et al.*, 2009) may provide more accurate and precise TP predictions. Application of inorganic fertilizer or manure can also promote soil P saturation and increase delivery of dissolved P to streams (Staver and Brinsfield, 2001). The poor performance of most watershed models in predicting P loads suggests a critical need for field and modeling research to understand the relative importance of different sediment and phosphorus transport processes.

Errors in Model Implementation. Our independent review of output from each model also helped identify and correct many errors and inconsistencies in the model implementations. For every model, we found errors in the model output caused by oversights in running the model or in summarizing its output for analysis. These included errors in data entry, database queries, or unit conversions. The mistakes were not evident when examining a single model, but became clear when predictions from several models were compared.

Sources of Differences Among Models

Land Use or Land Cover Data. Some of the differences among models in load predictions came from large and systematic differences among the land type data sets used to drive different models (Tables 6 and 7). Clearly, there are fundamental differences in land classification that yield differences in the proportions of cropland and developed land among model inputs. These are the land types responsible for most nonpoint TN and TP, so the differences in their proportions are likely an important source of differences in modeled TN and TP discharges. Some of the differences among the data sets arose from processing land type maps to provide model input. The SERC, SERC-LM, and PLM models used published data without any modifications. The SPARROW and CBP input data were synthesized from several land cover maps combined with county agricultural census data and counts of septic systems. The modified data sets had higher proportions of developed land and less cropland than the unmodified land cover maps. The MDP data sets, which were interpreted from aerial photography, also had higher percentages of developed land than the unmodified data sets derived from satellite imagery, such as the EPA-EMAP (1994) data (Table 6). The differences remind us that land type data do not come from simple, direct measurements, but are instead derived from interpreting aerial photography or from applying classification models to remotely sensed data.

The dates of the land data sets range from 1973 to 2000 (Table 6); so some of the differences among them come also from land use change. Most watershed models use information from a single land use or land cover data set. Even many models that include detailed representation of temporal responses to precipitation and temperature still assume that land use remains constant through time. In our model set, only the CBP5 and PLM models incorporated information on land use change to dynamically account for its effects in multi-year simulations (but the estimated temporal changes in land use proportions are similar in magnitude to the differences in proportions among data sets for similar dates; Table 6). Given the important effects of land types on water and nutrient discharges, more models need to account for land changes, especially in applications over multi-year periods.

More research is needed to classify land use and land cover more accurately and consistently, to quantify the uncertainty in those classifications, and to propagate those uncertainties through watershed models to measure how they affect the uncertainty of predicted loads. The current uncertainties in land characterization confound our interpretation of predicted nutrient source allocations or impacts from alternative land use management scenarios (Huisman *et al.*, 2009). Recent progress in refining land classification models with ancillary data (e.g., Pyke, 2010; P. Claggett, USGS, unpublished data) may provide more reliable land type data for the Chesapeake Bay region.

Stream Measurements of Nutrient Loads. Disparities in measured loads also contributed to differences among model predictions. In particular, the SERC model was a poor predictor of the USGS observed TP loads (Table 12), despite the model's strong calibration results (Jordan *et al.*, 2003; Weller *et al.*, 2003). This discovery led us to compare the USGS streamflow and nutrient load data with the independent SERC measurements used to calibrate the SERC and SERCLM models. Nutrient loads reported by the USGS are derived from a log-linear regression model that estimates nutrient loads from measured nutrient concentrations, measured streamflow, and a function of time that represents seasonality and possible linear trends (Cohn *et al.*, 1992). The USGS measured nutrient concentrations during periodic short sampling events using a spatially integrated approach in which samples were collected across the stream width and depth and then composited. Streamflow was monitored continuously. The SERC data came from automated samplers, which measured stream depth continuously and collected weekly flow-weighted composite samples for nutrient analysis (Jordan *et al.*, 1997c), a method which has been reported to provide direct and accurate estimates of loads (Stone *et al.*, 2000; Harmel *et al.*, 2006). Weekly mean flow rates and flow-weighted mean concentrations were multiplied to estimate weekly loads for the sampling period between August 1997 and July 1999.

A comparison of SERC and USGS monthly observed TN loads revealed that many of the SERC measurements fell outside the 95% prediction interval for the USGS observations provided using ESTIMATOR (Cohn *et al.*, 1989), but there was still a strong correlation ($R^2 = 0.90$) and data were grouped symmetrically around the 1:1 line (Figure 5). However, the observed TP loads were less strongly correlated ($R^2 = 0.71$) and the USGS and SERC loads were systematically different. The geometric mean slope (appropriate when two variables are both measured with error) (Sprent and Dolby, 1980) of SERC TP *vs*. USGS TP is 1.97, indicating that SERC-measured TP discharges are roughly twice the USGS-measured TP discharges. The relationship also explains why the SERC and SERCLM models (which were calibrated with the SERC measurements) made predictions that were strongly correlated with, but systematically greater than the USGS measurements (see Figure 3). If the SERC TP measurements are correct, then models calibrated with the USGS TP measurements would significantly underestimate TP loads. The ESTIMATOR modeling approach applied using USGS to estimate TP loads from flow and concentration data may underestimate loads during high flow periods. Further research comparing volume-integrated composite sampling using ESTIMATOR modeling and quantifying the uncertainties in both types of measurements would help identify the most accurate way to measure TP loads.

Time Period Considered. Differences in the modeled time period also contributed to differences among model predictions. The clearest example of possible confusion arises in evaluating the performance of the SERC model. This simple statistical

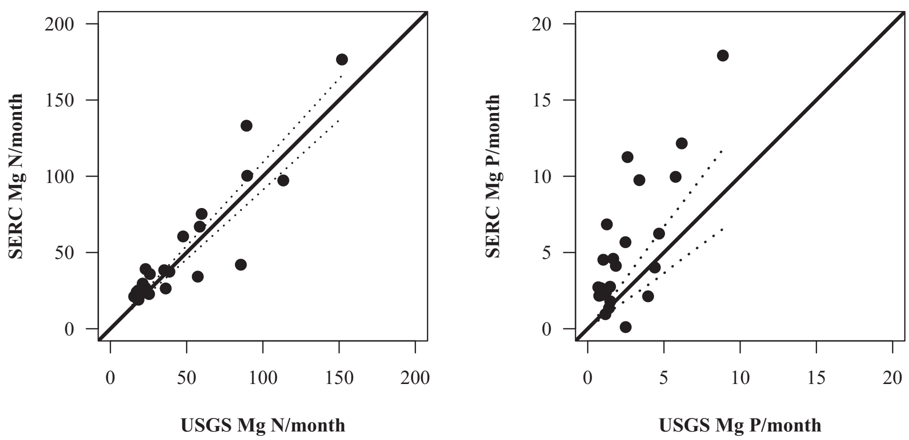

FIGURE 5. Comparison of USGS and Smithsonian Environmental Research Center (SERC) Observations of Monthly Nutrient Discharges from the Bowie Watershed. The SERC and USGS measurements would be equal along the diagonal line. The dotted lines are the 95% confidence limits for the USGS estimates of observed loads. The confidence limits are plotted against the vertical axis to show how differences between the SERC and USGS observations compare to the 95% confidence limits of the USGS loads.

model was calibrated using extensive empirical information from the Patuxent, so it is not surprising that it is among the best models for annual time series predictions of flow, TN, and TP (Table 13). Given its good performance for annual time series predictions, its much poorer performance for average annual predictions seems surprising. However, the explanation is simple. The two years predicted using the SERC model (August 1997-July 1999) are not typical and provide a poor estimate of average annual loads. This problem does not arise for the other models because they were designed to predict the annual average (MDP or SPARROW) or because they estimate the annual average from much longer annual time series (7–17 years, PLM, CBP, and SERCLM). In general, simulation periods need to be long enough to capture climatic variation, especially when calculating the long-term mean discharges most often used in applying watershed models to management questions.

Comparing Models as Published

We were not able to compare our models using exactly the same inputs and outputs. The models, we compared could not be fully standardized because they took very different approaches to modeling nutrient discharges from watersheds. For example, the models use fundamentally different kinds of information on land types and nutrient sources (Tables 6 and 7), and some models attribute nutrients directly to particular activities, such as fertilizer or manure application rather than using land type as a surrogate (Preston and Brakebill, 1999). Across the model ensemble, the land type data sets are not commensurate (see Weller *et al.*, 2003), so each model must be run with the land data for which it was developed. It would be wrong to apply a model calibrated with one type of land data to predict nutrient loads for watersheds described with a different source of land data (Weller *et al.*, 2003). The different choices of land data or nutrient generating activities to consider represent different ideas of how to represent nutrient sources, and the choices are essentially elements of model structure that cannot be eliminated. Similarly, we could not compare model outputs for exactly the same time periods. Some models (MDP, SPARROW) made only average annual predictions, not estimates for specific years. The SERC statistical model can only produce estimates during the years of its underlying empirical data (August 1997 through July 1999) and cannot make predictions for other years.

Our analysis, then, did not focus strictly on differences due to mathematical structure. Instead, we compared the predictions as published and as interpreted for management implications. This approach does demand some care in interpretation, such as the caution noted previously about annual average results based on different ranges of years, particularly, the short 2-year range of the SERC model. However, despite the limitations, comparing models as published is necessary and useful. All the models have been analyzed to make published inferences about the sources and magnitudes of nutrient loads, and some of the model predictions have been used to guide management decisions. We need to understand where the models agree and disagree, regardless of whether differences arise from mathematical structure or from differences in input data or time period considered. Considering all these differences provides the most complete information on uncertainty and confidence that the accumulated knowledge of all the models together can provide.

Multiple Models in Watershed Management

We must emphasize that the models we compared were designed for different purposes, had very different structures, and considered different geographic extents. Some of our models were for the Patuxent only (SERC, SERCLM, and PLM), one applied to the state of Maryland (MDP), and two modeled the entire Chesapeake Bay watershed (SPARROW and CBP), an area roughly 70 times larger than the Patuxent watershed. It is not surprising that empirical models customized for the Patuxent sometimes performed better in matching measured Patuxent discharges than did simulation models calibrated for the entire Chesapeake Bay watershed.

The CBP models have capabilities that none of the other models can provide, and those capabilities reflect its unique role in regional planning, decision making, and environmental regulation. The CBP models have been linked to a model of nutrient deposition from the atmosphere and to a model of estuarine circulation and ecological processes in the Bay. The combined modeling system estimates the maximum watershed nutrient loads that still support legally acceptable water quality in the Bay (the Chesapeake Bay TMDL, see USEPA, 2010b). The CBP model has been applied to partition the necessary watershed load reductions to states and local governments. The model accounts for a broad array of nutrient sources and watershed management actions, and it can predict the nutrient load reductions that might be achieved by different management alternatives. For these reasons, the CBP model is central to past and ongoing management and regulation in the Chesapeake Bay, and none of the other models in our ensemble could replace it in the regulatory process.

The existence of a dominant and very capable model, however, does not mean that a single model is sufficient and other models should be ignored. Our ensemble analysis demonstrates that there is much to be learned from comparing models, even for a system in which model development and application has focused strongly on a dominant model. Examining the range of estimates for common endpoints provided a way to quantify uncertainty in the model predictions, and averaging across all models generated more reliable estimates of flow and nutrient discharge than selecting and relying on any single model. Comparing the models also helped identify and correct problems and revealed gaps in scientific knowledge that require further research. Furthermore, considering a single model and not presenting its uncertainties can damage credibility if implemented strategies fail to meet the predicted outcomes (Breuer *et al.*, 2009). Ensemble modeling provides a robust, transparent mechanism for building public credibility (Leamer, 1983; Layton and Lee, 2006).

The Patuxent is a much studied watershed (see review in Weller *et al.*, 2003), so our ensemble analysis benefited from having published results available from many modeling programs. This reduced the cost of our analysis, overcoming one of the main limits on applying multiple models in watershed management (Pappenberger and Beven, 2006). However, analyzing the models as published also limited our ability to quantitatively attribute the differences among model predictions to the possible causes: differences in model structure, differences in the input data, or differences in the time period considered. Developing and sharing more standardized input data would improve ensemble modeling and its ability to inform science and management.

An ongoing program of ensemble modeling cannot be sustained by a single research group, but instead requires long-term funding and support of collaborative research (Jakeman *et al.*, 2006). In the Chesapeake Bay region, recent reports have called for adopting a multiple model approach to watershed management (Friedrichs *et al.*, 2011; STAC, 2011), and the proposal to create a Chesapeake Bay modeling laboratory (NRC, 2011; STAC, 2011) could provide the collaborative environment needed to support effective applications of multiple modeling.

Adaptive Management

The uncertainty documented by the large ranges among model predictions is definitely not an excuse for forgoing or delaying watershed management actions. The proper response to uncertainty is adaptive management (Boesch, 2002; Stankey *et al.*, 2005; Williams *et al.*, 2007), not inaction. In the case of the Chesapeake Bay, the legal mandate to address impairments is clear (USEPA, 2010c), major sources of nutrients are well known, and management actions that can reduce those sources have been identified. Proceeding through successive adaptive management cycles provides an effective way to move forward and address impairments in the face of uncertainties like those evident in our model comparison. As discussed above, using multiple models helps reinforce confidence in some model predictions and helps identify where additional monitoring or research is needed to reduce uncertainty and increase confidence. With these advantages, using multiple models can improve the outcomes of adaptive management more efficiently over time than relying on a single model (Williams *et al.*, 2007).

SUPPORTING INFORMATION

Additional results of our modeling analysis can be found in the online version of this article.

Data S1. Monthly Time Series Analysis. Methods. Results.

Figure S1. Time series predictions of monthly water, TN, and TP discharges from the Bowie basin *vs.* measured monthly discharges.

Please note: Neither AWRA nor Wiley-Blackwell is responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

ACKNOWLEDGMENTS

This research was funded by the National Oceanic and Atmospheric Administration Coastal Oceans Program (grant numbers NA66RG0129 and NA03NOS4780008), National Science Foundation (grant numbers BSR-9085219 and DEB-9317968), and the Smithsonian Institution Environmental Sciences Program. John Brakebill of the U.S. Geological Survey provided information and model output for the SPARROW models.

LITERATURE CITED

Alexander, R.B., J.W. Brakebill, R.E. Brew, and R.A. Smith, 1999. Enhanced River Reach File 1.2 (ERF1). http://water.usgs.gov/GIS/metadata/usgswrd/XML/erf1.xml, *accessed* July 2012.

Alexander, R.B., P.J. Johnes, E.W. Boyer, and R.A. Smith, 2002. A Comparison of Models for Estimating the Riverine Export of Nitrogen from Large Watersheds. Biogeochemistry 57:295-339.

Beaulac, M.N. and K.H. Reckhow, 1982. An Examination of Land-Use — Nutrient Export Relationships. Water Resources Bulletin 18:1013-1024.

Beven, K., 2007. Towards Integrated Environmental Models of Everywhere: Uncertainty, Data and Modelling as a Learning Process. Hydrology and Earth System Sciences 11:460-467.

Beven, K. and J. Kirby, 1979. A Physically-Based, Variable Contributing Area Model of Basin Hydrology. Hydrological Sciences Bulletin 24:43-69.

Bicknell, B.R., J.C. Imhoff, J.L. Kittle, A.S. Donigian, Jr., and R.C. Johanson, 2001. Hydrologic Simulation — Program FORTRAN (HSPF): User's Manual for Release 12. Aqua Terra Consultants, Mountain View, California.

Bloschl, G., 2006. Hydrologic Synthesis: Across Processes, Places, and Scales. Water Resources Research 42: W03S02, doi:10.1029/2005WR004319:1-3.

Boesch, D.F., 2002. Challenges and Opportunities for Science in Reducing Nutrient Over-Enrichment of Coastal Ecosystems. Estuaries 25:886-900.

Boomer, K.B., D.E. Weller, and T.E. Jordan, 2008. Empirical Models Based on the Universal Soil Loss Equation Fail to Predict Sediment Discharges from Chesapeake Bay Catchments. Journal of Environmental Quality 37:79-89.

Bormann, H., L. Breuer, T. Graff, and J.A. Huisman, 2007. Analysing the Effects of Soil Properties Changes Associated with Land Use Changes on the Simulated Water Balance: A Comparison of Three Hydrological Catchment Models for Scenario Analysis. Ecological Modelling 209:29-40.

Bormann, H., L. Breuer, T. Graff, J.A. Huisman, and B. Croke, 2009. Assessing the Impact of Land Use Change on Hydrology by Ensemble Modelling (LUCHEM) IV: Model Sensitivity to Data Aggregation and Spatial (Re-)distribution. Advances in Water Resources 32:171-192.

Brakebill, J.W. and S.D. Preston, 1999. Digital Data Used to Relate Nutrient Inputs to Water Quality in the Chesapeake Bay Watershed, Version 1.0. Report USGS OFR 99-60, 10 pp., U.S. Geological Survey.

Brakebill, J.W. and S.D. Preston, 2003. A Hydrologic Network Supporting Spatially Referenced Regression Modeling in the Chesapeake Bay Watershed. Environmental Monitoring and Assessment 81:73-84.

Brakebill, J.W. and S.D. Preston, 2004. Digital Data Used to Relate Nutrient Inputs to Water Quality in the Chesapeake Bay Watershed, Version 3.0. Report USGS OFR 2004-1433, 15 pp., U.S. Department of the Interior, U.S. Geological Survey.

Brakebill, J.W., S.D. Preston, and S.K. Martucci, 2001. Digital Data Used to Relate Nutrient Inputs to Water Quality in the Chesapeake Bay Watershed, Version 2.0. Report USGS OFR 01-251, 17 pp., U.S. Department of the Interior, U.S. Geological Survey.

Brandt, S.B. and D.M. Mason, 2003. Effect of Nutrient Loading on Atlantic Menhaden (Brevoortia tyrannus) Growth Rate Potential in the Patuxent River. Estuaries 26:298-309.

Breitburg, D.L., T.E. Jordan, and D. Lipton, 2003. Preface-From Ecology to Economics: Tracing Human Influence in the Patuxent River Estuary and Its Watershed. Estuaries 26:167-170.

Breuer, L. and J.A. Huisman, 2009. Assessing the Impact of Land Use Change on Hydrology by Ensemble Modeling (LUCHEM). Advances in Water Resources 32:127-128.

Breuer, L., J.A. Huisman, P. Willems, H. Bormann, A. Bronstert, B.F.W. Croke, H.G. Frede, T. Graff, L. Hubrechts, A.J. Jakeman, G. Kite, J. Lanini, G. Leavesley, D.P. Lettenmaier, G. Lindstrom, J. Seibert, M. Sivapalan, and N.R. Viney, 2009. Assessing the Impact of Land Use Change on Hydrology by Ensemble Modeling (LUCHEM). I: Model Intercomparison with Current Land Use. Advances in Water Resources 32: 129-146.

Cerco, C.F., 1995. Response of Chesapeake Bay to Nutrient Load Reductions. Journal of Environmental Engineering 121:549-556.

Clark, J.S., 2007. Models for Ecological Data: An Introduction. Princeton University Press, Princeton, New Jersey.

Cohn, T.S., D.L. Caulder, E.J. Gilroy, L.D. Zynjuk, and R.M. Summers, 1992. The Validity of a Simple Log-Linear Model for Estimating Fluvial Constituent Loads: An Empirical Study Involving Nutrient Loads Entering Chesapeake Bay. Water Resources Research 28:2353-2364.

Cohn, T.S., L.L. DeLong, E.J. Gilroy, R.M. Hirsch, and D.K. Wells, 1989. Estimating Constituent Loads. Water Resources Research 25:937-942.

Costanza, R., A.A. Voinov, R. Boumans, T. Maxwell, F. Villa, L. Wainger, and H. Voinov, 2002. Integrated Ecological Economic Modeling of the Patuxent River Watershed, Maryland. Ecological Monographs 72:203-231.

Crawford, N.H. and R.K. Linsley, 1966. Digital Simulation in Hydrology: Stanford Watershed Model IV. Report Technical Report No. 39, 210 pp., Department of Civil Engineering, Stanford University, Palo Alto, California.

Darrell, L.C.D., B.F. Majedi, J.S. Lizárraga, and J.D. Blomquist, 1998. Nutrient and Suspended-Sediment Concentrations, Trends, Loads, and Yields from the Nontidal Part of the Susquehanna, Potomac, Patuxent, and Choptank Rivers, 1985–96. Report USGS Water-Resources Report 98-4177, 38 pp., U.S. Geological Survey.

Dearing, J.A. and R.T. Jones, 2003. Coupling Temporal and Spatial Dimensions of Global Sediment Flux through Lake and Marine Sediment Records. Global and Planetary Change 39:147-168.

Devereux, O.H., K.L. Prestegaard, B.A. Needelman, and A.C. Gellis, 2010. Suspended-Sediment Sources in an Urban Watershed, Northeast Branch Anacostia River, Maryland. Hydrological Processes 24:1391-1403.

Dezetter, A., S. Girard, J.E. Paturel, G. Mahe, S. Ardoin-Bardin, and E. Servat, 2008. Simulation of Runoff in West Africa: Is There a Single Data-Model Combination that Produces the Best Simulation Results? Journal of Hydrology 354:203-212.

Donigian, Jr, A.S., B.R. Bicknell, A.S. Patwardhan, L.C. Linker, C.H. Chang, and R. Reynolds, 1994. Chesapeake Bay Program Watershed Model Application to Calculate Bay Nutrient Loadings – Final Facts and Recommendations. Modeling Subcommittee of the Chesapeake Bay Program Report CBP/TRS 57/96, EPA 903-R-94-042. U.S. Environmental Protection Agency, Annapolis, Maryland.

Duan, Q.Y., N.K. Ajami, X.G. Gao, and S. Sorooshian, 2007. Multi-Model Ensemble Hydrologic Prediction Using Bayesian Model Averaging. Advances in Water Resources 30:1371-1386.

EPA-EMAP (U.S. Environmental Protection Agency Environmental Monitoring and Assessment Program), 1994. Chesapeake Bay Watershed Pilot Project, U.S. EPA-EMAP, Research Triangle Park, North Carolina.

Fox, G.A. and G.V. Wilson, 2010. The Role of Subsurface Flow in Hillslope and Stream Bank Erosion: A Review. Soil Science Society of America Journal 74:717-733.

Friedrichs, M., C. Cerco, C. Friedrichs, R. Hood, D. Jasinski, W. Long, K. Sellner, and G. Shenk, 2011. Chesapeake Bay Hydrodynamic Modeling: A Workshop Report. STAC Publication 11-04. Chesapeake Research Consortium. Edgewater, Maryland.

Givens, G.H., 1999. Multicriterion Decision Merging: Competitive Development of an Aboriginal Whaling Management Procedure. Journal of the American Statistical Association 94:1003-1014.

Glibert, P.M., C.A. Heil, J.M. O'Neil, W.C. Dennison, and M.J.H. O'Donohue, 2006. Nitrogen, Phosphorus, Silica, and Carbon in Moreton Bay, Queensland, Australia: Differential Limitation of Phytoplankton Biomass and Production. Estuaries and Coasts 29:209-221.

Gneiting, T. and A.E. Raftery, 2005. Atmospheric Science — Weather Forecasting with Ensemble Methods. Science 310:248-249.

Goetz, S.J., C.A. Jantz, S.D. Prince, A.J. Smith, R. Wright, and D. Varlyguin, 2004. Integrated Analysis of Ecosystem Interactions with Land Use Change: The Cheapeake Bay Watershed. *In*: Ecosystems and Land Use Change, R.S. DeFries, G.P. Asner, and R.A. Houghton (Editors). American Geophysical Union, Washington, D.C., pp. 263-275.

Gordon, W.S., J.S. Famiglietti, N.L. Fowler, T.G.F. Kittel, and K.A. Hibbard, 2004. Validation of Simulated Runoff from Six Terrestrial Ecosystem Models: Results from VEMAP. Ecological Applications 14:527-545.

Gourley, J.J. and B.E. Vieux, 2005. A Method for Evaluating the Accuracy of Quantitative Precipitation Estimates from a Hydrologic Modeling Perspective. Journal of Hydrometeorology 6:115-133.

Gutierrez-Magness, A.L., J.E. Hannawald, L.C. Linker, and K.J. Hopkins, 1997. Chesapeake Bay Watershed Model Application and Calculation of Nutrient and Sediment Loadings, Appendix E. Report Report # EPA 903-R-97-019, 142 pp., U.S. Environmental Protection Agency Chesapeake Bay Program Office, Annapolis, Maryland.

Harmel, R.D., K.W. King, B.E. Haggard, D.G. Wren, and J.M. Sheridan, 2006. Practical Guidance for Discharge and Water Quality Data Collection on Small Watersheds. Transactions of the ASABE 49:937-948.

Hartzell, J.L., T.E. Jordan, and J.C. Cornwell, 2010. Phosphorus Burial in Sediments along the Salinity Gradient of the Patuxent River, a Subestuary of the Chesapeake Bay (USA). Estuaries and Coasts 33:92-106.

Homer, C., J. Dewitz, J. Fry, M. Coan, N. Hossain, C. Larson, N. Herold, A. McKerrow, J.N. VanDriel, and J. Wickham, 2007. Completion of the 2001 National Land Cover Database for the Conterminous United States. Photogrammetric Engineering and Remote Sensing 73:337-341.

Homer, C., C.Q. Huang, L.M. Yang, B. Wylie, and M. Coan, 2004. Development of a 2001 National Land-Cover Database for the United States. Photogrammetric Engineering and Remote Sensing 70:829-840.

Hsu, K.L., H. Moradkhani, and S. Sorooshian, 2009. A Sequential Bayesian Approach for Hydrologic Model Selection and Prediction. Water Resources Research 45:15.

Huisman, J.A., L. Breuer, H. Bormann, A. Bronstert, B.F.W. Croke, H.G. Frede, T. Graff, L. Hubrechts, A.J. Jakeman, G. Kite, J. Lanini, G. Leavesley, D.P. Lettenmaier, G. Lindstrom, J. Seibert, M. Sivapalan, N.R. Viney, and P. Willems, 2009. Assessing the Impact of Land Use Change on Hydrology by Ensemble Modeling (LUCHEM) III: Scenario Analysis. Advances in Water Resources 32:159-170.

Jakeman, A.J., R.A. Letcher, and J.P. Norton, 2006. Ten Iterative Steps in Development and Evaluation of Environmental Models. Environmental Modelling and Software 21:602-614.

Jordan, T.E., D.L. Correll, and D.E. Weller, 1997a. Effects of Agriculture on Discharges of Nutrients from Coastal Plain Watersheds of Chesapeake Bay. Journal of Environmental Quality 26:836-848.

Jordan, T.E., D.L. Correll, and D.E. Weller, 1997b. Nonpoint Source Discharges of Nutrients from Piedmont Watersheds of Chesapeake Bay. Journal of the American Water Resources Association 33:631-645.

Jordan, T.E., D.L. Correll, and D.E. Weller, 1997c. Relating Nutrient Discharges from Watersheds to Landuse and Streamflow Variability. Water Resources Research 33:2579-2590.

Jordan, T.E., D.E. Weller, and D.L. Correll, 2003. Sources of Nutrient Inputs to the Patuxent River Estuary. Estuaries 26:226-243.

Kadane, J.B. and N.A. Lazar, 2004. Methods and Criteria for Model Selection. Journal of the American Statistical Association 99:279-290.

Koop, G. and L. Tole, 2004. Measuring the Health Effects of Air Pollution: To What Extent Can We Really Say that People Are Dying from Bad Air? Journal of Environmental Economics and Management 47:30-54.

Langland, M.J., J.D. Blomquist, L.A. Sprague, and R.E. Edwards, 1999. Trends and Status of Flow, Nutrients, and Sediments for Selected Nontidal Sites in the Chesapeake Bay Watershed, 1985-98. Report U.S. Geological Survey Open File Report 99-451, 64 pp., U.S. Department of Interior, Lemoyne, Pennsylvania.

Langland, M.J., P.L. Lietman, and S. Hoffman, 1995. Synthesis of Nutrient and Sediment Data for Watersheds within the Chesapeake Bay Drainage Basin. U.S. Geological Survey. Lemoyne, Pennsylvania.

Layton, D.F. and S.T. Lee, 2006. Embracing Model Uncertainty: Strategies for Response Pooling and Model Averaging. Environmental and Resource Economics 34:51-85.

Leamer, E.E., 1983. Let's Take the Con Out of Econometrics. The American Economic Review 73:31-43.

Linker, L.C., G.W. Shenk, R.L. Dennis, and J.S. Sweeny, 2000. Cross-Media Models of the Chesapeake Bay Watershed and Airshed. Water Quality and Ecosystem Modeling 1:91-122.

Linker, L.C., G.W. Shenk, P. Wang, and R. Batiuk, 2008. Integration of Modeling, Research, and Monitoring in the Chesapeake Bay Program. *In*: The Management of Water Quality and Irrigation Technologies, J. Albiac and A. Dinar (Editors). Earthscan, London, United Kingdom, pp. 41-60.

Linker, L.C., C.G. Stigall, C.H. Chang, and A.S. Donigian, Jr., 1996. Aquatic Accounting: Chesapeake Bay Watershed Model Quantifies Nutrient Loads. Water Environment and Technology 8:48-52.

Liu, Z.J. and D.E. Weller, 2008. A Stream Network Model for Integrated Watershed Modeling. Environmental Modeling and Assessment 13:291-303.

Liu, Z.-J., D.E. Weller, D.L. Correll, and T.E. Jordan, 2000. Effects of Land Cover and Geology on Stream Chemistry in Watersheds of Chesapeake Bay. Journal of the American Water Resources Association 36:1349-1365.

Liu, Z.J., D.E. Weller, T.E. Jordan, D.L. Correll, and K.B. Boomer, 2008. Integrated Modular Modeling of Water and Nutrients from Point and Nonpoint Sources in the Patuxent River Watershed. Journal of the American Water Resources Association 44:700-723.

Lung, W.S. and S. Bai, 2003. A Water Quality Model for the Patuxent Estuary: Current Conditions and Predictions under Changing Land-Use Scenarios. Estuaries 26:267-279.

Lung, W.S. and A.J. Nice, 2007. Eutrophication Model for the Patuxent Estuary: Advances in Predictive Capabilities. Journal of Environmental Engineering-ASCE 133:917-930.

Maxwell, T. and R. Costanza, 1997. A Language for Modular Spatio-Temporal Simulation. Ecological Modelling 103: 105-113.

McIntyre, N., H. Lee, H. Wheater, A. Young, and T. Wagener, 2005. Ensemble Predictions of Runoff in Ungauged Catchments. Water Resources Research 41, W03202, W12434, doi:10.1029/2005WR004289, pp. 1-14.

MDP (Maryland Department of Planning), 2003a. 1973 Land Use/Land Cover for Maryland. http://planning.maryland.gov/OurWork/landUseDownload.shtml, *accessed* July 2012.

MDP (Maryland Department of Planning), 2003b. 2002 Land Use/Land Cover for Maryland. http://planning.maryland.gov/OurWork/landUseDownload.shtml, *accessed* July 2012.

MDP (Maryland Department of Planning), MDE (Maryland Department of the Environment) and DNR (Maryland Department of Natural Resources), 2007. Water Resources Element of the Comprehensive Plan Guidance Document Planning for Water Supply. Wastewater Management and Stormwater Management, Baltimore, Maryland, 84 pp.

Miller, R.C., D.P. Guertin, and P. Heilman, 2004. Information Technology in Watershed Management Decision Making. Journal of the American Water Resources Association 40:347-357.

MOP (Maryland Office of Planning), 1993. Nonpoint Source Assessment and Accounting System: Final Report for the FFY, '91 Section 319 Grant. Maryland Office of Planning, Baltimore, Maryland.

MOP (Maryland Office of Planning), 1995. Development and Application of the Nonpoint Source Assessment and Accounting System: Final Report for the FFY, '92 Section 319 Grant. Maryland Office of Planning, Baltimore, Maryland.

Moriasi, D.N., J.G. Arnold, M.W. Van Liew, R.L. Bingner, R.D. Harmel, and T.L. Veith, 2007. Model Evaluation Guidelines for Systematic Quantification of Accuracy in Watershed Simulations. Transactions of the Asabe 50:885-900.

Mukundan, R., D.E. Radcliffe, J.C. Ritchie, L.M. Risse, and R.A. McKinley, 2010. Sediment Fingerprinting to Determine the Source of Suspended Sediment in a Southern Piedmont Stream. Journal of Environmental Quality 39:1328-1337.

Nash, J.E. and J.V. Sutcliff, 1970. River Flow Forecasting Through Conceptual Models Part I — A Discussion of Principles. Journal of Hydrology 10:282-290.

NOAA-CCAP (National Oceanographic and Atmospheric Administration Coastal Change and Analysis Program), 2006. Coastal Change Analysis Program Regional Land Cover. http://www.csc.noaa.gov/digitalcoast/data/ccapregional/, *accessed* July 2012.

Noe, G.B. and C.R. Hupp, 2009. Retention of Riverine Sediment and Nutrient Loads by Coastal Plain Floodplains. Ecosystems 12:728-746.

NRC (National Research Council), 2011. Achieving Nutrient and Sediment Reduction Goals in the Chesapeake Bay: An Evaluation of Program Strategies and Implementation for Nutrient Reduction to Improve Water Quality. The National Academies Press. Washington, D.C. (Prepublication copy).

Pappenberger, F. and K.J. Beven, 2006. Ignorance Is Bliss: Or Seven Reasons Not to Use Uncertainty Analysis. Water Resources Research 4:2.

Phillips, T.J. and P.J. Gleckler, 2006. Evaluation of Continental Precipitation in 20th Century Climate Simulations: The Utility of Multimodel Statistics. Water Resources Research 42, W03202, doi:10.1029/2005WR004313, pp. 1-10.

Preston, S.D. and J.W. Brakebill, 1999. Application of Spatially Referenced Regression Modeling for the Evaluation of Total Nitrogen Loading in the Chesapeake Bay Watershed. Report WRIR-99-4054, pp. 1-12, U.S. Geological Survey, Baltimore, Maryland.

Prosser, I., P. Rustomji, B. Young, C. Moran, and A. Hughes, 2001. Constructing River Basin Sediment Budgets for the National Land and Water Resources Audit. Report Technical Report Number 15/01, pp. 1-34, CSIRO Land and Water, Canberra, Australia.

Pyke, C., 2010. Review of Land-Use and Land Cover Dataset and Methodology. STAC Publication, Chesapeake Research Consortium, Inc., Edgewater, Maryland.

Radcliffe, D.E., J. Freer, and O. Schoumans, 2009. Diffuse Phosphorus Models in the United States and Europe: Their Usages, Scales, and Uncertainties. Journal of Environmental Quality 38:1956-1967.

Redfield, A.C., 1958. The Biological Control of Chemical Factors in the Environment. American Scientist 46:205-221.

Sivakumar, B., 2008. The More Things Change, the More They Stay the Same: The State of Hydrologic Modelling. Hydrological Processes 22:4333-4337.

Smith, R.A., G.E. Schwarz, and R.B. Alexander, 1997. Regional Interpretation of Water-Quality Monitoring Data. Water Resources Research 33:2781-2798.

Smith, V.H., S.B. Joye, and R.W. Howarth, 2006. Eutrophication of Freshwater and Marine Ecosystems. Limnology and Oceanography 51:351-355.

Sprent, P. and G.R. Dolby, 1980. The Geometric Mean Functional Relationship. Biometrics 36:547-550.

STAC (Scientific and Technical Advisory Committee), 2011. Review of the LimnoTech Report, Comparison of Load Estimates for Cultivated Cropland in the Chesapeake Bay Watershed. Committee for the ANPC/LimnoTech Review, Scientific and Technical Advisory Committee to the Chesapeake Bay Program. STAC Publication 11-02, Chesapeake Research Consortium, Inc., Edgewater, Maryland.

Stankey, G.H., R.N. Clark, and B.T. Bormann, 2005. Adaptive Management of Natural Resources: Theory, Concepts, and Management Institutions. Gen. Tech. Rep. PNW-GTR-654, 73 pp., U.S. Department of Agriculture, Forest Service, Pacific Northwest Research Station, Portland, Oregon.

Staver, K.W. and R.B. Brinsfield, 2001. Agriculture and Water Quality on the Maryland Eastern Shore: Where Do We Go from Here? BioScience 51:859-868.

Stone, K.C., P.G. Hunt, J.M. Novak, M.H. Johnson, and D.W. Watts, 2000. Flow-Proportional, Time-Composited, and Grab

Sample Estimation of Nitrogen Export from an Eastern Coastal Plain Watershed. Transactions of the American Society of Agricultural Engineers 43:281-290.

Tassone, J., D.M. Weller, R.E. Hall, and N.M. Edwards, 1998. Smart Growth Options for Maryland's Tributary Strategies. Maryland Office of Planning, Baltimore, Maryland.

Tele Atlas, 2004. Dynamap 2000 Road Network Database. Spatial Insights, Inc., Bethesda, Maryland.

Tobias, J.L. and M. Li, 2004. Returns to Schooling and Bayesian Model Averaging: A Union of Two Literatures. Journal of Economic Surveys 18:153-180.

U.S. Department of Commerce and U.S. Census Bureau, 2005. TIGER/Line Files, 2005 (First Edition). http://www.census.gov/geo/www/tiger/shp.html, *accessed* July 2012.

USEPA (U.S. Environmental Protection Agency), 2010a. Chesapeake Bay Phase 5 Community Watershed Model. http://ches.communitymodeling.org/models/CBPhase5/index.php, *accessed* April 2011.

USEPA (U.S. Environmental Protection Agency), 2010b. Chesapeake Bay Total Maximum Daily Load for Nitrogen, Phosphorus and Sediment, U.S. Environmental Protection Agency, Annapolis, Maryland.

USEPA (U.S. Environmental Protection Agency), 2010c. Chesapeake Bay Total Maximum Daily Load for Nitrogen, Phosphorus and Sediment. http://www.epa.gov/reg3wapd/pdf/pdf_chesbay/FinalBayTMDL/CBayFinalTMDLExecSumSection1through3_final.pdf, *accessed* April 4, 2011.

USGS (U.S. Geological Survey), 1999. USGS National Hydrography Data Base (NHD): U.S. Geological Survey Fact Seet 106-99. http://egsc.usgs.gov/isb/pubs/factsheets/fs10699.html, *accessed* July 15, 2010.

USGS (U.S. Geological Survey), 2011a. USGS Water Data for the Nation. waterdata.usgs.gov/nwis, *accessed* February 2009.

USGS (U.S. Geological Survey), 2011b. Water Resources Applications Software — Summary of HSPF. http://water.usgs.gov/cgi-bin/man_wrdapp?hspf, *accessed* November 2011.

Valiela, I., J.L. Bowen, and K.D. Kroeger, 2002. Assessment of Models for Estimation of Land-Derived Nitrogen Loads to Shallow Estuaries. Applied Geochemistry 17:935-953.

de Vente, J. and J. Poesen, 2005. Predicting Soil Erosion and Sediment Yield at the Basin Scale: Scale Issues and Semi-Quantitative Models. Earth-Science Reviews 71:95-125.

Viney, N.R., H. Bormann, L. Breuer, A. Bronstert, B.F.W. Croke, H. Frede, T. Graff, L. Hubrechts, J.A. Huisman, A.J. Jakeman, G.W. Kite, J. Lanini, G. Leavesley, D.P. Lettenmaier, G. Lindstrom, J. Seibert, M. Sivapalan, and P. Willems, 2009. Assessing the Impact of Land Use Change on Hydrology by Ensemble Modelling (LUCHEM) II: Ensemble Combinations and Predictions. Advances in Water Resources 32:147-158.

Vogelmann, J.E., S.M. Howard, L. Yang, C.R. Larson, B.K. Wylie, and N. Van Driel, 2001. Completion of the 1990's National Land Cover Data Set for the Conterminous United States from Landsat Thematic Mapper Data and Ancillary Data Sources. Photogrammetric Engineering and Remote Sensing 67:650-662.

Voinov, A., R. Costanza, C. Fitz, and T. Maxwell, 2007. Patuxent Landscape Model: 2. Model Development — Nutrients, Plants, and Detritus. Water Resources 34:268-276.

Voinov, A., C. Fitz, R. Boumans, and R. Costanza, 2004. Modular Ecosystem Modeling. Environmental Modelling and Software 19:285-304.

Voinov, A.A., H. Voinov, and R. Costanza, 1999. Surface Water Flow in Landscape Models: 2. Patuxent Watershed Case Study. Ecological Modelling 119:211-230.

Vrugt, J.A. and B.A. Robinson, 2007. Treatment of Uncertainty Using Ensemble Methods: Comparison of Sequential Data Assimilation and Bayesian Model Averaging. Water Resources Research 43, W01411, doi:10.1029/2005WR004838, pp. 1-15.

Walter, R.C. and D.J. Merritts, 2008. Natural Streams and the Legacy of Water-Powered Mills. Science 319:299-304.

Weller, D.E., T.E. Jordan, D.L. Correll, and Z.J. Liu, 2003. Effects of Land-Use Change on Nutrient Discharges from the Patuxent River Watershed. Estuaries 26:244-266.

Wells, R.R., C.V. Alonso, and S.J. Bennett, 2009. Morphodynamics of Headcut Development and Soil Erosion in Upland Concentrated Flows. Soil Science Society of America Journal 73:521-530.

Wilkinson, S.N., I.P. Prosser, P. Rustomji, and A.M. Read, 2009. Modelling and Testing Spatially Distributed Sediment Budgets to Related Erosion Processes to Sediment Yields. Environmental Modelling and Software 24:489-501.

Williams, B.K., R.C. Szaro, and C.C. Shapiro, 2007. Adaptive Management: The U.S. Department of the Interior Technical Guide. Adaptive Management Working Group, US Department of Interior, Washington, D.C.

Wright, J., 2003. Bayesian Model Averaging and Exchange Rate Forecasts. Journal of Econometrics 146:329-341.

Supporting Information

USING MULTIPLE WATERSHED MODELS TO PREDICT WATER, NITROGEN, AND PHOSPHORUS DISCHARGES TO THE PATUXENT ESTUARY

Kathleen M.B. Boomer, Donald E. Weller, Thomas E. Jordan, Lewis Linker, Zhi-Jun Liu, James Reilly, Gary Shenk, and Alexey A, A. Voinov

Respectively, Ecologist, Senior Ecologist, and Senior Ecologist, Smithsonian Environmental Research Center, 647 Contees Wharf Road, Edgewater, Maryland 21037-0028; Modeling Coordinator, U. S. Environmental Protection Agency Chesapeake Bay Program, Annapolis, MD 21403; Associate Professor, Department of Geography, University of North Carolina, Greensboro, NC 27402-6170; Planner, Maryland Department of Planning, Baltimore, MD 21201 [Reilly now at Reilly Consulting, Lafayette Hill, PA 19444]; Integrated Analysis Coordinator, U. S. Environmental Protection Agency Chesapeake Bay Program, Annapolis, MD 21403; and Associate Professor, The Gund Institute for Ecological Economics, University of Vermont, Burlington, Vermont 05405 [Voinov now at International Institute for Geo-information Science and Earth Observation, P.O. Box 6, 7500 AA Enschede, The Netherlands] (E-Mail/Boomer: boomerk@si.edu).

Monthly Time Series.

Methods. We summarized monthly time series predictions of Bowie discharges for the SERC, CBP, and PLM implementations. We plotted monthly discharge estimates against observed values and calculated Nash-Sutcliffe values of model performance (Eq. 2, but with *i* indexing months rather than years).

Results

Results. Five models could predict monthly loads at Bowie. The best model again varied by constituent (Figure S1), and the patterns of model performance were different from the average annual and annual time series patterns. The PLM model best predicted flow (NS = 1), although all models (except CBP4, which overestimated monthly flow) produced good predictions (NS > 0.8). In contrast, the PLM was least reliable (NS = -0.58) for predicting monthly TN loads, and the SERC statistical model best predicted TN loads (NS = 0.82). For TP, the CBP4 and CBP5 models performed best compared to the other models, though only slightly better than simply using the mean observed TP (NS = 0.07 and 0.09, respectively). The models generally over-predicted monthly TP, especially during high discharge months. In particular, the SERC model consistently over-predicted TP loads (NS = -3.62), but the predictions were strongly correlated with USGS observed TP loads (R^2 = 0.92). The systematically higher monthly TP load estimates from the SERC and SERCLM models can be traced to a difference in TP measurements used to calibrate the SERC models and the USGS TP measurements (see Discussion).

FIGURE S1. Time series predictions of monthly water, TN, and TP discharges from the Bowie basin versus measured monthly discharges. The numbers are Nash-Sutcliffe (NS) performance coefficients. Predictions would equal observations along the diagonal line. The dotted lines on the TN and TP plots are the 95% confidence limits for the USGS estimates of observed loads. The confidence limits are plotted along the vertical axis to show how differences between model predictions and observed loads compare to the 95% confidence limit of the observed loads.