Comparison of Estuarine Water Quality Models for Total Maximum Daily Load Development in Neuse River Estuary

Craig A. Stow[1]; Chris Roessler[2]; Mark E. Borsuk[3]; James D. Bowen[4]; and Kenneth H. Reckhow[5]

Abstract: The North Carolina Division of Water Quality developed a total maximum daily load (TMDL) to reduce nitrogen inputs into the Neuse River Estuary to address the problem of repeated violations of the ambient chlorophyll *a* criterion. Three distinct water quality models were applied to support the TMDL: a two-dimensional laterally averaged model, a three-dimensional model, and a probability (Bayesian network) model. In this paper, we compare the salient features of all three models and present the results of a verification exercise in which each calibrated model was used to predict estuarine chlorophyll *a* concentrations for the year 2000. We present six summary statistics to relate the model predictions to the observed chlorophyll values: (1) the correlation coefficient; (2) the average error; (3) the average absolute error; (4) the root mean squared error; (5) the reliability index; and (6) the modeling efficiency. Additionally, we examined each model's ability to predict how frequently the 40 μg/L chlorophyll *a* criterion was exceeded. The results indicate that none of the models predicted chlorophyll concentrations particularly well. Predictive accuracy was no better in the more process-oriented, spatially detailed models than in the aggregate probabilistic model. Our relative inability to predict accurately, even in well-studied, data-rich systems underscores the need for adaptive management, in which management actions are recognized as whole-ecosystem experiments providing additional data and information to better understand and predict system behavior.

DOI: 10.1061/(ASCE)0733-9496(2003)129:4(307)

CE Database subject headings: Estuaries; Water quality; North Carolina; Comparative studies.

Introduction

There is an ongoing debate among water quality modelers concerning the attributes of models that make them useful as tools for decision support. For example, many modelers argue that because nature is complex, models must also be complex, and spatially and temporally detailed representation is necessary. An alternative view is that nature is too complex to ever be fully described with a set of differential equations, so simple, probabilistic models are a better approach.

[1]Duke Univ., Nicholas School of the Environment and Earth Sciences, Environmental Science and Policy Division, Durham, NC 27708; Address as of August 2003, Dept. of Environmental Health Sciences, Arnold School of Public Health, Univ. of South Carolina, Columbia, SC 29208. E-mail: cstow@duke.edu

[2]North Carolina Dept. of Environment and Natural Resources, Division of Water Quality, Raleigh, NC 27699.

[3]Duke Univ., Nicholas School of the Environment and Earth Sciences, Environmental Science and Policy Division, Durham, NC 27708; presently, Dept. of Systems Analysis, Integrated Assessment, and Modeling (SIAM), Swiss Federal Institute for Environmental Science and Technology (EAWAG), 8600 Dübendorf, Switzerland.

[4]Civil Engineering Dept., Univ. of North Carolina at Charlotte, Charlotte, NC 28223.

[5]Duke Univ., Nicholas School of the Environment and Earth Sciences, Environmental Science and Policy Division, Durham, NC 27708.

Note. Discussion open until December 1, 2003. Separate discussions must be submitted for individual papers. To extend the closing date by one month, a written request must be filed with the ASCE Managing Editor. The manuscript for this paper was submitted for review and possible publication on June 13, 2002; approved on August 2, 2002. This paper is part of the ***Journal of Water Resources Planning and Management***, Vol. 129, No. 4, July 1, 2003. ©ASCE, ISSN 0733-9496/2003/4-307–314/$18.00.

In addition to choosing a level of process representation, water quality modelers must consider the applicability of the model to the pending policy decision, as expressed in questions such as: "Does the model address the environmental attributes upon which the decision is to be based?", "Can it be supported by available data?", and "Does it facilitate conceptual understanding of the system and identification of major uncertainties?"

In this paper, we compare three different estuarine nutrient response models with respect to the above issues. While it is unlikely that this comparison will resolve the issue of model selection, we do believe that it provides a significant opportunity to critically examine three modeling approaches. This model comparison opportunity arose from the development of the nitrogen total maximum daily load (TMDL) for the Neuse River Estuary in North Carolina. The three models that have been applied to the Neuse Estuary for this purpose are

- A two-dimensional, laterally averaged process model referred to as the Neuse Estuary Eutrophication Model (NEEM) (see Bowen 2003),
- A three-dimensional Water Analysis Simulation Program (WASP) application (see Wool and Davie 2003),
- A Bayesian probability network model named the Neuse Estuary Bayesian Ecological Response Network (Neu-BERN) (see Borsuk et al. 2003).

The two-dimensional model (NEEM) is based on CE-QUAL-W2 (Cole and Buchak 1995) and represents a conventional mass-balance simulation approach in which daily values of response variables are predicted on a two-dimensional (longitudinal and vertical) grid. The NEEM was the first model to be supported by the North Carolina Division of Water Quality (NCDWQ) for assessing the nitrogen TMDL, in part because a previous version of CE-QUAL-W2 had already been calibrated to the Neuse Estuary (Bowen 2000). The three-dimensional model

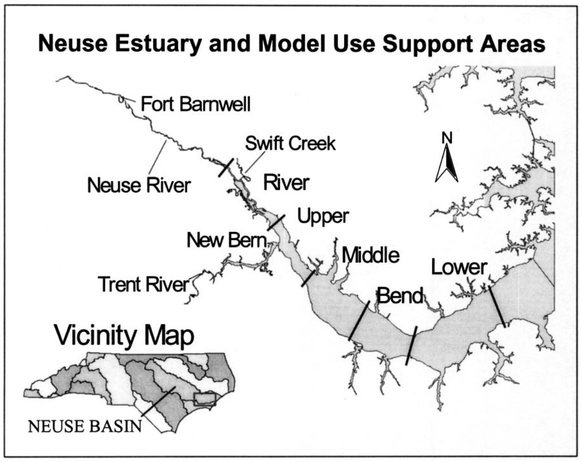

Fig. 1. Map of Neuse River watershed and estuary. Lines across estuary depict boundaries of five modeled sections.

(WASP) was applied to the Neuse Estuary by scientists working in United States Environmental Protection Agency (USEPA) Region IV. This modeling effort was undertaken in expectation of a significant cross-channel component to estuarine dynamics and phytoplankton growth. Both WASP and CE-QUAL-W2 are currently supported by the USEPA for TMDL development (Office of Water 1997a). The Bayesian network model (Neu-BERN), on the other hand, was initiated as an academic study to evaluate whether graphical, probabilistic models could be successfully applied to water quality management decisions. The Neu-BERN model characterizes aggregate spatial and temporal behavior. Consequently, predictions generated by the model are best interpreted as representing a range of expected system behaviors and not exact conditions at a specific time and location.

The three models applied to TMDL development for the Neuse Estuary represent three different perspectives, and we hope that a comparison of their scientific and decision support characteristics can provide some general insights for future model selection and development.

Problem Background—Eutrophication in the Neuse River Estuary, North Carolina

The Neuse River Estuary (Fig. 1) has experienced symptoms of eutrophication, including algal blooms, fishkills, and extensive hypoxia and anoxia, for a long time. Tales of fishkills trace back into the early 20th century, though systemic water quality problems are apparently more recent. Hobbie and Smith (1975) present monitoring data from the early 1970s, documenting high-nutrient levels, algal blooms, and occasional hypolimnetic hypoxia, suggesting that eutrophication symptoms were at least a periodic problem before regular ambient monitoring began in the late 1970s. Eutrophication became a more severe water quality concern in the lower Neuse River and estuary in the late 1970s and early 1980s with the proliferation of nuisance algal blooms (Paerl 1983; Christian et al. 1986; Paerl 1987). More recently, symptoms including fishkills and toxic microorganisms have dominated public attention.

North Carolina's 303(d) list, or list of impaired waters, has included the Neuse Estuary since at least 1990 and identifies chlorophyll *a* as the water quality characteristic that violates an established criterion of 40 $\mu g\,L^{-1}$. Violations are most common in the oligohaline sections above the midestuary bend (Fig. 1). Because the estuary appears on the 303(d) list, a TMDL is required for the responsible pollutant (Office of Water 1999). Nitrogen has been identified as the primary factor limiting algal biomass (as measured by chlorophyll *a*) (Paerl and Bowles 1987; Rudek et al. 1991; Paerl et al. 1995); consequently, the TMDL is being established for total nitrogen.

The North Carolina Division of Water Quality (NCDWQ) has taken a phased approach to developing a nitrogen management strategy for the estuary. In 1997, NCDWQ developed the Neuse Nutrient Sensitive Waters Management Strategy (commonly referred to as the "Neuse Rules") that sought to reduce total nitrogen loading to the Neuse Estuary by the year 2003. The targeted nitrogen load reduction under the Neuse Rules was 30% relative to 1991–1995 levels. This was based on the best professional judgment of a panel of scientists, evaluation of historical trend analyses, and results of algal dilution bioassays. The Neuse Rules were not a TMDL. They were an interim strategy, initiated at the state level, for achieving nitrogen load reductions to the estuary.

Subsequently, as the result of a lawsuit settlement, the USEPA required NCDWQ to develop a TMDL for the Neuse Estuary. NCDWQ complied with this request, and a TMDL requiring a 30% reduction relative to 1991–1995 was conditionally approved by USEPA in July, 1999 (NCDENR 1999). The 1999 TMDL was approved with the agreement that, by 2001, NCDWQ would complete a second phase of the TMDL based on more detailed analysis. The 2001 TMDL would make use of the extensive monitoring and modeling conducted between 1996 and 2000 as part of the Neuse River Estuary Modeling and Monitoring Project (ModMon). ModMon has been the collaborative effort of a number of scientists from several universities [University of North Carolina (UNC) Chapel Hill, including the Institute of Marine Sciences (IMS), UNC-Charlotte, Duke, East Carolina, and N.C. State], government agencies (National Marine Fisheries, N.C. Division of Water Quality, U.S. Geological Survey), and private industry (Weyerhaeuser Corporation). Both the NEEM and Neu-BERN model development were supported under the ModMon project.

The goal of the Phase II TMDL for the Neuse Estuary is compliance with the North Carolina chlorophyll *a* criterion of 40 $\mu g\,L^{-1}$. Based on USEPA guidance for use support determination (Office of Water 1997b), compliance is achieved if violation of the criterion occurs in fewer than 10% of the samples collected in a specified area and time. The NCDWQ has defined five estuarine sections for this purpose (Fig. 1), each of which is to be evaluated periodically. The period 1991–1995 continues to be the baseline against which future load reductions will be measured, as this is the period when eutrophication became increasingly evident and led to the development of the Neuse Rules.

Model Development

Much of the planning and assessment that led to the Phase II TMDL decision focused on model development. The NCDWQ hosted a series of meetings that included representatives of a number of groups, including: model developers, scientists from ModMon, a subset of stakeholders from a broader TMDL stakeholder panel, and regulatory personnel from USEPA-Region IV and NCDWQ. The objectives of these meetings were to ensure comparability among the three models, collaboratively resolve any modeling or data difficulties, and provide an open forum for sharing model development activities with regulatory personnel and stakeholders.

While model comparison is of longer-term academic interest, it is also relevant to near-term TMDL assessment; if one of the three models clearly performs best, then the predictions from that model should carry greatest weight in TMDL decision making. Thus, to the extent possible, the models employed the same input data and provided predictions at the same spatial and temporal scale. Because NEEM and WASP were constructed similarly, their developers were able to take additional steps to ensure that prediction differences were not due to differences in assumptions (e.g., they used the same boundary conditions).

NEEM and WASP are similar in that both are deterministic, mass balance simulation models that track an assortment of water quality variables related to eutrophication. The key difference between the two models is that WASP has been configured for compartments in three dimensions, rather than two for NEEM. In addition to having longitudinal segments and vertical layers, WASP divides the estuary laterally (e.g., from bank to bank). In other words, while NEEM aggregates predictions in the lateral dimension, WASP makes spatially explicit predictions. See Bowen (2003) and Wool and Davie (2003) for more details on these models.

The Bayesian network model, Neu-BERN, expresses linkages among variables as conditional probabilities rather than deterministic relationships. Thus, natural variability and uncertainty are explicitly included in the analysis and characterized probabilistically. One feature of a Bayesian network is its decomposability, which allows the conditional probabilities in the model to be estimated using separate submodels (Reckhow 1999). While Neu-BERN was primarily developed to focus on broader ecological endpoints (Borsuk et al. 2003), one component was structured to predict algal response (chlorophyll *a*) to nitrogen inputs. Unlike the other two models, which predict chlorophyll concentrations at particular points in space, the Neu-BERN model assumes that the variation among individual measurements within each estuary section (Fig. 1) on each day result from an unpredictable (stochastic), rather than predictable (systematic), process that can be best represented by a probability distribution. The characteristics of this distribution are predicted as a function of (1) Neuse River flow, (2) Neuse River total N concentration, (3) estuarine water temperature, (4) location in the estuary, and (5) the institution collecting the chlorophyll sample. This last term was included to account for consistent differences that were observed between the chlorophyll samples collected by the two sampling institutions. Samples collected by Weyerhaeuser were found to be 29% higher on average than those collected by UNC-IMS. The reason for this difference could not be specifically identified, but may relate to slight differences in sample collection depth (see Borsuk et al. 2002a for details).

The data necessary to develop the response models differed with each model, particularly between the two process models (NEEM and WASP), and the probability network model (Neu-BERN). However, all used riverine loading data and some estuary calibration data. The loading data came from NCDWQ ambient monitoring stations on the Neuse River (at Fort Barnwell), the Trent River (at Pollocksville), and Swift Creek (near New Bern). Fort Barnwell is the primary station for monitoring riverine load, as discharge is continuously measured there, and 85% of the Neuse watershed drains through this point. The estuary data used to develop the models primarily came from the ModMon project. The key dataset that this project provided was the weekly midchannel chemical and physical sampling conducted by Weyerhaeuser and UNC-IMS. The measured variables included the typical suite of physical and chemical constituents associated with eutrophication assessments. The NEEM and WASP models used data on water surface elevation, freshwater inflow, meteorological condition (wind, solar radiation, cloud cover), salinity, water temperature, nitrogen and phosphorus, chlorophyll, and dissolved oxygen for calibration, while Neu-BERN required flow, river nitrogen, chlorophyll, and water temperature data. The requirement of NEEM and WASP for downstream water elevation data precluded their application to periods before mid-1997 when the downstream water level recorder was installed. Additional data on similar variables at seven side-channel locations in the estuary were collected by researchers at North Carolina State University (NCSU). These data were used in calibrating WASP but were not available for development of Neu-BERN or NEEM.

NEEM was calibrated using the following objectives for guidance: (1) to match the cumulative distribution function of the observed and predicted values as closely as possible; (2) to maximize the R^2 for the total model (entire estuary); and (3) to achieve a mean error of zero. EPA calibrated WASP with the following objectives in mind: (1) to parametrize the model to best represent the nutrient gradients in the river/estuary; (2) to account for seasonal variability; and (3) to predict the chlorophyll concentrations in time and space. EPA judged these objectives using qualitative (best professional judgment of visual plots) and quantitative (unspecified statistical measures) methods (USEPA 2001). Calibrating the algal component of the Neu-BERN model involved statistically fitting a nonlinear equation to the available monitoring data using nonlinear least-squares and analogous Bayesian methods (Borsuk et al. 2002a).

Model Verification Methods

In our experience, it is common for modelers to calibrate a model to a set of observed data, visually compare some time-series plots of predictions and observations, declare the model to be ‘‘reasonable,’’ and proceed to use the model for inference or decision-support. Model verification or confirmation (Reckhow and Chapra 1983)—a comparison of model predictions with observations that represent conditions distinct from those represented by the calibration data—is often omitted or done in a perfunctory manner. In this modeling exercise, all three models were calibrated using data collected prior to the year 2000. Each model was then used to predict surface chlorophyll concentrations in 2000, and these predictions were then compared to the year 2000 measured values. Model predictions for NEEM consisted of the chlorophyll values predicted to occur at noon on the day of sample collection in the model grid cell corresponding to the sample location. WASP predictions consisted of the chlorophyll values predicted for the same time and location of sample collection. Predictions of Neu-BERN consist of full probability distributions, rather than single values. However, for model comparison, the median of this distribution was chosen as the point prediction corresponding to all midchannel sampling locations within each section for each sampling day.

A variety of model fit statistics are available for evaluating model performance (Reckhow et al. 1990). For this comparison, we calculated the following summary statistics for each model:

1. r—the correlation coefficient of the model predictions and observations

$$r=\frac{\sum_{i=1}^{n}(O_i-\overline{O})(P_i-\overline{P})}{\sqrt{\sum_{i=1}^{n}(O_i-\overline{O})^2\sum_{i=1}^{n}(P_i-\overline{P})^2}}$$

2. RMSE—the root mean squared error

$$\mathrm{RMSE}=\sqrt{\frac{\sum_{i=1}^{n}(P_i-O_i)^2}{n}}$$

3. RI—the reliability index

$$\mathrm{RI}=\exp\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\log\frac{O_i}{P_i}\right)^2}$$

4. AE—the average error

$$\mathrm{AE}=\frac{\sum_{i=1}^{n}(P_i-O_i)}{n}=\bar{P}-\bar{O}$$

5. AAE—the average absolute error

$$\mathrm{AAE}=\frac{\sum_{i=1}^{n}|P_i-O_i|}{n}$$

6. MEF—the modeling efficiency

$$\mathrm{MEF}=\frac{\left(\sum_{i=1}^{n}(O_i-\bar{O})^2-\sum_{i=1}^{n}(P_i-O_i)^2\right)}{\sum_{i=1}^{n}(O_i-\bar{O})^2}$$

where n=number of observations; O_i=ith of n observations; P_i=ith of n predictions; and $\bar{O}$ and $\bar{P}$=observation and prediction averages, respectively. All observations and predictions were log-transformed before calculating the fit statistics so that differences between predicted and observed values would not be highly skewed and dominated by a small proportion of high values.

The correlation coefficient r measures the tendency of the predicted and observed values to vary together linearly. It can range from -1 to 1, with negative values indicating that the observed and predicted values tend to vary inversely. It should be recognized that even if the correlation is close to 1, the predicted and observed values may not match each other, they only tend to vary similarly.

The root mean squared error, average error, and average absolute error are all measures of the size of the discrepancies between predicted and observed values. Values near zero indicate a close match. The average error is a measure of aggregate model bias, though values near zero can be misleading because negative and positive discrepancies can cancel each other. The average absolute error and the root mean squared error both accommodate the shortcoming of the average error by considering the magnitude rather than the direction of each discrepancy. Together these three statistics provide an indication of model prediction accuracy.

The reliability index (Leggett and Williams 1981) quantifies the average factor by which model predictions differ from observations. An RI of 2.0 indicates that a model predicts the observations within a multiplicative factor of two, on average. Ideally, the RI should be close to one. When the root mean squared error has been calculated for log transformed values of the predictions and observations, then the RI is the exponentiated RMSE.

The modeling efficiency measures how well a model predicts relative to the average of the observations. It is related to the RMSE according to $\mathrm{MEF}=1-\mathrm{RMSE}^2/s^2$ where s^2 is the variance of the observations. A value near one indicates a close match between observations and model predictions. A value of zero indicates that the model predicts individual observations no better than the average of the observations. Values less than zero indicate that the observation average would be a better predictor than the model results. We calculated this statistic two ways. First we calculated the modeling efficiency (MEF) using the average of all observations as $\bar{O}$, then we calculated MEF by section, using the average of the observations in each estuary section as $\bar{O}$ (note that when all sections are considered together the MEF and the MEF by section are the same).

Because inclusion or removal of a waterbody from the 303(d) list is contingent on achieving an acceptable number of chlorophyll standard violations, it is important to consider a model's ability to predict this quantity. Therefore, we compared the number of exceedances of the chlorophyll *a* criterion (40 μg/L) occurring in the observations and in the corresponding predictions of each model for the year 2000. Predictions of the exceedance frequency resulting from the Neu-BERN model were made according to the methods described by Borsuk et al. (2002a).

Model Verification Results

Results from the model verification exercise are humbling (Table 1). When all five estuary sections are considered together, r values for all three models are greater than 0.7, suggesting a general agreement in model prediction and observation variability. However, r values within individual sections were generally much lower and declined moving down estuary, with NEEM and WASP exhibiting slightly negative values in the lower section. None of the models tracked observed behavior very well in the upper, middle, and bend sections (Figs. 2), the regions of highest historical chlorophyll *a* concentration. The relatively high aggregate r values, as compared to the relatively low values within sections, indicate that most of the variability captured by the models is the systematic longitudinal variation that occurs moving from the river to lower sections.

A similar intuition is gained from the MEF and MEF by section calculations (Table 1). The MEF by section indicates minimal predictive gains relative to the observation means within each section. In fact, the MEF by section is negative for all three models in the middle, bend, and lower sections. When the MEF is calculated using the grand mean of all observations, all values are positive, except in the lower section. Together, these confirm that the principal chlorophyll *a* variability captured in these models is section-to-section variation, rather than the spatial or temporal variation within each section.

Comparisons of the RMSE, AE, and AAE indicate that WASP tended to do best among the three models in the river and upper sections, nearer upstream boundary conditions. This performance advantage diminished through the middle and bend sections (Table 1), and was reversed in the lower section. With all sections combined, none of the models offers clear or consistent predictive superiority, based on these three statistics.

Values of the reliability index indicative that predictions differed from observations by a factor of approximately 2 to 3 (Table 1). Among all estuary sections, Neu-BERN had a slightly lower value than NEEM or WASP, but not within every section.

The difference in calibration objectives among the models may explain some of the differences in verification results. NEEM was calibrated to match the full range of observed chlorophyll values, especially the upper tail of the distribution at which violations occur. This led to a consistent overprediction of the remainder of the values in both the calibration (Bowen 2003) and verification exercises (Table 1 and Fig. 2). WASP, on the other hand, em-

Table 1. Model Summary Statistics for Year 2000 Chlorophyll *a* Verification

Model[a]	n[b]	Observed exceedances	Predicted exceedances	r[c]	Root mean squared error	Reliability index	Average error	Average absolute error	Modeling efficiency	Modeling efficiency by section[d]
River Section	94	0								
NEEM			0	0.60	0.99	2.70	−0.06	0.80	0.68	0.06
WASP			0	0.87	0.54	1.71	−0.17	0.32	0.91	0.72
Neu-BERN			0	0.26	1.00	2.72	−0.13	0.84	0.67	0.04
Upper Section	72	9								
NEEM			10	0.65	0.96	2.61	−0.17	0.79	0.24	0.24
WASP			5	0.71	0.83	2.30	−0.16	0.61	0.43	0.42
Neu-BERN			7	0.65	0.91	2.47	−0.35	0.70	0.32	0.32
Middle Section	60	10								
NEEM			9	0.40	0.92	2.51	0.004	0.77	0.10	−0.63
WASP			1	0.29	0.76	2.14	−0.17	0.59	0.39	−0.11
Neu-BERN			10	0.36	0.77	2.16	−0.33	0.59	0.37	−0.14
Bend Section	60	12								
NEEM			16	0.07	0.99	2.70	0.13	0.78	0.03	−1.33
WASP			0	0.11	0.86	2.36	−0.28	0.64	0.28	−0.75
Neu-BERN			12	0.22	0.72	2.06	−0.29	0.59	0.49	−0.23
Lower Section	69	8								
NEEM			13	−0.11	0.89	2.43	0.20	0.73	−0.03	−1.06
WASP			0	−0.05	1.31	3.71	−0.24	0.77	−1.24	−3.48
Neu-BERN			10	0.16	0.72	2.06	−0.35	0.55	0.32	−0.36
All Sections	355	39								
NEEM			48	0.74	0.91	2.60	0.01	0.78	0.41	0.41
WASP			6	0.76	0.88	2.41	−0.20	0.57	0.50	0.50
Neu-BERN			39	0.76	0.85	2.34	−0.28	0.67	0.53	0.53

[a]NEEM=Neuse estuary eutrophication model; WASP=water analysis simulation program; Neu-BERN=Neuse Estuary Bayesian ecological response network.

[b]Number of observations, exceedances refers to exceedances of the 40 μg/L chlorophyll criterion.

[c]Correlation coefficient of predicted and observed values.

[d]Modeling efficiency (using the overall mean of chlorophyll observations); MEF by section=modeling efficiency (using the mean of the chlorophyll observations for each estuary section).

ployed calibration procedures that amounted to temporally fitting predictions "through the middle of the data." Thus, predictions tended to underpredict the occurrence of extreme values, including standard violations (Table 1 and Fig. 2). Neu-BERN had similar, though more statistically based, objectives as WASP, and thus would also underpredict extremes if only the median predictions were considered. However, the model uncertainty term (see #1 below) accounts for this, leading to more accurate estimates of exceedance frequencies (Table 1).

Model Application to Total Maximum Daily Load Development

A detailed presentation of how the NCDWQ used the three models to guide the final TMDL decision can be found in the final TMDL report (NCDENR 2001). The following discussion is a summary of this process.

Phase II TMDL development required that 1991–1995 be used as a baseline period for determining future nitrogen load reductions. However, both NEEM and WASP required boundary condition data that were not collected until mid-1997. Thus, neither model could directly simulate the 1991–1995 baseline period. Therefore, both models were used to estimate the nitrogen load reduction that would be necessary to meet the chlorophyll criterion based on simulations of the period 1998–2000. NEEM results indicated a 5% nitrogen load reduction relative to this time period would be adequate to meet the chlorophyll criterion, while WASP indicated that no reduction was necessary.

To relate the 1998–2000 period that NEEM and WASP could simulate to the baseline period of 1991–1995, the NCDWQ compared long-term nitrogen load estimates from these two periods (Stow and Borsuk 2003; Stow et al. 2001). These analyses indicated that the average annual mass loads during the two periods were almost identical. However, while the mass loads were nearly the same, the 1998–2000 period included substantial river flow inputs from several hurricanes. A statistical correction to remove the year-to-year variation in river flow indicated that nitrogen concentrations in the Neuse River had decreased between the periods 1991–1995 and 1998–2000. This concentration decrease is likely attributable to decreases in nitrogen inputs to the river from point sources, as documented in discharge monitoring reports, and to the implementation of best management practices by many nonpoint sources, pursuant to the Neuse Rules. In essence, the average annual nitrogen mass load from 1998–2000 was comparable to the average annual load from 1991–1995 because increased river flow from 1998–2000 compensated for decreased nitrogen concentration. Thus, the NCDWQ judged that when selecting a TMDL relative to the 1991–1995 baseline period, the reductions of 0 and 5% from 1998–2000 indicated by NEEM and

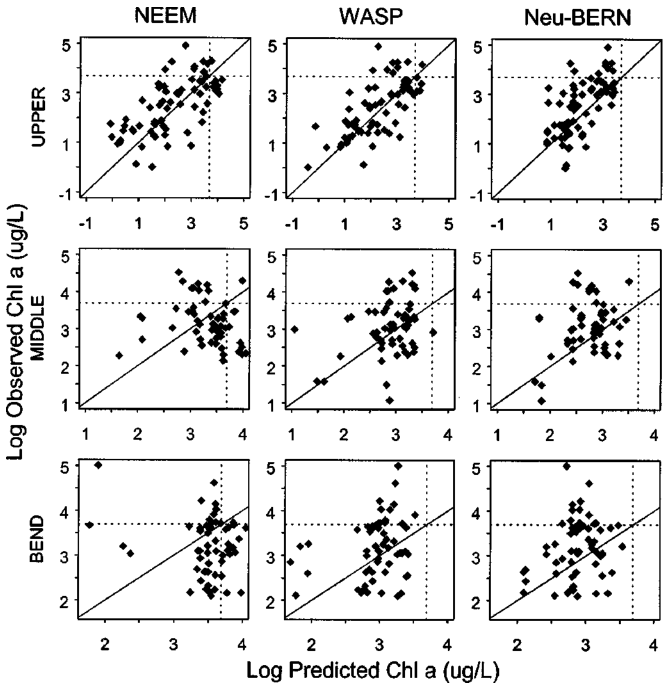

Fig. 2. Observed versus predicted chlorophyll *a* values from year 2000 verification study for all three models. Upper, middle, and bend sections are where most severe eutrophication problems occur. Solid lines depict a 1:1 correspondence, dashed lines depict 40 μg/L chlorophyll *a* criterion. Note that predictions for Neu-BERN model are median values of predicted distribution, thus predicted exceedances from Neu-BERN do not appear in these plots.

WASP, respectively, should be corrected for the nitrogen input reductions that had occurred in the intervening years.

Neu-BERN could simulate the 1991–1995 baseline period directly, but the suggested nitrogen reductions differed depending on whether the model was fit with or without a term to accommodate the interinstitutional differences in chlorophyll *a* measurement. With this accomodation, Neu-BERN indicated that a 45% nitrogen load reduction relative to 1991–1995 would be required (Borsuk et al. 2002b). Without it, Neu-BERN indicated reductions of approximately 15%. Because Neu-BERN predictions provide quantitative information on uncertainty, the target reduction was also dependent on the size of the margin of safety chosen (the confidence of compliance). This choice is a policy decision, dependent on the risk tolerance of the public and decision makers (Borsuk et al. 2002b). Thus, a range of target nitrogen reductions could be derived from the Neu-BERN model, dependent on some explicit decisions regarding model structure and the desired margin of safety.

Given the range of predictions indicated by the three models and analyses by Stow et al. (2001) and Stow and Borsuk (2003) which indicated that nitrogen reductions were already underway, the final TMDL prescribed a continuation of the 30% nitrogen load reduction established in the Neuse Rules (NCDENR 2001).

Differences among Models

Differences in TMDL implications of the three models may result from any of a number of differences in model structure and assumptions. Whether these model differences are considered advantages or disadvantages is somewhat context and perspective dependent. Some of the more important model differences include:

1. Neu-BERN considers unexplained variability due to model error, natural variability, and knowledge uncertainty in estimating exceedance probabilities (see Borsuk et al. 2002b for details) while NEEM and WASP do not. Due to practical difficulties in estimating error terms, neither NEEM nor WASP permits error propagation. It is unknown how the preliminary estimated reduction from NEEM and WASP would change if this type of analysis were included.
2. NEEM has a sediment diagenesis model that simulates the long-term impacts of sediment changes, whereas Neu-BERN and WASP do not have that capability. Neu-BERN's predictions characterize the immediate effects of riverine load reductions during the model periods. WASP uses an additional constant flux of nutrients from the sediment. Based on the NEEM results, the required nitrogen reduction difference between scenarios including and omitting sediment diagenesis is small and the uncertainty in these predictions is very high. Therefore, we cannot presently estimate how predictions from Neu-BERN or WASP would change if sediment diagenesis were considered.
3. Only Neu-BERN is able to directly address the 1991–1995 period, which NCDWQ considers to be the reference time period. Comparing the Neu-BERN reductions from 1991–1995 with the Neu-BERN reductions from 1998–2000 suggests that direct consideration of 1991–1995 conditions leads to greater required reductions, perhaps as a result of reductions that had already been achieved in the intervening years.
4. NEEM and WASP separately account for inorganic and organic nitrogen, Neu-BERN considers only total nitrogen. Inorganic nitrogen is generally considered more available for uptake by algae, so if management actions preferentially eliminate inorganic nitrogen, then Neu-BERN will likely estimate higher post N-reduction chlorophyll *a* levels than WASP and NEEM.
5. Neu-BERN included an explicit term to accommodate differences in chlorophyll *a* results obtained from UNC-IMS and Weyerhaeuser. These differences were not explicitly addressed in NEEM or WASP, but were partially accommodated by modeling chlorophyll concentration differentially with depth.
6. The calibration of WASP included side-channel data collected and analyzed by NCSU. These chlorophyll *a* data appear to be somewhat higher than the ModMon midchannel data, though this is based on a limited (1.5 years) comparison. We might expect that chlorophyll values in the side channel would be higher due to shallower depth and reduced mixing. However, because there are no overlapping sites where all institutions collected samples, the exact cause and magnitude of the measured differences is unclear.
7. WASP models three spatial dimensions, as opposed to two for NEEM and one for Neu-BERN. The third dimension (across the estuary) does not appear to be important in terms of chlorophyll *a* exceedances, as the results from WASP do not vary significantly in that direction. Side-channel locations may have an ecological importance that is not reflected by the current regulatory structure, which does not explicitly address lateral differences. If this should become an important consideration for future decision making, then NEEM and Neu-BERN would require additional capabilities.
8. In model applications, NEEM and WASP require upstream

and downstream boundary conditions, including chlorophyll *a* values, as inputs. Neu-BERN requires only the upstream forcing variables of flow, total nitrogen concentration, and water temperature.

9. As an explicitly data-based model, Neu-BERN is subject to the usual criticisms directed at empirical models, including those related to application of the model to previously unobserved conditions. NEEM and WASP, on the other hand, consist of process formulations intended to have information content beyond that contained in the observational data. This should increase their potential to correctly predict future conditions if the a priori information content is correct. However, neither of the process-based models was fully specified a priori. Both models required significant calibration to data.

When comparing the various advantages/disadvantages and apparent implications of differences in structure and assumptions of these models it is important to recognize that each model is an extremely crude representation of complex system behavior, based on an incomplete understanding of the system. There are likely to be unexpected feedbacks and changes in system behavior with time that are not captured in any of these models. Ongoing theoretical work suggests the possibility that aquatic ecosystems experience nonlinear shifts between alternate stable states such as oligotrophy and eutrophy, with asymmetric pathways between these states (Scheffer 2002). The practical implication of this work for systems such as the Neuse is that the nitrogen reduction needed to return the estuary to a desirable state may not be the same as the nitrogen increase that put the system in the current undesirable state. It is not clear that the current generation of models being used for TMDL development, including the models used in this study, describe such nonlinear, asymmetric shifts. Thus, any discussion of the relative attributes of these models should be tempered with the realization, reinforced by the verification results (Table 1), that these models are primarily tools for decision making, not archetypes of estuary behavior.

Discussion and Conclusions

Mathematical models have long been advocated as a rigorous means for evaluating the effectiveness of alternative strategies for managing water quality, but our verification results underscore the realization that predicting ecosystem behavior is inexact. Based on the results we have presented, it may appear that the three water quality models did little to contribute to the TMDL decision. Indeed, the model verification exercise showed that relatively little confidence could be attached to the specific predictions of any of the models, and the TMDL decision was not changed from the initial 30% reduction strategy contained in the Neuse Rules as a result of the modeling effort. However, despite their predictive shortcomings, these models were essential in the TMDL decision process. In fact, there are few alternatives to mathematical models for reconciling the likely outcomes of alternative management options. Direct experimentation at an appropriate scale is usually implausible, and the uncertainty associated with extrapolation from smaller scale experiments is difficult to quantify. Similarly, aquatic ecosystems are typically so idiosyncratic in their individual behavior that extending inference among systems can be misleading. However, the utility of models is to provide quantitative guidance rather than a definitive number.

In addition to providing guidance for the TMDL, the modeling effort helped facilitate the exchange of ideas among the stakeholders. The models facilitated an articulation of beliefs and disagreements among stakeholders regarding processes and feedbacks important in the functioning of the river and estuary, as well as an appreciation of the attributes of the system that were considered important by different members of the community. The models also provided a better appreciation of the scientific uncertainties inherent in predicting the response of the estuary to various management options.

The ability to forecast ecosystem behavior is extremely important for management decision making (Clark et al. 2001, National Research Council 2001). But our results indicate that, even in a well-studied, data-rich system, accurate prediction is difficult. Ecosystems exhibit behavior characteristic of self-organized, complex adaptive systems, which may be difficult to predict, as well as behavior driven by more foreseeable environmental processes (Levin 1999). Differentiating the predictable from the unpredictable behavior remains a modeling frontier.

The concept of adaptive management (Walters 1986) arose from the realization that predicting ecosystem behavior is inherently uncertain. Rather than use uncertainty as an excuse for inaction, pending further study, adaptive management is a call to action. Adaptive management advocates the concept of ''learning by doing''—using management actions as large-scale experiments to enhance what is learned through further study. Management is considered to be a process through time rather than a single set of actions.

In a recent report (National Research Council 2001), the National Research Council Committee To Assess the Scientific Basis of the Total Maximum Daily Load Approach to Water Pollution Reduction recommended adaptive implementation (adaptive management) as a key element for improvement of the TMDL program. Consistent with our experiences, the NRC TMDL committee observed that water quality models used for TMDL development typically yield predictions that are highly uncertain; as a result, TMDLs may be expected to either fail to achieve designated uses or to exceed the necessary pollutant load reduction. To address this shortcoming, the committee urged that TMDLs be implemented in an adaptive manner, with postimplementation monitoring providing feedback to improve the TMDL over time. We agree with the committee's recommendation.

Adaptive management may be unpalatable to some, who feel that management decisions should provide resolution. But Holling et al. (2002) argue that stochastic factors, multistable behaviors, and the presence of environmental variables operating at different temporal scales will cause adaptive management to outperform strategies that seek optimal stable targets. With our relative inability to make precise, accurate forecasts, as this analysis demonstrates, it is difficult to see any practical alternative.

Acknowledgments

The writers thank the Neuse River Modeling and Monitoring (ModMon) project participants and the U.S. Geological Survey for providing the data used to develop and test the models. Support for ModMon was provided by the Water Resources Research Institute of the University of North Carolina, the State of North Carolina General Assembly, and the North Carolina Department of Environment and Natural Resources. Additionally, MEB was supported by an EPA STAR graduate fellowship.

References

Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2002a). ''The confounding effects of nitrogen load on eutrophication of the Neuse River

Estuary, North Carolina.'' in press.
Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2002b). ''Predicting the frequency of water quality standard violations: A probabilistic approach for TMDL development.'' *Environ. Sci. Technol.*, 36(10), 2109–2115.
Borsuk, M. E., Stow, C. A., and Reckhow, K. H. (2003). ''Integrated approach to TMDL development for the Neuse River Estuary using Bayesian probability network model (Neu-BERN).'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, in press.
Bowen, J. D. (2000). ''Calibration performance of a two-dimensional, laterally-averaged eutrophication model of a partially mixed estuary.'' *Estuarine and Coastal Modeling: Proc., 6th Int. Conf.*, M. L. Spaulding and A. F. Blumberg, eds., ASCE, New York, 1244–1258.
Bowen, J. D. (2003). ''A CE-Qual-W2 model of the Neuse Estuary for TMDL development.'' *J. Water Resour. Plan. Manage. Div., Am. Soc. Civ. Eng.*, in press.
Christian, R. R., Bryant, W. L., and Stanley, D. W. (1986). ''The relationship between river flow and Microcystis aerunginosa blooms in the Neuse River, North Carolina.'' *Rep. No. 223*, Water Resources Research Institute of the University of North Carolina, Raleigh, N.C.
Clark, J. S., et al. (2001). ''Ecological forecasts: An emerging imperative.'' *Science*, 293, 657–660.
Cole, R. W., and Buchak, E. M. (1995). ''CE-QUAL-W2: A two-dimensional, laterally averaged, hydrodynamic and water quality model. Version 2.0.'' *Instructional Rep. EL-95-1*, U.S. Army Engineer Waterways Experiment Station, Vicksburg, Miss.
Hobbie, J. E., and Smith, N. W. (1975). ''Nutrients in the Neuse River Estuary.'' *Publication Number UNC-SG-75-21*, Univ. of North Carolina Sea Grant Program, Raleigh, N.C.
Holling, C. S., Carpenter, S. R., Brock, W. A., and Gunderson, L. H. (2002). ''Discoveries for sustainable futures.'' *Panarchy, understanding transformations in human and natural systems*, L. H. Gunderson and C. S. Holling, eds., Island Press, Washington, D.C.
Leggett, R. W., and Williams, L. R. (1981). ''A reliability index for models.'' *Ecol. Modell.*, 13, 303–312.
Levin, S. A. (1999). *Fragile dominion, complexity and the commons*, Perseus, Cambridge, Mass.
National Research Council. (2001). *Assessing the TMDL approach to water quality management*, National Academy Press, Washington, D.C.
North Carolina Department of Environment and Natural Resources (NCDENR). (1999). ''Total maximum daily load for total nitrogen to the Neuse River Estuary, North Carolina.'' North Carolina Dept. of Environment and Natural Resources, Raleigh, N.C.
NCDENR. (2001). ''Preliminary reduction target assessment for the Neuse River Estuary.'' North Carolina Dept. of the Environment and Natural Resources, Raleigh, N.C.
Office of Water. (1997a). ''Compendium of tools for watershed assessment and TMDL development.'' *EPA-841-B-97-006*, U.S. Environmental Protection Agency, Washington, D.C.
Office of Water. (1997b). ''Guidelines for preparation of the comprehensive state water quality assessments.'' *EPA-841-B-97-002*, U.S. Environmental Protection Agency, Washington, D.C.
Office of Water. (1999). ''Protocol for developing nutrient TMDLs.'' *EPA-841-B-99-007*, U.S. Environmental Protection Agency, Washington, D.C.
Paerl, H. W. (1983). ''Factors regulating nuisance blue-green bloom potential in the lower Neuse River, NC.'' *Rep. No. 188*, UNC Water Resources Research Institute, North Carolina State Univ., Raleigh, N.C.
Paerl, H. W. (1987). ''Dynamics of blue-green algal (Microcystis aeruginosa) blooms in the lower Neuse River, North Carolina: Causative factors and potential controls.'' *Rep. No. 229*, Water Resources Research Institute of the University of North Carolina, Raleigh, N.C.
Paerl, H. W., and Bowles, N. D. (1987). ''Dilution bioassays: Their application to assessments of nutrient limitation in hypereutrophic waters.'' *Hydrobiologia*, 146, 265–273.
Paerl, H. W., Mallin, M. A., Donahue, C. A., Go, M., and Peierls, B. L. (1995). ''Nitrogen loading sources and eutrophication of the Neuse River Estuary, North Carolina: Direct and indirect roles of atmospheric deposition.'' *Rep. No. 291*, Water Resources Research Institute of the University of North Carolina, Raleigh, N.C.
Reckhow, K. H. (1999). ''Water quality prediction and probability network models.'' *Can. J. Fish. Aquat. Sci.*, 56(7), 1150–1158.
Reckhow, K. H., and Chapra, S. C. (1983). ''Confirmation of water quality models.'' *Ecol. Modell.*, 20, 113–133.
Reckhow, K. H., Clements, J. T., and Dodd, R. C. (1990). ''Statistical evaluation of mechanistic water-quality models.'' *J. Environ. Eng.*, 116(2), 250–268.
Rudek, J., Paerl, H. W., Mallin, M. A., and Bates, P. W. (1991). ''Seasonal and hydrological control of phytoplankton nutrient limitation in the lower Neuse River Estuary, North Carolina.'' *Mar. Ecol.: Prog. Ser.*, 75, 133–142.
Scheffer, M., Westley, F., Brock, W. A., and Holmgren, M. (2002). ''Dynamic interaction of societies and ecosystems—linking theories from ecology, economy, and sociology.'' *Panarchy, understanding transformations in human and natural systems*, L. H. Gunderson and C. S. Holling, eds., Island Press, Washington, D.C.
Stow, C. A., and Borsuk, M. E. (2003). ''Assessing TMDL effectiveness using flow-adjusted concentrations: A case study of the Neuse River, NC.'' *Environ. Sci. Technol.*, in press.
Stow, C. A., Borsuk, M. E., and Stanley, D. W. (2001). ''Long-term changes in watershed nutrient inputs and riverine exports in the Neuse River, North Carolina.'' *Water Res.*, 35(6), 1489–1499.
United States Environmental Protection Agency (USEPA). (2001). ''U.S. Environmental Protection Agency, Region IV, Neuse River/Estuary Hydrodynamic and water quality monitoring report.'' Atlanta.
Walters, C. J. (1986). *Adaptive management of renewable resources*, Macmillan, New York.
Wool, T., and Davie, S. (2003). ''An EFDC WASP model of the Neuse Estuary for TMDL development.'' *J. Water Resourc. Plann. Manage. Div., Am. Soc. Civ. Eng.*, in press.