On the effect of scaling conceptual model complexity on stochastic response for water quality modeling

G. T. Parker

ABSTRACT

This paper extends previous work comparing the response of water quality models under uncertainty. A new model, the River Water Quality Model no. 1 (RWQM1), is compared to the previous work of two commonly used water quality models. Additionally, the effect of conceptual model scaling within a single modelling framework, as allowed by RWQM1, is explored under uncertainty. Model predictions are examined using against real-world data for the Potomac River with a Generalized Likelihood Uncertainty Estimation used to assess model response surfaces to uncertainty. Generally, it was found that there are tangible model characteristics that are closely tied to model complexity and thresholds for these characteristics were discussed. The novel work has yielded an illustrative example but also a conceptually scaleable water quality modelling tool, alongside defined metrics to assess when scaling is required under uncertainty. The resulting framework holds substantial, unique, promise for a new generation of modelling tools that are capable of addressing classically intractable problems.

Key words | conceptual scaling, river, model, water quality, watershed, uncertainty

INTRODUCTION

Background

Watershed-level management of nutrient and other pollutants presents challenges both in terms of source and runoff assessments but also the impacts of these loads on water quality targets and kinetics throughout dendritic river systems. Riverine water quality models have therefore become important tools in assessing impacts of point and non-point sources and managing water issues. Existing tools used to address river water quality issues range from very basic formulations of the original Streeter-Phelps (SP) model (1925) to vastly more complex systems involving dozens of state variables, processes and kinetic parameters (e.g. Reichert et al. 2001).

In Parker et al. (2009) a comparison of the behaviour of two conceptual models of water quality kinetics under uncertainty was presented. The models presented in that work were QUAL2E (Brown & Barnwell 1987) and MIKE11 (DHI 1992), each of which have a long and broad history of application to these sorts of problems. Though quite similar in terms of complexity and constituents tracked, the two models demonstrated appreciably different characteristics under uncertainty, as demonstrated in their respective model response surfaces. These differences were shown to have important implications on the use and selection of a conceptual model for a given context.

Problem outline

The shifts in model response under uncertainty have implications for decisions pertaining to the conceptual model itself as well. By considering conceptual model complexity as a ‘scaleable’ element, analogous to spatio-temporal discretization, the question arises as to the trade-offs associated with model robustness versus complexity under stochastic conditions. For many problems, the choice of conceptual scale is unclear, with several equally suitable (or equally unsuitable, given issues of data availability and parameter identifiably) alternatives available (e.g. Wichern et al. 2006). In this paper, the model surface response under uncertainty is examined for a water quality model with a scaleable conceptual model complexity. The water quality model selected for this task was
the River Water Quality Model no. 1 (RWQM1) formulation (Reichert et al. 2001) which can be scaled from a simple SP type model to an advanced conceptual model with several dozen processes and state variables tracked (see Vanrolleghem et al. 2001). The advantage of the RWQM1 model is that its formulation specifically allows for a greater consistency in major conceptual elements (i.e. processes, state variables, kinetic formulations) across a spectrum of complexities when compared with two entirely distinct models.

METHODOLOGY

Approach

A FORTRAN implementation of RWQM1, as defined in Reichert et al. (2001) and Vanrolleghem et al. (2001), was incorporated into the qual-IT modelling toolkit. The qual-IT toolkit was developed as part of previous research (e.g. Parker et al. 2009), and allows, in particular, the measurement of model surface response characteristics under uncertainty. This is assessed by conducting a generalized likelihood uncertainty estimation (GLUE) behavioural analysis of the combined set of conceptual model, model parameters and calibration data used. These model characteristics are examined and discussed for the model framework here.

Study area and data

The Potomac watershed in the Eastern United States covers an area of approximately 38,000 km². The Potomac River itself stretches over 600 km from Fairfax Stone in West Virginia to Point Lookout in Maryland. The average river flow at the Washington, D.C location is about 310 m³/s, with minimum and maximum recorded flows of 17 m³/s (September 1966) and 12,000 m³/s (March 1956), respectively. The watershed flows into Chesapeake Bay, becoming tidal where the river crosses the Fall Line into the Coastal Plain province. Downstream from here, the water becomes increasingly brackish with a gradual increase in salinity. This study limits itself to the freshwater and non-tidal portion of the Potomac. Figure 1 shows a simplified map of the region, and in particular the dendritic river system that flows to the study region’s outlet in the East.

From 1991 to 1996, the U.S. Geological Survey’s (USGS) National Water Quality Assessment (NAWQA) program began extensive work in the basin to monitor surface and ground water quality and ecology. That dataset is used here, and also draws on over 50 USGS stations to provide information on stream flow quantity at up to daily timesteps for individual stations. Water quality information was gathered at 5 ‘integrator’ and 5 ‘indicator’ stations on a monthly basis. Point Sources loading the river system were identified from U.S. EPA (2008).

Water quality simulation

Typical constituents of interest in water quality models include dissolved oxygen (DO), nutrients (e.g. nitrogen, phosphorus), and toxics (e.g. pesticides). An extensive review of water quality modelling approaches in beyond the scope of this outline, (but can be found in e.g. Rauch et al. 1998 or Bowie et al. 1985). Previous work has compared the behaviour of two different conceptual models (QUAL2E and MIKE11). The resulting analysis showed that important differences in model behaviour are associated with both the level of complexity and the specific conceptual formulation of key processes (see Parker et al. 2009). In this paper a similar methodology is used to compare a kinetic model (RWQM1) that has somewhat more consistent conceptual formulation whose complexity can be scaled (through addition or removal of processes, for instance) as needed for a specific problem. The RWQM1 formulations examined herein are based on 4 levels (‘Simple’ and ‘Extended’ versions of SP, ‘Similar to Qual2e’ and ‘Simplified [Full] RWQM1’) presented in Vanrolleghem et al. (2001). The specific implementation was coded in FORTRAN on the qualit platform (previously presented in Parker et al. 2009) and is dubbed qualit-RWQM.

Table 1 shows a comparison of the complexity of each examined model, with a corresponding designation and

![Figure 1](https://iwaponline.com/wst/article-pdf/63/2/360/445260/360.pdf)
several measures of complexity levels. The second column presents rank of complexity, and the following 3 columns describe the dimensionality of a parameter/state-variable matrix describing each model. These are qualitative, relative measures of model complexity that serve to compare one model to another. \( I_c \) and \( I'_c \) are, respectively, the complexity measures presented in Snowling & Kramer (2001) and Parker (2010), which attempt to quantify how much more complex one conceptual model is than another.

Output of the models was compared for stochastic simulation of DO, BOD and ammonia-N (NH4). Initial ranges for kinetic parameters are drawn from Vanrolleghem et al. (2001), with other model parameters estimated from literature values (Bowie et al. 1985).

### Determination of model surface response

The GLUE methodology used was as described previously in Parker et al. (2009) and others. GLUE approaches as first defined by Beven & Binley (1992) seek to identify the set of behavioural models supported by a given set of data. Since many model structures and parameter sets are typically of comparable predictive ability, the set of these models can be considered together as the surface response of the specified model components. The GLUE concept has commonly been used in hydrology and, more recently, also been adapted to riverine water quality problems (Parker et al. 2009; Zheng & Keller 2007).

In this technique, the focus is on estimating posterior distributions for predicted variables directly from the outputs of a set of representative models, rather than from a model and an additive residual model. The resulting set of models is further constrained by the calibration data and criteria (cf. Beven 2001 and Rode et al. 2007), and contains information on the simulation output variability for all representative input sets. The representative set can be viewed as the range of water quality models that conform to the calibration criteria, which in turn are constrained by the specific dataset used.

The specified model components selected for this project were the qualit-RWQM implementation at 4 levels of complexity and the Potomac River NAWQA data. The other major component which forms the model is the calibration criteria selected, and the likelihood function used.

#### Absolute Relative Error (ARE)

Behavioural (or ‘representative’) models were defined as those matching validation values at the system outlet and water quality stations for constituents including DO, rapidly biodegradable organic matter (BOD) and ammonium (NH4) with an absolute relative error (ARE) beneath 30%. A simple binary (pass/fail based on ARE) likelihood function was used to calculate surface model dispersion statistics.

The ARE calibrations are based on the absolute error relative to data at each calibration point, when compared with annual and seasonal averages, using the simple relation found in Equation 1.

\[
ARE = \frac{(C_{\text{model}} - C_{\text{data}})}{C_{\text{data}}} \tag{1}
\]

### Simulation matrix

A base set of runs was conducted for each conceptual model, consisting of 2 million randomized parameter sets from the selected sample space. These initial runs were divided into parallel jobs across several processors, and then reassembled for further processing. These tasks are performed by the qualit-RWQM toolset. The resulting run sets share similar characteristics and sample size, suggesting that our selection criteria and initial run size were sufficient to avoid some potential sampling issues.
Surface response metrics

The basic statistical measure used to describe the response surface in this paper is the coefficient of variation (CV, sometimes also termed ‘relative variability’), which expresses the dispersion of a data set around its mean (cf. Equation 2).

\[ c_v = \frac{\sigma}{\mu} \] (2)

The coefficient of variation is useful because it allows the degree of variation in several series of data to be compared, and is therefore often used to assess relative volatility or risk.

An additional metric that is often used in discussing model performance is the root mean squared error (RMSE). The use of RMSE as a simulation performance metric should not be confused with its use as a calibration criteria, which has previously been argued against in the case of stochastic modeling (e.g. Parker et al. 2010). Nevertheless, the metric is a measure of the predictive value of the performance of the set of representative parameter members when compared to observed data. The RMSE is calculated according to the following equation:

\[ RMSE = \sqrt{\frac{\sum_{i=1}^{n} (c_{\text{model}_i} - c_{\text{data}_i})^2}{n}} \] (3)

where \( c_{\text{data}_i} \) and \( c_{\text{model}_i} \) are the \( i \)th observed and simulated constituent concentration, respectively.

RESULTS AND DISCUSSION

Results

Model surface response showed considerable differences, both between average ‘deterministic’-type instances of each conceptual model, but, more significantly, between their response surfaces. An illustrated example of one analysis is shown here in Figure 2 and discussed below.

Figure 2 shows the change in surface response dispersal (measured using normalized CVs) at various flow rates throughout the watershed. This plot reveals some interesting insight about how the models perform and contrast under uncertainty. The more simple model tend to perform worse (higher CVs) than the more complex model at very small flow rates (i.e. first order streams and the like), however the situation is decidedly less clear as the flow rates increase, and differences in model complexity become outweighed by model structure. Additionally, the range of flow scales (and thus drainage area scales) at which each conceptual model performs best, in terms of narrow response surfaces, varies considerably, even between simpler models. The very simple
model are optimal in this regard above the very small flows for just about one-third of the examined flow rates, while at higher flow the ‘Qual2e-like’ representation outperforms both more complex and simpler conceptual models. These results imply an important link between model scale and uncertainty, requiring careful balancing of scale versus complexity, and pose other very similar themes to previous work as discussed in Parker (2009) for hydrological models.

Figure 3 shows the change of RMSE in model vs. observed DO at 4 complexity levels. The model complexity index is $I_c$ as listed in Table 1.

Figure 3 again shows differences in model performance relative to location. RMSE has broadly increased as the examined location nears the outlet, which is likely based on data-limited issues and the increasing size of the upstream watershed area. More relevant to this paper is that the RMSE at the upstream location tends to increase slightly with model complexity, but at the downstream location this trend seems to be reversed. This suggests that the model complexity is becoming increasingly advantageous under circumstances of increasing watershed area and data branching.

**DISCUSSION**

It is well understood that conceptual model complexity is an important factor which determines the characteristics, and ultimately utility, of simulation results. Snowling & Kramer (2003), and others have argued that a ‘balance’ point exists that balances increasing sensitivity and decreasing model error as complexity increases for deterministic cases. They present an ‘X’ figure comparing on the axes model utility vs. complexity, with the middle of the ‘X’ representing the balance point of the trade-off between increasing model sensitivity and decreasing error. This paper examines this argument, and extends the analysis to stochastic conditions, which in any case are probably more relevant when discussing model sensitivity (as this generally involved pseudo-stochastic analysis in the form of e.g. Monte-Carlo simulations), as well.

The results show that a simple ‘balance point’ analysis based on conceptual model characteristics themselves, is insufficient to adequately characterize the complexity effects. Both the CV and RMSE results (as presented in Figures 2 and 3, respectively) illustrate that data mixing and information flow (which follows literal flow, both upstream and downstream, in the case of river systems) are also significant factors in determining which level of complexity should be preferred. This should not be surprising, as it intuitively makes sense that models with additional complexity will not only be able to describe more complex systems better (which may be of use at points downstream or under otherwise more testing circumstances) but will also tend to show additional model inertia and rigidity (as defined in Parker et al. 2009). The phenomenon can be further illustrated by the considering the schematic shown in Figure 4, which shows two different conceptual models under identical conditions, and the network of ‘flows’ reflecting kinetic dependencies between state variables.

This schematic was generated using the QualViz toolset presented in Parker (2010). A similar schematic could be constructed at every location throughout the examined watershed, but is not necessary to discuss the salient point. One effect of increasing model complexity at a given location (but also between locations) is to create a more dense network of interdependencies between state variables, even those that may not be of direct interest or reported.
With this point clarified, one can now reconsider Figure 2, which shows the change in model response at various levels of complexity vs. flow rates. The most significant change in the model behaviours occurs at intermediate flows, above and below which all models seem to perform more consistently. This can be explained with the insight that these intermediate scales are those at which the most significant mixing of lower order branches occur, but also where sufficient (calibration) data exists to add rigidity to the upstream models. The ‘dip’ is what happens as the models must abruptly reconcile new data inputs (from other branches) in the mixing locations. This is again, analogous to Parker et al. (2009).

What this implies is that, in any meaningful sense, the ‘best’ model depends to a significant degree on data conditions (upstream -and- downstream) as well as the conceptual structure of the model itself. In effect, the physical (meaning the layout of the dendritic river) and conceptual (meaning the complexity and assumptions of the model) network both dictate when and if a given model may be optimum at each position. In terms of Snowling & Kramer (2001), the ‘balance’ point is subject to shift in locations and data conditions throughout the watershed, as are the sensitivity and error curves themselves.

CONCLUSIONS

The suggested methodology can be used to test conceptual model validity and especially robustness under uncertainty for specified problems. As the science and scope of water quality modelling has progressed, the range of conceptual model complexities has also increased considerably. The results of this study will help users select and assess models (e.g. QUAL2E vs. MIKE11 vs. RWQM1) for varied problems, and refine appropriate data collection and monitoring schemes for a given problem. The use of a scaleable (or ‘nested’) model under uncertainty, as presented here, is novel and further allows for progressive increase or decreases in model complexity as required to meet specific trade-off criteria in terms of stochastic robustness versus conceptual model scope. Such a tool has immense potential relevance to a wide range of applications, but judicious use requires both a firm intuitive grasp and tailored metrics that reflect the behavioural characteristics of the scaled models when faced with calibration data and a specific physical description of the area examined. These conceptual elements together ‘flow’ through and shape the model in a sort of ‘information hydrology’ which directly impacts the aggregate model performance.

REFERENCES


U.S. EPA 2008 Chesapeake Bay Phase 5 Community Watershed Model In preparation EPA XXX-X-XX-008 Chesapeake Bay Program Office, Annapolis MD.


