Gu, Chuanhui, and Newsha K. Ajami (2010). Complexity In Microbial Metabolic Processes In Soil Nitrogen Modeling: A Case For Model Averaging. Stochastic Environmental Research and Risk Assessment, 24(6) 24.6: 831-44. Print. Version of record available from Springer Verlag. [ISSN: 1436-3240], [DOI: 10.1007/s00477-010-0381-4]

Complexity In Microbial Metabolic Processes In Soil Nitrogen Modeling: A Case For Model Averaging

Newsha K. Ajami · Chuanhui Gu

1 Introduction

With increasing concerns over environmental problems caused by human disturbances, accurate and reliable predictions of nutrient cycling are essential for sustainable resource management practices. In the last few decades, much effort has been invested in developing biogeochemical models with different complexity that capture nutrient cycling across the scales (Li et al. 1992; Parton et al. 2001; Maggi et al. 2008). Regardless of the detail at which the chemical, biological and physical processes are taken into account, all these models are, in the best case, selective mathematical approximations of processes that are essentially complex.

Choosing a level of complexity has been a routine concern by biogeochemical modelers (e.g. Johnson and Omland 2004; Homann et al. 2000) as Biogeochemists have usually struggled to decide what level of complexity
in models is actually warranted (e.g., Kimmins et al. 2008). However, this so-called best-model selection really does not solve the significant variability between model performances (Homann et al. 2000). Thus, the standard notion of choosing the right complexity level (e.g., Lawrie and Hearne 2007) doesn’t overcome the model’s structural inadequacy. Furthermore, when there is nearly equivalent support in the observed data for multiple models, it is problematic to choose one model over another. Breuer et al. (2008) recently reviewed some of the nitrogen hydro-biogeochemical models and discussed the importance of addressing various sources of uncertainty in such models, especially the model structural (i.e. complexity) uncertainty. Multi-model averaging can lead to reduction of model selection bias and account for model selection uncertainty.

The need for accounting for model structure uncertainty by multi-model averaging have motivated researchers in various fields such as economic, weather and hydrological forecasting to consider multi-model methods that aim to obtain consensus prediction from a set of models by different averaging schemes (Bates and Granger 1969; Dickinson 1973; Krishnamurti et al. 1999; Shamseldin et al. 1997; Stockdale 2000; Raftery et al. 2005). One way to aggregate multiple deterministic model outputs is to employ deterministic techniques (i.e. regression based approaches) (Shamseldin and O’Connor 1999; Piedelievre 2000; Georgakakos et al. 2004; Ajami et al. 2006). None-the-less, the deterministic averaging does not provide an indication of the relative quality of the different model structures and, thus, it is difficult to include it in a quan-titative uncertainty analysis in terms of probabilities (Refsgaard et al. 2006).

Bayesian Model Averaging (BMA; Hoeting et al. 1999), on the other hand, is a probabilistic multi-model averaging technique. It uses a statistical scheme to infer from an ensemble of competing deterministic predictions the probabilistic prediction that possesses more skill and reliability than the original ensemble members (Madigan et al. 1996; Raftery 2003). BMA assigns weight to each model which reflects the relative model performance given the observations. The BMA weights sum up to unity. BMA methods were used in various forecasting applications such as surface water hydrological forecasting (Duan et al. 2007; Ajami et al. 2007; Vrugt and Robinson 2007; W ‘o hling and Vrugt 2008; Hsu et al. 2009), ground water modeling (Neuman 2003; Neuman and Wierenga 2003; Ye et al. 2004; Poeter and Anderson 2005; Refsgaard et al. 2007; Ye et al. 2008; Rojas et al. 2009) and weather forecasting (Raftery et al. 2003, 2005) and have shown promising results in dealing with model predictive uncertainty.

Soil nitrogen cycling is complex in the sense that it involves many interactive physical, chemical, and biological processes (e.g., Gu et al. 2009). Models have been constructed to predict soil N transport and fate with varying complexity of processes, including water dynamics, carbon cycling, plant uptake, chemical transport, and N transformation pathways (Chen et al. 2008). The model uncertainty can inevitably arise from different conceptualizations and subsequent mathematic representations of those processes.

We here focused on soil nitrogen microbial metabolic processes as an appropriate example to investigate the general biogeochemical dilemma about complexity. It was outside the limits of this study to investigate all the possible model complexity mentioned previously. The process used here represents only a subset of N transformation processes, but arguably most dominant process that controls soil nitrogen fate (Maggi et al. 2008). In the current study, we described the nitrogen microbial meta-bolic processing (1) lumped microbial kinetics or (2) di-saggregating into functional groups (e.g., denitrifiers and ammonia oxidizers). Models with different levels of microbial complexity can give different model results, but none may be deemed entirely adequate (Homann et al. 2000). Therefore, it provides an ideal test of how BMA can lead to new insights about model uncertainties and system behavior.

The objective of this paper is two fold. First, we explore the importance of a model’s complexity level in developing an accurate bio-geochemical representation of state variables including soil gas emissions and pore water solute concentration. This has been a focus of many ecological and biological studies (Homann et al. 2000; Kimmins et al. 2008). Here we will examine the plausibility of the notion of selecting the “right” complexity level to address all different processes within a field (Lawrie and Hearne 2007). We later illustrate that changing complexity levels does not conceal the model’s structural deficiencies. Therefore, the second objective focuses on addressing the model predictive uncertainty in biogeochemical models by using Bayesian Model Averaging approach. We developed three different versions of biogeochemical N cycle model THOUGHREACT-N (Maggi et al. 2008; Gu et al. 2009; Xu 2008). Each version of the model presents different complexity levels in representing the N-cycling in our study field. The models were used to describe the N cycling in a tomato field in the Western Sacramento County, CA, site from which we measured different quantities over time. Each model was calibrated twice on these data sets using Parameter ESTimation (PEST) software (Doherty 2004) considering two different weighting sets in our objective function. Later we investigated the importance of accounting for predictive uncertainty using Bayesian Model Averaging to combine the existing six various deterministic model outputs.
2 Methods

2.1 The biogeochemical cycle of nitrogen

The reactions responsible for N cycling are numerous and mainly mediated by variegate microorganisms that extensively inhabit near-surface soils. These microorganisms can transform nitrogen via more than one pathway (e.g., Wrage et al. 2001; Shrestha et al. 2002), and under various conditions of temperature, pH, water content, substrate, electron acceptor, and inhibitor concentrations (e.g., Knowles 1982).

The reaction patterns accounted for in this work are biological nitrification (i.e., NH₄⁺ → NO₃⁻ → NO₂⁻ → NO → N₂O → N₂), and denitrification (i.e., NO₃⁻ → NO₂⁻ → NO → N₂O → N₂) reaction chains, and chemical N decomposition (i.e., NO₃⁻ → HNO₃ → HNO₄ → NO₃⁻ → HNO₃ → NO₃⁻ → HNO₃ → NO₃⁻ → NO₃⁻ → NO₃⁻ → NO₃⁻ → NO₃⁻ → NO). Nitrification of ammonium (NH₄⁺) into nitrate (NO₃⁻) consists of two oxidation reactions mediated respectively by ammonia oxidizing autotrophic bacteria (AOB; e.g., Nitrosomonas and Nitrosospira) and by nitrite oxidizing autotrophic bacteria (NOB; e.g., Nitrobacter and Nitrosospira) (e.g., Salsac et al. 1987; Arp and Stein 2003). AOB and NOB are active in moist soils under oxic conditions, while they become inactive at low soil moisture content and water potential, and under anoxic conditions (e.g., Rosswall 1982; Rodrigo et al. 1997). Denitrification of NO₃⁻ into N₂ consists of a sequence of redox reactions mainly mediated by heterotrophic denitrifier bacteria (DEN; e.g., Pseudomonas, Thiobacillus) that consume dissolved organic carbon (DOC) and use NO₃⁻, NO₂⁻, NO, and N₂O as electron acceptors (e.g., Payne 1973; Knowles 1982; Rosswall 1982). The metabolism of DEN, with concomitant CO₂ production, is limited by available DOC, and is favored under anoxic conditions while declining under water drought stress. Some AOB are capable of carrying out part of the denitrification reactions in anoxic conditions by means of the same reactions as DEN (e.g., Tortoso and Hutchinson 1990; Wrage et al. 2001). Chemical decomposition of HNO₃ into HNO₂ and NO was taken into account as demonstrated in Venterea and Rolston (2000a).

In linking the N cycle to part of the C cycle in a simplified manner, we assume that background organic carbon dissolution sustains a DOC pool as in Li et al. (1992). Although DOC is known to comprise a large number of complex and more or less recalcitrant mole-cules, we consider CH₃O as the available DOC substrate for simplicity (e.g., Chen and MacQuarrie 2004). DOC is competitively consumed by AOB and DEN during denitrification, and by other heterotrophic and aerobic microorganisms (AER) during respiration, resulting in CO₂ production.

2.2 Experimental data

Experimental data were collected from a furrow irrigated tomato field during July–August 1998 in western Sacramento County, California. Fertilization consisted of an application of anhydrous ammonia, NH₃, injected at 5 cm depth on 1st July. Irrigation occurred on 11th July and lasted for 4 days for a total water volume of 146 m³/ha (Venterea and Rolston 2000b). Available data consisted of water saturation (volume of water per volume of pores), pH, concentrations of NH₄⁺, N O₃⁻, and NO₃⁻ solutes at 0–5 cm depth, and NO, N₂O, and CO₂ fluxes measured at various times for 20 days after fertilization (more details can be found in Venterea and Rolston (2000a, b)). The specific data set was chosen because it covered the main reactants, products, and intermediates in soil N cycling, while we acknowledge that the precise model structure can only be approximated rather than determined for any arbitrary field data.

2.3 Multi modeling approach

In this work, we considered three different models of the biogeochemical N cycle; each model was derived from the underlying model TOUGHREACT-N (Maggi et al. 2008; Gu et al. 2009) but each of them was characterized by a different level of complexity. The three models TRN1 (complex with high complexity), TRN2 (middle with intermediate complexity), and TRN3 (simple with low complexity) were comprised of mechanistic representations of physical, chemical, and biological aspects. Complex model TRN1 used here was previously presented in Maggi et al. (2008) (Fig. 1). The middle model TRN2 (Fig. 1) differed from TRN1 in the description of denitrification, which was performed only by DENs as compared to TRN1. By consequence, the metabolism of AOBs was simplified in TRN2 in a way such to not contribute to CO₂ production from consumption of CH₃O as during nitrifier denitrification. The simple model TRN3 (Fig. 1) differed from TRN1 because it employed an equivalent microbial biomass lumping all microbial biomass diversity into one species. This microbial biomass (defined by B in Fig. 1) was alone responsible for the full biogeochemical cycle of nitrogen and drastically simplified the complex respiration scheme of TRN1 and TRN2, the network on electron donor and acceptor feedback, and the inhibition effects.

In each of the three models, the soil was modeled through a one-dimensional vertical mesh of the top 60 cm of column discretized by dz = 1.25 cm. The column depth encompasses the dynamically active zone for N cycle reactions in the agricultural field experiment described by (Venterea and Rolston 2000b). The initial soil–water
saturation, primary species and biomass concentrations within the soil column were assigned according to the measured values or obtained from calibration (Maggi et al. 2008, Table 2). The Dirichlet boundary condition was assigned to the bottom boundary. Partial pressures of the gaseous species at the soil surface were kept constant and equal to 0.2 bar for O$_2$ (g) and to $4 \times 10^{-4}$ bar for CO$_2$ (g), and equal to zero for all other gases. Surface fluxes of NO, N$_2$O (g), CO$_2$ (g), and O$_2$ (g) were computed from soil surface concentration gradients. The soil moisture dynamics was modeled with the full mass-conserving Richards equation taking into account the water tension-saturation relationship according to van Genuchten (1980). Transport of chemical species was modeled with the advective-diffusive Fick’s law in both liquid and gaseous phases. Biologically-mediated chemical reactions were modeled with the Michaelis–Menten kinetics coupled with the Monod kinetics for the microbial biomass dynamics and accounting for electron donor, acceptor, and inhibitor concentrations. The chemical reactions of aqueous complexation, gaseous dis-solution and exsolution, and solute adsorption and desorption were described as equilibrium reactions. Note that temporal dynamics of soil pH is directly simulated by tracing production and consumption of protons based upon stoichiometric reaction equations (Maggi et al. 2008, Table 1b and 1c). A detailed description of the formulations used in these models, which is beyond the purpose of this work, can be found in Maggi et al. (2008).

The soil physical and hydraulic properties were previously determined by calibrating the model to the observed soil water saturation dynamics (Maggi et al., 2008). A silt loam soil was used with density of 2600 kg m$^{-3}$, porosity of 0.6, permeability of $1.82 \times 10^{-13}$ m$^2$, residual water saturation of 0.001, and van Genuchten parameter of 0.62. The derivation of chemical transport parameters were described in Maggi et al. (2008). The rest of the biogeochemical parameters are subject to calibration and will be discussed in the following sections.

2.4 Calibration approach

Automatic calibration was conducted using the Parameter Estimation (PEST) software (Doherty 2004) which implements a particularly robust variant of the Gauss–Marquardt–Levenberg method of nonlinear parameter estimation. During the calibration process, the different sets of parameters were adjusted at different model complexity levels (Table 1). Seven subobjective functions that were based on multiple variable groups were combined to form a single composite objective function. Weights were assigned to each subobjective function to ensure that the contributions of each group to the multiple objective functions were almost equal so that no subobjective function dominated the inversion process. It is acknowledged that the process of assigning weights to components of a multicomponent objective function can be subjective. In an
attempt to remove such subjectivity, two different weighting strategies were applied as a part of the calibration process (Table 2). The objective function, $U$, can be stated as the weighted sum of squares of subobjective functions:

$$U = \sum_{i=1}^{m} \delta_{i} r_{i} b_{i}^{2}$$

where $\delta_{i}$ is the weight attached to the $i$th objective variables, $m$ the number of variables, and $r_{i}$ expresses the $i$th subobjective function that represents the difference between the model output and the specific observation group. Weights were inversely proportional to the mean absolute values of observations. The values for $\delta_{i}$ were used as empirical weights with a goal to make the contributions of different model variables to $U$ similar in size and, therefore, give all measured variables a similar influence on the estimates of the parameters (Table 2).

Detailed process based N biogeochemical models such as TOUGHREACT-N are usually overparameterized with respect to given sets of observations. The involvement of too many parameters may lead to nonuniqueness in their estimation and possibly poor performance of the estimation due to consequential numerical instability. Temporary parameter immobilization technique was used to tackle this problem (Doherty 2004). In implementing this scheme PEST selects the most insensitive parameter, and temporarily removes it from the optimization process. With the dimensionality of estimable parameter space thus reduced, the parameter upgrade vector is recalculated. A model run is then conducted to compute the new objective function. Unless the new objective function has fallen by a significant amount, the next most troublesome parameter is temporarily frozen, and the parameter upgrade calculation procedure is repeated.

### 2.5 Bayesian model averaging

Bayesian model averaging (BMA) is a statistical scheme that can drive probabilistic ensemble predictions from competing individual deterministic predictions (Madigan et al. 1996; Raftery et al. 2003). The BMA expected value is a combination of a set of competing models such that the more skillful models receive higher weights. The variance of the BMA reflects the model predictive uncertainty in the
predicted BMA. BMA combines models for each output variable based on the assumption that at each particular time step, there is only one best ensemble member or model. Since we have more than one output variable in this study, BMA was applied separately to combine the model ensembles for each individual model variables.

Let’s consider the forecasted variable, $y$, (or predictand) and the ensemble of all considered model predictions, $f = \{f_1, f_2, \ldots, f_l\}$. $P_t(y_j|x_k, D)$ is the posterior distribution of $y$ given model prediction $x_k$ and observational data set $D$. The posterior distribution of the BMA prediction is therefore given as:

$$p(y_j|x_1; f_1; \ldots; f_l; D) \propto p(y_j|x_k; D) \cdot p(x_k|D)$$

where $p(x_k|D)$ is the posterior probability of model prediction $x_k$ being the best one, also known as the likelihood of model prediction, $x_k$ being the correct prediction given the observational data, $D$. If we denote $w_k = p(x_k|D)$, we obtain $\sum_{k=1}^{K} w_k = 1$: $w_k$ values are nonnegative and reflect relative contribution of each ensemble member to the predictive skill of BMA over the training period. The posterior mean and variance of the BMA prediction can be expressed as (Raftery et al. 2003, 2005):

$$E_y [y|x_1; f_1; \ldots; f_l; D] \propto \sum_{k=1}^{K} w_k f_k$$

$$\text{Var}_y [y|x_1; f_1; \ldots; f_l; D] \propto w_k f_k$$

In essence, the expected BMA prediction is the average of individual predictions weighted by the likelihood that an individual model is correct given the observations. The BMA prediction receives higher weights from better performing models as the likelihood function measures the agreement between predictions and the observations. In addition, the BMA variance is an uncertainty measure of the BMA prediction. The right hand side of Eq. 4 contains two terms: the between-model-variance which presents between model spread (ensemble spread) and the within-model-variance ($r^2$) which measures the expected uncertainty conditional on one of the forecasts being the best at any given time step. The within model variance or ($r^2$) accounts for under-dispersivity of the ensembles. Equation 4 is a better description of predictive uncertainty than that in a non-BMA scheme, which estimates uncertainty based only on the ensemble spread (i.e., only the between-model variance is considered), and consequently results in under-dispersive predictions (Raftery et al. 2003).

To implement the BMA scheme numerically, several assumptions are made for this study. First, the model ensemble used in BMA was assumed to be representative of the entire model space and that the individual ensemble members were independent of each other. Second, the conditional probability $p(x_k; D)$ was assumed Gaussian. The Gaussian assumption was made here for computational simplicity and BMA scheme could be applied assuming other probability distributions. We estimate the model parameters, $w_k$ and $r^2$, $k = 1, 2, \ldots, K$ (denoting $h = \{w_k k = 1, 2, \ldots, K\}$, $r^2$ by the maximum likelihood from the training period. The maximum likelihood estimator, estimates the value of the parameters under which the observed data is most likely to be observed. For both algebraic simplicity and numerical stability, here we maximize the logarithm of the likelihood function (log-likelihood functions) instead of the likelihood function. The log-likelihood function was approximated as:

$$\delta_{h|D} \propto \log \sum_{k=1}^{K} w_k \cdot p(x_k; D)$$

Because there is no analytical solution for $h$, an iterative procedure that made use of the Expectation-Maximization (EM) algorithm was used for this purpose (Raftery et al. 2005). The EM algorithm casts the maximum likelihood problem as a “missing data” problem. The missing data may not be actual data. Rather, it can be a latent variable that needs to be estimated. For this study, a latent variable $z_{k,t}$ is introduced. If the $k$th model ensemble is the best prediction considering the value of the parameters, at time $t$, $z_{k,t} = 1$; otherwise $z_{k,t} = 0$. Even though the estimates of the $z_{k,t}$ (Fig. 2) are not necessarily integers, at any time $t$, only one $z_{k,t}$ can be equal to 1 and the rest is equal to 0. As in the namesake, the EM algorithm alternates between the E (or expectation) step and the M (or maximization) step. It starts with an initial guess, $h^{0}$, for parameter, $h$. In the E step, $z_{k,t}$ is estimated given the current guess of $h$. In the M step, $h$ is estimated given the current values of the $z_{k,t}$. The EM steps are repeated until certain convergence criteria are satisfied. The EM algorithm is illustrated in Fig. 2. For a more detailed description of the EM algorithm, readers are referred to McLachlan and Krishnan (1997).

2.6 Model evaluation

The performance of the Bayesian Model averaging approach should be evaluated in two terms. The first measure is to determine the overall accuracy of the predictive mean in capturing observations. Second, we should measure the consistency of the Probability Density Function (PDF) of prediction with the actual observations. To evaluate the accuracy of individual model predictions
Initialize the weight $w_0^k = 1/K$

Compute the initial likelihood $K(0) = \log(w_p(y | y_t)) \sum_{j=1}^{j-1} \sum_{t=1}^{t_k} z_{jt}^{k_t}$

Select the number of iterations ($j$), training period ($t$), number of models ($K$), set $j=1, t=1$

Start

Read the models $M_k, k=1, ..., K$, and their simulations $M_k = (y_{k1}, ..., y_{kt})$

Calculate the standard deviation of the individual model simulation from observed values, $\sigma^{(0)}, k=1, ..., K$

Initialize the weight $w_k = 1/K$

Compute the initial likelihood $\xi^{(0)} = \sum_{i} \log(\sum_{k} w_k^{(0)} p(y_i | y_t))$

Calculate the weights (likelihood of the model $K$ to be true),

$Z_{jt}^{(j)} = \frac{w_k g(\bar{y}_i | y_t, \sigma^{(j-1)})}{\sum_{l=1}^{t_k} w_l g(\bar{y}_i | y_t, \sigma^{(j-1)})} \Rightarrow w_{jt}^{(j)} = \frac{1}{n} \sum_{j=1}^{j-1} Z_{jt}^{(j)}$

Update the variance (standard deviation)

$\sigma^{(j-1)}_{t} = \frac{1}{T} \sum_{t=1}^{T} (\bar{y}_t - y_t)^2$

Update the likelihood

$\xi^{(j)} = \sum_{i} \log(\sum_{k} w_k^{(j)} p(\bar{y}_i | y_t))$

$(\xi^{(j)} - \xi^{(j-1)}) < CC$

No $\rightarrow j = j+1$

Yes

Stop

Fig. 2 EM flow chart for each variable

RMSE $\frac{\sum_{i} T_{i} y_{t}^{obs} - y_{t}^{est}}{T}$

ABSE $\frac{\sum_{i} T_{i} y_{t}^{obs} - y_{t}^{est}}{T}$

where $y_{t}^{obs}$ and $y_{t}^{est}$ are observed and estimated output values, $T$ denotes the number of observations. In order to evaluate the consistency of the uncertainty bounds against the observations, we need to determine what percentage of observation point’s fall within the 95% uncertainty bounds, which here represents predictive uncertainty. The 95%
uncertainty bounds are expected to be wide enough to capture 95% of the observation points.

3 Results

3.1 Calibration of TOUGHREACT-N

The calibrated parameter values for the simple, middle, and complex models for weighting scheme 1 are presented in Table 1. In general, the parameters governing the nitrogen dynamics of the system differ largely across the complexity levels. Figure 3 shows calibration results of the simple, middle, and complex models with two weightings. These results show that none of the models are sufficient to fully reproduce the observed temporal patterns of chemical concentrations and fluxes. The pulse behaviors of N solutes and gases were difficult to reproduce. Exact fitting of the peak concentration or flux was compromised in order to insure a good fit of the simulation to the data sampled throughout the whole calibration period. The performance of the middle model is much better than the simple model regarding pH and NO$_3^-$ concentration, while there are no other significant improvements compared to the simple model. The complex model showed improved performance compared to the simple and middle models in terms of capturing the temporal dynamics of soil gaseous fluxes (i.e. NO, N$_2$O, CO$_2$). However, the simulations are still poor: the peaks of NH$_4^+$ and NO$_2^-$ concentrations, the CO$_2$ flux and the final NO$_3^-$ concentration are all underestimated. It is also worth to note that the calibration results are very sensitive to the two different weighting sets. The first weighting scheme emphasizing solute concentration variables produced better performance on capturing soil solute dynamics, while gaseous fluxes were better predicted by the second weighting scheme that stressed on soil gaseous flux variables (Fig. 5).

3.2 Assessment of predictive uncertainty

The generated set of six ensemble members was used in the Bayesian Model Averaging approach. The goal was to see if the application of multi-model combination approaches such as BMA can help to improve the prediction of solute concentrations and gas emissions by accounting for the model uncertainty. The disparity in the ensemble set for each output variable and their inaccuracy in comparison to observed data (Fig. 3) demonstrates the incompetence of the individual models in capturing the entire chemical and biological processes in the field. By considering various model structural formats we expected that some of these structural deficiencies can be captured. However, we

Fig. 3 The calibration results for the simple, middle, and complex model with two different weighting sets
should note that the skill of BMA is still dependant on the skill of individual member models (ensembles) that participate in multi-model combination. BMA’s 95% uncertainty bounds for all the solute concentrations and gas emissions capture most of the observation point (Fig. 4).

Figure 5 depicts DRMS and DABS of the ensemble models as well as BMA predictive mean. In this figure we also have included the result of simple model averaging which simply combines the ensemble members, assigning equal weights to all of them. BMA clearly outperforms the simple model averaging approach in most cases which emphasizes the importance of more sophisticated ways of combination and weight estimations.

We applied BMA to the six models that were associated with every output (seven of outputs) of the THROUGH REACT-N. The estimated weights are presented in Fig. 6. Not every model is included in the combination, since BMA assigns negligible weights to the models that do not contribute significantly to the combination results. Figure 5 depicts that intermediate and complex models are the most selected models among all contributing models (both nine times) while the simple models are the least selected ones (only five times). The percent coverage of 95% uncertainty bounds for all outputs is also presented here (Fig. 7).

4 Discussion

In this study we investigated the use of Bayesian Model Averaging method to improve the prediction of soil N cycling with respect to soil N solute concentrations and N gas emissions. We generated three different modeling schemes with different levels of complexity (Fig. 1) based on THOUGHREACT-N model. These models were calibrated using PEST automatic calibration algorithm based on two different sets of weights for objective function. This created six different deterministic model outputs which comprised the ensemble. The ensembles showed a diverse behavior of the system which shows the sensitivity of the system to the different mathematical representations and parameter sets. The model performances appear to be sensitive to model complexity levels (Figs. 4, 5), which is consistent with the finding that model complexity is one of most important causes of model prediction variability (Homann et al. 2000). To assess model predictive uncertainty, we used a BMA combination scheme to weigh individual deterministic model predictions based on the likelihood of the prediction being correct given the observations. The BMA weights are the estimated posterior model probabilities, representing each models relative skill (Fig. 6). Models with negligible

Fig. 4 BMA predictive mean (stared-line) and the associated model predictive uncertainty bounds associated with them (gray area). The ensemble models are presented in solid lines and observed values are presented as a dotted-line
We applied BMA scheme on seven of the THOUGH-REACT-N output variables individually. The deterministic model ensembles (solid lines) along with the BMA predictive mean (stared line) and uncertainty bound (gray area) for each output were presented in Fig. 4 with the observed values (dotted line). One can see that skill of BMA is limited to the range of the ensemble members. When none of the models capture an event, the combination of them is not going to predict that specific event either (Fig. 4). The uncertainty bounds tend to capture most of the observations in this study while they are visually not too wide (Fig. 4). One of the main drawbacks in this study has been the length of the observed data. With only nine data points available, it is very hard to train the BMA model perfectly putting aside evaluating it. This period does not capture many different events within the field. Even with this limitation we can see the BMA has done a very good job in capturing many observation points (Fig. 7).
while they stayed narrow enough and also they follow the trend of the observations (Fig. 4). In an ideal case with long enough data set, one would expect that the 95% uncertainty bounds would capture 95% of the observation points. Higher or lower percent coverage is an indication of too wide or too narrow uncertainty bounds. However, in this study because of the small number of data points every point that falls outside of the uncertainty bounds has a significant impact on the percent coverage. The 95% uncertainty bounds seem to be capturing 90% to 100% of the observation points with the exception of N₂O (Fig. 7). The uncertainty bounds on N₂O are too narrow compared to the uncertainty bounds on the other output variables, just capturing 80% of the observations (Fig. 7). This indicates that the models included in the BMA modeling approach to predict N₂O are not accurately representing the model space. The contribution of each model differs from one set of observations to another. The estimated weights are shown in Fig. 6. The results show that different models have different strength so that it is impossible to select a single best model that matches all observations well. Generally, intermediate and complex models have been selected more often over simple ones. These results suggest that there is no single best complexity level.

Figure 8 shows the contribution of between model variance (light gray area) or in another word under-dispersion of the ensembles to the estimation of total model predictive uncertainty (dark gray area). This is evident just by considering different model structures and analyzing their spread, one can underestimate the predictive uncertainty. One of the strengths of BMA is inclusion of the

---

**Figure 7** Percent observations within the BMA 95% prediction uncertainty bounds

**Figure 8** Total model predictive uncertainty versus uncertainty drawn solely from ensemble spread
effect of both ensemble spread and under-dispersion in calculating the predictive uncertainty. The between-model variance in some degree represents the limitation of our ensemble pool, and importance of inclusion of other modeling schemes or improvement of existing models. The information about the susceptibility to model structure uncertainty of different state variables (model outputs) provides useful information for possible improvement of the model structure or to guide further data collection campaigns to optimally reduce model structure uncertainty. This is more evident in estimation of pH, NO$_2$ and NO$_3^-$, where we can see many observation point fall outside of selected ensemble spread by BMA. This means that the existing model structures in hand are incapable of capturing the exact physical and biological processes that relates to these state variables (model outputs).

Close look at the statistical summary of the model performances (Fig. 5) reveals that BMA outperforms the simple model averaging approach in almost all the cases (except the RMSE of NO$_2^-$ and NO$_3^-$) which assigns uniform weights to all the models. RMSE best examines the performance of the model in capturing extreme events while ABSE gives the overall performance of the models. In case of NO$_2^-$ and NO$_3^-$ simulations, looking at the RMSE values (Fig. 5), it seems that simple model aver-aging has done a better job in capturing the extreme events compared to BMA. Meanwhile studying ABSE values (Fig. 5), BMA outperform simple model averaging in overall. The trend of estimated model weights is not necessarily consistent with the trend of two statistical measures (RMSE and ABSE) we presented here. This is because in estimating the model weights, here we used maximum likelihood as our objective function which does not necessarily results in the same pattern as RMSE and ABSE. Analyzing Fig. 5, we can see that application of BMA improves prediction of gas emissions especially CO$_2$ and N$_2$O, more significantly compared to the predictions of solute concentrations. Again we should emphasize here that the strength of BMA is not just improving the predictive mean but also more importantly, BMA directly accounts for model predictive uncertainty, which is often neglected in traditional uncertainty analyses, and, therefore, it yields more accurate estimations of the predictive uncertainty compared to estimations based on a single model.

One limitation of the current study is its reliance on the deterministic calibration of parameters defined for each alternative conceptual model. In the case of multi-model combination approaches that include a deterministic calibration step, errors in the model structures will be compensated by biased parameter estimates in order to optimize model fit during the calibration (Beven 2006). Sensitivity of our model outputs to the estimated parameter sets using PEST (Fig. 3) is the evidence of such limitation. Therefore, accounting for parameter uncertainty using probabilistic parameter estimation approaches (Kuczera and Parent 1998; Vrugt et al. 2003) can help to reduce this effect. In general to gain the best and most accurate predictive mean and uncertainty estimation, it is important to assess all sources of uncertainty simultaneously (Ajami et al. 2007).

The current study has brought out the benefits of combining biogeochemical models with different complexity levels to produce a more robust result that better captures uncertainties and also lead to insight about the system behavior. There may be several model outputs that may contain information of direct bearing to different levels of complexity. In the current soil N model, for example, the simple model complexity did better for solute concentrations whereas the complex model did better for gaseous emissions. This might suggest a compromised model structure between disaggregated microbial functional groups (i.e., AOB, DEN, and NOB) and lumped microbial kinetics. Thus, the standard concepts of “Reducing model complexity” (Lawrie and Hearne 2007) and “choosing the right complexity” (Homann et al. 2000) do not reconcile the inherent model deficiency and neglect uncertainty in the choice of models. This uncertainty may be important especially if there are several models that differ in predictions but have similar fitting criterion scores (Gibbons et al. 2008). The alternative approach proposed by this study is multi-model ensemble by BMA, which indeed enhances prediction capacity, so as to outperform the best participating single model. In a real model prediction practice, it may be difficult or impossible to know which model is the best, since model performance varies with predictand. The best model is usually not the same for all variables. In such a setting, model averaging can provide a pragmatic method for issuing an optimum prediction.

Acknowledgements This work was supported by the Berkeley Water Center (BWC) and by the Laboratory Directed Research and Development (LDRD) funding from Berkeley Lab, provided by the Director, Office of Science, of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors would like to acknowledge Dr. Federico Maggi for providing us with the THOUGHREACT-N model. We also would like to thank Dr. Hornberger, Dr. Ye and the anonymous reviewers for their valuable and constructive comments.

Open Access This article is distributed under the terms of the Creative Commons Attribution Noncommercial License which permits any noncommercial use, distribution, and reproduction in any medium, provided the original author(s) and source are credited.

References
Ajami NK, Duan Q, Gao X, Sorooshian S (2006) Multi-model combination techniques for hydrological forecasting: application to distributed model intercomparison project results. J Hydrometeorol 7(4):755–768
