Stochastic Environmental Research and Risk Assessment (2018) 32:2809–2819
https://doi.org/10.1007/s00477-018-1592-3

ORIGINAL PAPER

Relative model score: a scoring rule for evaluating ensemble simulations with application to microbial soil respiration modeling

Ahmed S. Elshall[1,8] · Ming Ye[1,2,3] · Yongzhen Pei[3] · Fan Zhang[4] · Guo-Yue Niu[5,6] · Greg A. Barron-Gafford[5,7]

Published online: 4 August 2018

Abstract
This paper defines a new scoring rule, namely relative model score (RMS), for evaluating ensemble simulations of environmental models. RMS implicitly incorporates the measures of ensemble mean accuracy, prediction interval precision, and prediction interval reliability for evaluating the overall model predictive performance. RMS is numerically evaluated from the probability density functions of ensemble simulations given by individual models or several models via model averaging. We demonstrate the advantages of using RMS through an example of soil respiration modeling. The example considers two alternative models with different fidelity, and for each model Bayesian inverse modeling is conducted using two different likelihood functions. This gives four single-model ensembles of model simulations. For each likelihood function, Bayesian model averaging is applied to the ensemble simulations of the two models, resulting in two multi-model prediction ensembles. Predictive performance for these ensembles is evaluated using various scoring rules. Results show that RMS outperforms the commonly used scoring rules of log-score, pseudo Bayes factor based on Bayesian model evidence (BME), and continuous ranked probability score (CRPS). RMS avoids the problem of rounding error specific to log-score. Being applicable to any likelihood functions, RMS has broader applicability than BME that is only applicable to the same likelihood function of multiple models. By directly considering the relative score of candidate models at each cross-validation datum, RMS results in more plausible model ranking than CRPS. Therefore, RMS is considered as a robust scoring rule for evaluating predictive performance of single-model and multi-model prediction ensembles.

Keywords Scoring rule · Continuous ranked probability score · Bayes factor · Log-score · Dispersion · Reliability

1 Introduction

Environmental modeling is subject to conceptual mofrl uncertainty, and using multiple models to address the model uncertainty is becoming a common practice. Multi-model analysis such as model averaging (Ajami et al. 2007; Exbrayat et al. 2013; Hargreaves et al. 2012; Liu et al. 2016; Lu et al. 2015; Poeter and Anderson 2005; Poeter et al. 2005; Wenger et al. 2013; Winter 2010; Winter and Nychka 2010; Ye et al. 2010; Zeng et al. 2018) or model selection (Elshall

✉ Ming Ye
mye@fsu.edu

[1] Department of Scientific Computing, Florida State University, Tallahassee, FL, USA

[2] Department of Earth, Ocean, and Atmospheric Science, Florida State University, Tallahassee, FL, USA

[3] School of Computer Science and Software Engineering, Tianjin Polytechnic University, Tianjin 300387, China

[4] Key Laboratory of Tibetan Environmental Changes and Land Surface Processes, Institute of Tibetan Plateau Research, Chinese Academy of Sciences, Beijing 100101, China

[5] Biosphere 2, University of Arizona, Tucson, AZ, USA

[6] Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ, USA

[7] School of Geography and Development, University of Arizona, Tucson, AZ, USA

[8] Present Address: Department of Geology and Geophysics, and Water Resources Research Center, University of Hawaii Manoa, Honolulu, HI, USA

and Tsai 2014; Foglia et al. 2013; Nowak et al. 2012; Schöniger et al. 2014; Zhang et al. 2014; Wöhling et al. 2015) evaluates and ranks multiple candidate models based on their expected utility. A utility function is an important component in multi-model analysis as it defines the gain of the following situations: (1) choosing one model over another, (2) choosing one model averaging method over another, and (3) choosing model averaging over individual models. For the convenience of discussion, we will not distinguish model averaging from individual models in this paper, unless specified otherwise. In other words, the discussion of this paper applies to the results of both individual models and model averaging. Examples of utility functions are a model adequacy criteria for model selection, a model weighting criteria for model averaging, or a predictive performance metric for prediction evaluation.

Predictive performance metrics are of particular use in multi-model analysis, since a typical feature of a better model is that the model makes better predictions. Therefore, predictive performance metrics can be used for ranking alternative models. Since future events are unknown in practice, evaluating model predictive performance is always done in a manner of cross-validation, in which a dataset is split into two parts: calibration data for model calibration and cross-validation data for evaluation of predictive performance. The discussion of predictive performance hereinafter is in the context of cross-validation. The evaluation requires measuring to what extent the cross-validation data can be predicted, reflecting a model's capability of not only predicting future events but also measuring the uncertainty associated with the predictions. Therefore, model predictions are routinely probabilistic in nature, taking the form of a probability distribution for each prediction obtained using Monte Carlo (MC) methods through repeated random sampling (Dawid 1984; Gneiting and Raftery 2007; Kumar 2011).

Probabilistic predictive performance can be assessed by at least three basic criteria: mean, dispersion, and reliability of the prediction ensemble (Hill and Tiedeman 2007). With respect to mean, the prediction ensemble is said to be *accurate* when the ensemble is centered on the cross-validation data. With respect to dispersion, the prediction ensemble is said to be *precise* when the prediction interval has a narrow band. With respect to reliability, the prediction ensemble is said to be more *reliable* when the prediction interval bands bracket more cross-validation data. Figure 1 illustrates the differences between the three criteria. Figure 1a–c shows predictive performance failure with respect to reliability (i.e., bracketing the cross-validation data), despite that the predictions in Fig. 1b have adequate mean and dispersion. Figure 1d, e shows adequacy with respect to reliability, while the predictions in Fig. 1d are poor with respect to mean and dispersion. The predictions in Fig. 1f are adequate with respect to all the three criteria. To aid assessing this adequacy, several single-criterion metrics focusing on a single prediction criterion are commonly used (Ajami et al. 2007; Gulden et al. 2008; Zhang et al. 2014). With respect to mean, squared residual error metrics such as root mean squared error (Anderson and Woessner 1992) and Nash-Sutcliffe model efficiency (Nash and Sutcliffe 1970) are the most commonly used metrics (Zhang et al. 2014). With respect to dispersion, the commonly used metrics to measure the precision of the prediction interval include ensemble standard deviation (Yokohata et al. 2012) and sharpness (Smith et al. 2010). With respect to reliability, several simple to complex reliability metrics have been used (Hoeting et al. 1999; Annan and Hargreaves 2010; Annan et al. 2011; Smith et al. 2010; Oldenborgh et al. 2013). However, it is not uncommon to visually assess reliability (Laloy and Vrugt 2012), since there is usually a trade-off between reliability and dispersion. In the absence of any preference toward any single predictive performance criterion, the single-criterion metrics are generally insufficient for judging the overall predictive performance, since there is always puzzling trade-off between different criteria that focus on different aspects of prediction (as illustrated in detail below). Thus, in addition to using the single-criterion metrics, scoring rules are needed to provide a summary measure of the overall probabilistic predictive performance.

Gneiting and Raftery (2007) provided an excellent theoretical and critical review of different scoring rules, and they defined scoring rules as follows: "scoring rules assess the quality of probabilistic forecasts, by assigning a numerical score based on the predictive distribution and on the event or value that materializes". Scoring rules are useful for model elicitation and evaluation by ranking candidate models and prediction procedures based on their overall predictive performance. Among the scoring rules discussed in Gneiting and Raftery (2007), this study is focused on log-score (Good 1952), continuous ranked probability score (CRPS, Matheson and Winkler 1976; Hersbach 2000), and pseudo Bayes factors (Dawid 1984), which have been widely used in environmental science and engineering.

This paper defines a new scoring rule called relative model score (RMS) for evaluating the overall predictive performance. RMS ranks candidate models based on the overall quality of their probabilistic predictions by using the probability density function (PDF) of the prediction ensemble. Therefore, similar to other scoring rules, RMS simultaneously accounts for accuracy, dispersion and reliability. However, as discussed in Sect. 2 and demonstrated in Sect. 4, RMS outperforms log-score, CRPS, and pseudo Bayes factors. The comparison between RMS and the other

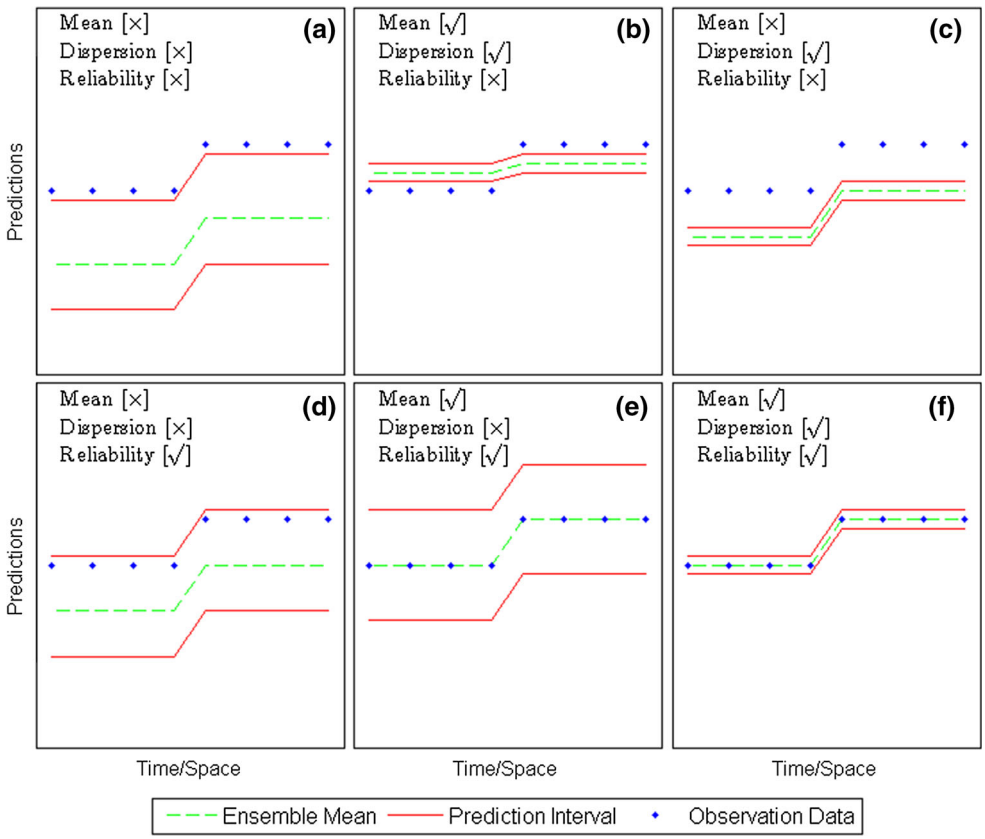

Fig. 1 Predictions with adequate (check sign) and inadequate (cross sign) predictive performance with respect to mean, dispersion and reliability

three scoring rules is focused on the numerical and practical issues that the log-score, CRPS, and pseudo Bayes factors cannot handle. Specifically speaking, log-score suffers from rounding error, CRPS may give inaccurate ranking, and pseudo Bayes factors are only applicable for evaluating models with the same likelihood functions. These issues among other (as explained in the next section) motive us to develop the RMS that has a simple form, but has not been reported in literature to the best of our knowledge.

2 Relative model score

This section first defines relative model score, discusses its numerical evaluation, and demonstrates its use to address the trade-off between accuracy, precision, and reliability. Subsequently, the advantages of RMS relative to log-score, CRPS, and pseudo Bayes factor are discussed.

2.1 Definition and numerical evaluation of RMS

RMS is a skill score, and its definition is similar to that of mean score discussed in Gneiting and Raftery (2007). In the context of fitting a parametric model, P_θ, for estimating parameter θ, based on data, $\mathbf{D} = \{D_1, D_2, \ldots, D_n\}$, the mean score measures the goodness-of-fit as

$$S_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} S(P_\theta, D_i), \tag{1}$$

which is aggregated over the score of individual datum $S(P_\theta, D_i)$. Here, $S(P_\theta, D_i)$ denotes a scoring rule that measures the accuracy of the probabilistic forecasts for P_θ to predict D_i. For example, if P_θ gives a probabilistic forecast, $p(D)$, for event, D, and the true outcome of the event is D_i (at a given location and time), then one can assign a score to P_θ as $S(P_\theta, D_i) = S(p, D_i)$, which can be either a linear score, $S(p, D_i) = p(D_i)$, or a log score, $S(p, D_i) = \log p(D_i)$. For Eq. (1) to be valid, $S(P_\theta, D_i)$ needs to be a strictly proper scoring rule, whose definition is referred to Gneiting and Raftery (2007). Examples of strictly proper score rule are the Brier score and the logarithm score for assessing and comparing probabilistic forecasts. In line with Eq. (1), to measure the overall predictive performance of any given model, M_j, relative to other models, RMS for datum, D_i, of the quantity that all the models predict is defined as

$$RMS(M_j, D_i) = \frac{p(D_i|\mathbf{Y}_{i,j}, M_j)}{\sum_{k=1}^{m} p(D_i|\mathbf{Y}_{i,k}, M_k)} \quad (2)$$

where m is the number of models, and $\mathbf{Y}_{i,j}$ is the ensemble predictions of D_i made by model M_j using Monte Carlo methods. The use of $p(D_i|\mathbf{Y}_{i,j}, M_j)$, a linear score, is intuitively appealing, although it is not a proper scoring rule (Gneiting and Raftery 2007). The probability density function, $p(D_i|\mathbf{Y}_{i,j},M_j)$, of D_i is conditioned on the prediction ensemble. For the n-dimensional cross-validation data, $\mathbf{D}$, that all the models predict, Eq. (2) becomes

$$RMS(M_j, \mathbf{D}) = \sum_{i=1}^{n} \frac{p(D_i|\mathbf{Y}_{i,j}, M_j)}{\sum_{k=1}^{m} p(D_i|\mathbf{Y}_{i,k}, M_k)} w_i, \quad (3)$$

where w_i is a weighing coefficient for each cross-validation datum. For any candidate model M_j, RMS ranges between 0 and 1, and the RMS values of all the models satisfy $\sum_{j=1}^{m} RMS(M_j|\mathbf{D}) = 1$. A model with a higher RMS value is considered to have a better overall predictive performance relative to other models with lower RMS values.

The evaluation of RMS requires assigning weighting coefficient w_i and calculating the probability $p(D_i|\mathbf{Y}_{i,j}, M_j)$. The weighting coefficient,w_i, provides the flexibility of meeting various modeling objectives. For example, in flood prediction, larger weighting coefficients are always assigned to data of extreme events. The simplest weighting coefficient is $w_i = 1/n$, assigning equal weight to n cross-validation data. Other weighting coefficients are possible. For example, the weighting coefficient of $w_i = D_i \Big/ \sum_{j=1}^{n} D_j$ can be used to consider the magnitude of each datum. In the numerical example of this study, we use the weighting coefficient of $w_i = 1/n$ for RMS and the other three scoring rules.

Calculating the probability $p(D_i|\mathbf{Y}_{i,j}, M_j)$ is straightforward. The first step is to estimate the PDF, $p(\mathbf{Y}_{i,j}|M_j)$, of the ensemble prediction $\mathbf{Y}_{i,j}$ at each cross-validation datum using either parametric or nonparametric density function. After the probability density function of $\mathbf{Y}_{i,j}$ is obtained, the probability $p(D_i|\mathbf{Y}_{i,j}, M_j)$ is evaluated by the linear interpolation between the two closest bin centers (given by the kernel sdensity function) that bracket D_i (as illustrated below). To estimate the PDF of $\mathbf{Y}_{i,j}$, this study uses the MATLAB ksdensity function that implements a nonparametric technique of kernel density estimation (Smith 2014, p. 75). Kernel density estimation is a smoothing process that provides nonparametric representation of the PDF of data without making assumptions about the distribution of the data as the case with parametric distributions. The kernel density estimator, $\hat{p}(x_0)$, of the PDF at data location x_0 is

$$\hat{p}(x_0) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x_i - x_0}{h}\right) \quad (4)$$

where x_i is a data location around x_0 within the neighborhood of size h (also called bandwidth), n is the number of data within the neighborhood, and $K(\cdot)$ is the kernel (a weight function). Three commonly used kernels are Gaussian, Epanechnikov, and tri-cube kernels. More details of kernel density estimation are referred to Silverman (1998). We implemented MATLAB ksdensity function using the Gaussian kernel smoother with an unbounded support that allows the estimated density to extend over the whole real line, and no errors were observed around boundaries.

The scoring of RMS is illustrated in Fig. 2 for six cross-validation data. The data are carbon efflux measurements used in the soil respiration modeling discussed in the next section. Each of the data is predicted by two models (denoted as A and B), and the corresponding PDFs are shown in Fig. 2. We select six prediction PDFs without and with trade-off between accuracy, precision and reliability. In Fig. 2a, the PDF of the prediction ensemble of Model A brackets the datum, and Model B does not bracket the datum; the $p(D_i|\mathbf{Y}_{i,j}, M_j)$ values of the two models are $p(D_i|\mathbf{Y}_{i,A}, M_A) = 39.05$ and $p(D_i|\mathbf{Y}_{i,B}, M_B) = 0$. For this datum, the scores $p(D_i|\mathbf{Y}_{i,j}, M_j) \Big/ \sum_{k=1}^{2} p(D_i|\mathbf{Y}_{i,k}, M_k)$ of Model A and Model B are one and zero, respectively. Reliability in this case is the decisive factor as prediction ensemble of Model B does not bracket the datum. Figure 2b shows that Model A and Model B are reliable, as they both bracket the datum with the density values of 61.32 and 136.27, respectively. For this datum, Model B has a higher score of 0.69 in comparison with the score of 0.31 of Model A. In this example, there is no trade-off between accuracy and precision, since Model B is more accurate (i.e., the ensemble mean is closer to the datum) and more precise (i.e., the PDF is narrower) than Model A. Figure 2c–f shows four examples that the trade-off between accuracy and precision becomes puzzling. Figure 2c, d shows two examples that accuracy becomes more decisive than precision. Figure 2e shows an interesting case that, while Model B is more accurate and precise than Model A, Model A has a higher score. This is actually desirable since choosing Model A for this datum involve less risk, as the datum is at the far tail of the PDF of Model B. Figure 2f shows another interesting example that the trade-off between accuracy and precision is nearly equal. In this example, both models have received nearly equal RMS. In the numerical example of the next section, we illustrate how RMS aggregates the scores of all cross-validation data into a single metric to facilitate overall model ranking.

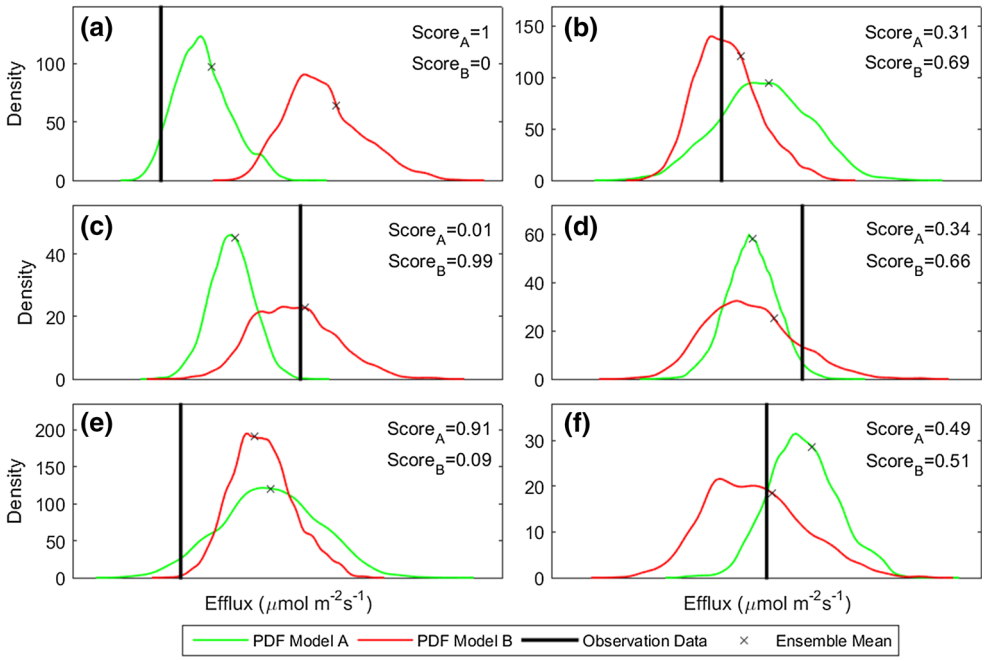

Fig. 2 Probability density function (PDF) of the prediction ensemble given by models A and B for six data, illustrating the trade-off between central mean tendency, dispersion, and reliability

2.2 Comparison with three other scoring rules

RMS has practical advantages in comparison to log-score, CRPS, and pseudo Bayes factor. Log-score is a logarithm scoring rule, and it is one of the most commonly used scoring rules for evaluating the overall predictive performance in groundwater hydrology (Lu et al. 2015; Shi et al. 2012; Xue and Zhang 2014; Ye et al. 2004) and for regression models in general (Hoeting et al. 1999). Log-score is a logarithm score, and defined as (Good 1952)

$$LS(M_j, \mathbf{D}) = -\log p(\mathbf{D}|\mathbf{Y}_j, M_j). \tag{5}$$

A drawback of log-score is that it suffers from rounding error, when certain cross-validation data are not bracketed by prediction ensemble and thus have extremely low probability, which is not uncommon in practice. The rounding error, as well known in numerical computing, is the difference between results produced using exact arithmetic and results produced using limited precision arithmetic. For example, a double-precision 64-bit floating-point number that is smaller than the smallest nonzero denormalized number $\varepsilon_{dn} \approx \pm 4.95 \times 10^{-324}$ is rounded to zero, which is also known as numerical underflow (Heath 1997). When data D_i is not bracketed by the prediction ensemble $\mathbf{Y}_{i,j}$, the corresponding $p(D_i|\mathbf{Y}_{i,j}, M_j)$ value may be smaller than ε_{dn} and thus rounded to zero. As a result, the corresponding log-score is $-\sum_{i=1}^{n} \log p(D_i|Y_{i,j}, M_j) = \infty$ because $\lim_{x\to 0^+} \log(x) = -\infty$, which does not permit the evaluation of model predictive performance. Although rounding error also happens to the numerical evaluation of RMS, its impact is negligible because RMS does not use $\log p(D_i|\mathbf{Y}_{i,j}, M_j)$ but $p(D_i|\mathbf{Y}_{i,j}, M_j)$.

CRPS is a continuous ranked probability score, and widely used in meteorology (Matheson and Winkler 1976). For an individual data, D_i, it is defined as

$$CRPS(M_j, D_i) = \int \left[F_{\mathbf{Y}_{i,j}}(y) - F_{D_i}(y)\right]^2 dy, \tag{6}$$

where y represents a value within the range of the ensemble $\mathbf{Y}_{i,j}$, $F_{\mathbf{Y}_{i,j}}$ is the cumulative distribution function (CDF) of the ensemble, and F_{D_i} is the CDF of the cross-validation datum that takes the form

$$F_{D_i} = \begin{cases} 0 & y < D_i \\ 1 & y \geq D_i \end{cases}. \tag{7}$$

For the n-dimensional data, $\mathbf{D}$, the rigorous definition of CRPS is (Gneiting and Raftery 2007)

$$CRPS(M_j, \mathbf{D}) = \int \left[F_{\mathbf{Y}_{i,j}}(\mathbf{y}) - F_{D_i}(\mathbf{y})\right]^2 d\mathbf{y}, \tag{8}$$

for all n-variate thresholds. The most commonly used form in practice is

$$\begin{aligned} CRPS(M_j, \mathbf{D}) &= \frac{1}{n}\sum_{i=1}^{n} CRPS(M_j|D_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \int \left[F_{\mathbf{Y}_{i,j}}(y) - F_{D_i}(y)\right]^2 dy, \end{aligned} \tag{9}$$

which is similar to the mean score discussed in Sect. 2.1.

For each cross-validation datum D_i, CRPS defines a predictive CDF scoring rule based on the squared area difference between the prediction CDF and the cross-

validation datum CDF. Then CRPS aggregates the scores of all cross-validation data into a single metric. Smaller CRPS indicates better predictive performance. The smallest CRPS value is zero, corresponding to the case that the prediction and cross-validation datum CDFs are identical, i.e., each realization of the prediction ensemble is identical to the cross-validation datum. This study evaluates Eq. (9) using the CRPS MATLAB toolbox developed by Shrestha (2014). Similar to RMS, CRPS does not suffers from the rounding error of log-score because CRPS uses CDF. RMS and CRPS measure the closeness of prediction distributions and corresponding cross-validation data using PDF and CDF, respectively. Yet RMS appears more plausible for the model comparison purpose as RMS considers the relative score of candidate models at each cross-validation datum.

The pseudo Bayes factor (Dawid 1984) is the same as the Bayes factor that has been used for model comparison and selection. Gneiting and Raftery (2007) related Bayes factor to scoring rule by defining a scoring rule as

$$S_i(M_j, \mathbf{D}) = \sum_{i=1}^{n} \log \Pr(\mathbf{D}|M_i). \tag{10}$$

The term *pseudo* refers to the use of cross-validation data rather than calibration data for evaluating the Bayes factor. For two models, the Bayes factor is

$$\text{Bayes factor} = \frac{\Pr(\mathbf{D}|M_i)}{\Pr(\mathbf{D}|M_j)} \tag{11}$$

where $\Pr(M_i)$ and $\Pr(M_i)$ are the prior model probability of models M_i and M_j, respectively, and $\Pr(\mathbf{D}|M_i)$ and $\Pr(\mathbf{D}|M_j)$ are the Bayesian model evidence (BME) of the two models, i.e., the likelihood that the models reproduce cross-validation data $\mathbf{D}$. The BME is defined as,

$$\Pr(\mathbf{D}|M_k) = \int p(\mathbf{D}|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) d\boldsymbol{\theta}_k, \tag{12}$$

where $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ is the joint likelihood of model M_k and its parameter set $\boldsymbol{\theta}_k$, and $p(\boldsymbol{\theta}_k|M_k)$ is the prior densities of parameters $\boldsymbol{\theta}_k$ under model M_k. The BME can be evaluated by using Laplace approximation or MC methods. To estimate BME this study uses the MC-based thermodynamic integration method (Lartillot and Philippe 2006; Liu et al. 2016), and the readers are referred to recent studies (Schöniger et al. 2014; Liu et al. 2016; Zeng et al., 2018) that survey numerical evaluation of BME.

Using the pseudo Bayes factor for model predictive performance has two limitations in comparison with RMS, log-score and CRPS. The first one is that pseudo Bayes factor needs to assume a distribution for the joint likelihood function, $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ because $\Pr(\mathbf{D}|M_k)$ is marginalized from $p(\mathbf{D}|\boldsymbol{\theta}_k, M_k)$ explicitly. Therefore, it cannot be used to rank model predictions produced by using different likelihood functions. Comparative studies of different likelihood functions are common in hydrology literature. Examples include the comparison of formal, informal and approximate Bayesian computation techniques (Vrugt et al. 2009; Sadegh and Vrugt 2013), assessment of different formal likelihood functions (Smith et al. 2010; Ricciuto et al. 2011), evaluation of different inference data approaches (Bulygina and Gupta 2011; Evin et al. 2014), and comparison of lumped versus segregated uncertainty quantification methodologies (Thyer et al. 2009; Renard et al. 2010). Using pseudo Bayes factor for such comparative studies is not applicable since different likelihood functions have different statistical meanings and different support scale for statistical hyper-parameters. The other limitation is that pseudo Bayes factor cannot be used to compare predictions of single-model ensembles (i.e., ensemble resulting from a single model) with predictions of multi-model ensembles (i.e., ensemble resulting from multiple models). Such comparisons are particularly desirable in the context of multi-model analysis (Yokohata et al. 2012). Therefore, being independent of likelihood functions and compatible with both multi-model and single-model ensembles, RMS, log-score, and CRPS have a wider range of applicability than pseudo Bayes factor does.

3 Illustrative examples

We present a challenging problem of model evaluation with six prediction ensembles generated by two individual soil respiration models (models 5C and 6C) using two likelihood functions (standard least square and skewed exponential power) and by model averaging of the two models for each likelihood function. The two soil respiration models are the five-carbon pool model (5C) and six-carbon pool model (6C) developed by Zhang et al. (2014) for simulating the nonlinear controls of soil temperature and moisture on microbial respiration. The models simulate the soil moisture control on soil organic carbon degradation through enzymatic catalysis and microbial uptake of degraded carbon. Model 6C is more complex than model 5C, as model 6C has an extra carbon pool with two additional parameters to account for carbon degradation through catalysis of enzyme in dry soil zones during dry periods. It is important to evaluate the predictive performance of the two models to determine whether the higher level of complexity for model 6C can be justified. For details about the two models, the readers are referred to Zhang et al. (2014).

To demonstrate the flexibility of using RMS for evaluating model prediction, we use two ensemble prediction schemes with two different likelihood functions. One likelihood function is the standard least squares (SLS)

likelihood that models the residual distribution using a Gaussian function. The other is the skewed exponential power (SEP) likelihood (Schoups and Vrugt 2010) that has two additional shape parameters to accounts for non-Gaussian residuals with varying degrees of skewness and kurtosis. This results in four combinations of physical models and likelihood functions, and they are denoted as 5C-SLS, 6C-SLS, 5C-SEP and 6C-SEP. In addition, we developed a multi-model ensemble using Bayesian model averaging (BMA, Hoeting et al. 1999) to average the predictions of the model 5C and model 6C under each likelihood function. This results in two additional models, denoted as BMA-SLS and BMA-SEP.

The dataset used in this example consists of 15,272 observations of carbon efflux obtained every half an hour over 1 year at an eddy-covariance flux tower (Zhang et al. 2014). The dataset is separated into two parts. The first two-third of the data are used to conduct Bayesian inverse modeling for each model and to evaluate the model weights for BMA. The last one-third of the data are used for cross-validation to evaluate predictive performance. By conducting a Markov chain Monte Carlo simulation using the DREAM code (Laloy and Vrugt 2012), the Bayesian inverse modeling gives posterior model parameter samples, which are used to produce the prediction ensembles for the cross-validation data. For calibration, to ensure convergence we run MCMC long enough with half million samples and burn-in the first 20% of the samples. For prediction, for each model a large sample size of half million is also needed for an accurate estimation of the PDFs and CDFs used in the scoring rules.

The prediction ensembles are created in different ways for the four individual models and the two averaged models. For each of the individual models, following Schoups and Vrugt (2010), we create the single-model ensemble by generating independent samples for the physical model parameters to account for parameter uncertainty and for the residual error model parameters to account for output uncertainty. The prediction ensemble of model averaging is obtained as the weighted average of the prediction ensembles of models 5C and 6C. The SLS-based model weights for 5C and 6C are 60.38% and 39.62%, respectively; they become 54.12% and 45.88%, when the SEP likelihood is used. Model weights are calculated by taking the harmonic mean of the posterior sample log-likelihoods, although using posterior samples can result in biased model weights (Schöniger et al. 2014; Liu et al. 2016). While other methods of evaluating model weights are available (Raftery et al. 2005; Diks and Vrugt 2010), this paper is not to evaluate techniques of evaluating model weights for BMA, but to demonstrate that RMS is applicable for comparing single-model ensemble and multi-model ensemble approaches.

The predictive performance of the six prediction ensembles are evaluated using RMS, log-score, CRPS, and BME. In addition, single-criterion metrics are also evaluated for understanding how the scoring rules provide a summary measure of the overall probabilistic predictive performance. Mean accuracy of the prediction ensemble is evaluated using root mean squared error $RMSE = \sqrt{\sum_{i=1}^{n} (D_i - \overline{Y_i})^2 / n}$, which aggregates the difference prediction ensemble mean $\overline{Y_i}$ and cross-validation datum D_i for n data into a single measure of accuracy. To evaluate dispersion, we use the sharpness metric $\text{Sharpness} = 1/n \sum_{i=1}^{n} [Max(\mathbf{Y}_i) - Min(\mathbf{Y}_i)]$ (Smith et al. 2010) to measure the average mean width of the prediction interval, where $\mathbf{Y}_i$ is prediction ensemble for cross-validation data, D_i. The smaller the sharpness, the less dispersed is the predication interval. Predictive coverage (Hoeting et al. 1999) is used to evaluate reliability by calculating the percentages of cross-validation data contained in a prediction interval, and the 95% credible interval is used in this study.

4 Results

This section first evaluates the predictive performance of the six prediction ensembles using RMS, and then shows the limitations of log-score, CRPS, and BME in evaluating the predictive performance. Figure 3 shows the means and 95% credible intervals of the six prediction ensembles during the cross-validation period along with their respective values of RMS and single-criterion metrics. In the absence of prior preference toward any criterion, using single-criterion metrics alone is insufficient for evaluating the predictive performance, because it is difficult to conclude which ensemble has the best predictive performance due to the trade-off between accuracy, precision and reliability. Taking models 5C-SEP (Fig. 3b) and BMA-SEP (Fig. 3f) as an example, while 5C-SEP is less accurate than BMA-SEP in terms of RMSE, 5C-SEP is more reliable than BMA-SEP in terms of predictive coverage; the two cases have the almost same dispersion as measured by sharpness. Thus, scoring rules are a necessary tool for assessing predictive performance of the six prediction ensembles.

RMS gives quantitative and unambiguous ranking for the predictive performance of the six prediction ensembles in the following descending order: BMA-SEP, 6C-SEP, BMA-SLS, 6C-SLS, 5C-SLS, and 5C-SEP. The ranking agrees with the visual inspection of the six prediction ensembles on the following aspects. First, the prediction ensembles of 6C-SEP and BMA-SEP stand out, because

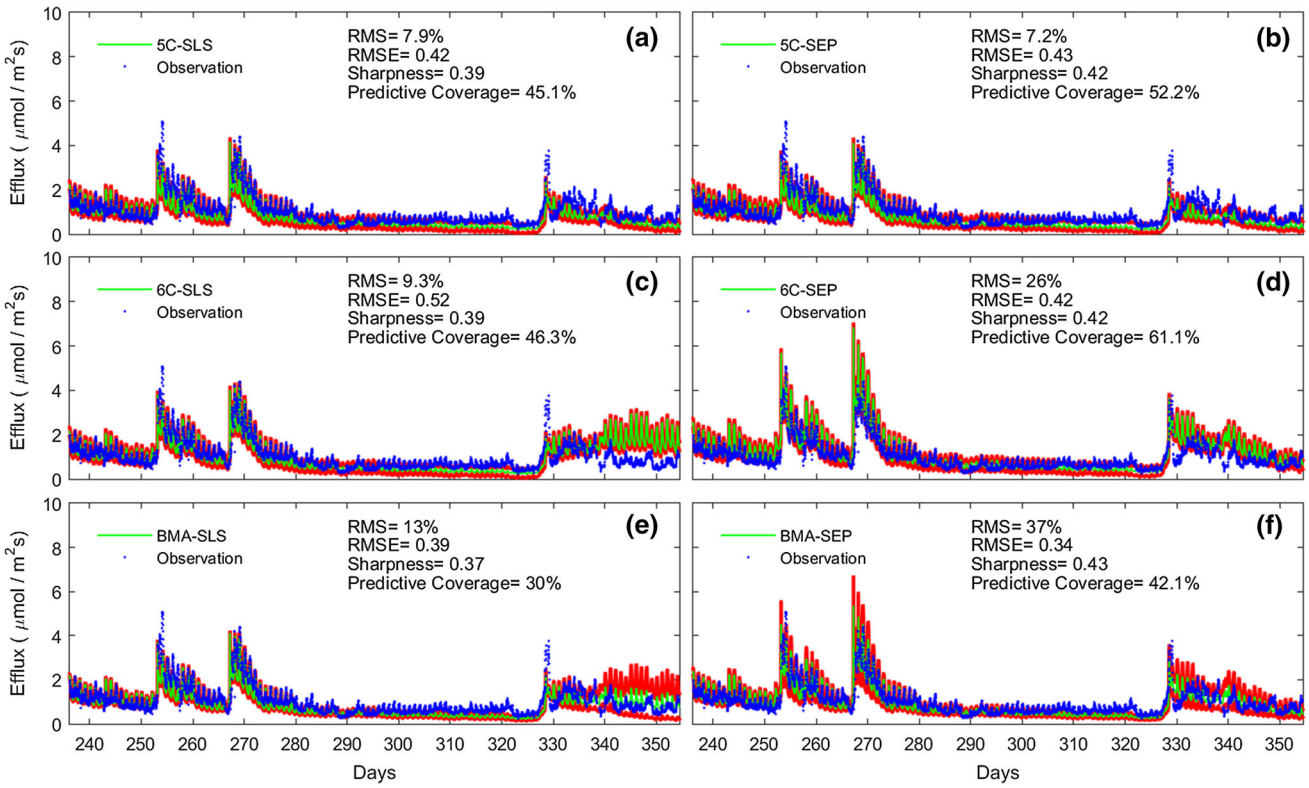

Fig. 3 Ensemble mean (green line), 95% credible interval (red line), cross-validation data (blue dots) and prediction performance metrics of **a** 5C-SLS, **b** 5C-SEP, **c** 6C-SLS and **d** 6C-SEP, **e** BMA-SLS, and **f** BMA-SEP. 5C and 6C denote two models, BMA stands for Bayesian model averaging (of 5C and 6C), and SLS and SEP are two likelihood functions

the other four ensembles fail to capture the peak carbon effluxes around 253 and 329 days. Between the two models, the prediction ensemble of BMA-SEP has higher RMS value than that of 6C-SEP, since the former has better predictions during the period after 328 days. Secondly, BMA-SLS is ranked higher than the remaining three models (5C-SLS, 5C-SEP, and 6C-SLS), because it has better predictions for the period after 328 days. Thirdly, model 6C-SLS is ranked above models 5C-SLS and 5C-SEP, which is consistent with the physical understanding that model 6C is better than model 5C, as explained in detail below. The higher RMS value of 6C-SLS may be attributed to the better predictive performance of 6C-SLS for the period of 330–340 days. Figure 3a, c indicate that predictive coverage may be the decisive statistics for determining the predictive performance of models 5C-SLS and 6C-SLS. Figure 3b, c indicate that sharpness is more decisive than the other two statistics for models 5C-SEP and 6C-SLS. Lastly, 5C-SLS and 5C-SEP have similar RMS values, and the difference in their predictions cannot be visually distinguished.

The discussion on the RMS values and corresponding model ranking leads to the following conclusions: (1) for a given likelihood function (SLS or SEP), model averaging has better predictive performance than the individual models; (2) for a given likelihood function (SLS or SEP), the high fidelity model 6C is rank better than the low fidelity model 5C, and (3) the SEP likelihood function will generally result in relatively similar (Fig. 3b) or better (Fig. 3d, f) overall prediction performance than the SLS likelihood function (Fig. 3a, c, e). These three assessments based on RMS are reasonable and justifiable. Since the first conclusion has been discussed extensively in the literature of multi-model analysis (Lu et al. 2013; Neuman 2003; Poeter and Anderson 2005; Tsai and Elshall 2013; Ye et al. 2004), the following discussion is only focused on the latter two conclusions. Model 6C has high fidelity with respect to process realism relative to model 5C. Accounting for enzymes degradation in soil during dry periods as represent by model 6C is theoretically more realistic than model 5C that excludes this process (Lawrence et al. 2009; Zhang et al. 2014; Manzoni et al. 2016; Allison and Goulden 2017). It is thus expected to find that model 6C maintain its superior predictive performance under different likelihood functions. For the SLS and SEP likelihood functions, since the PDF of SEP has a tail that is useful for robust parameter inference against outliers, SEP may potentially improves the predictive performance in comparison with SLS (Schoups and Vrugt 2010).

Table 1 Values of RMS, log-score, continuous ranked probability score, and log Bayesian model evidence and corresponding model ranking for two individual models (5C and 6C) with two likelihood functions (SLS and SEP) and Bayesian model average (BMA) of 5C and 6C

Models	5C-SLS	5C-SEP	6C-SLS	6C-SEP	BMA-SLS	BMA-SEP
Relative model score (RMS)						
Score	0.0789	0.0721	0.09319	0.2593	0.1299	0.3663
Model ranking	5	6	4	2	3	1
Log-score						
Score	Inf	Inf	Inf	Inf	Inf	Inf
Continuous ranked probability score (CRPS)						
Score	0.3147	0.3205	0.3631	0.2769	0.3184	0.4845
Model ranking	2	4	5	1	3	6
Log Bayesian model evidence (log-BME)						
Score	− 52,932	− 7473	− 33,025	− 6213	N/A	N/A
Model ranking	2 (SLS)	2 (SEP)	1 (SLS)	1 (SEP)	N/A	N/A

Table 1 lists the values of log-score, CRPS, and log RMS for the six prediction ensembles. The log-score values are infinite, because of the problem of rounding error explained in Sect. 2.2. Therefore, log-score cannot be used for model ranking in this study. While CRPS does not suffer from the rounding error, its model ranking is inaccurate. For example, CRPS ranks 5C-SLS as the second best model and BMA-SEP as the worst model, which is unreasonable based on the discussion above. For the log BME that is used for pseudo Bayes factor, it is only applicable to the individual models, but not to the model averaging. In addition, it is only applicable to the models with the same likelihood function, but not to the models with different likelihood functions. For models 5C-SLS and 6C-SLS, log BME favors 6C-SLS; for models 5C-SEP and 6C-SEP, log BME favors 6C-SEP. These rankings are consistent with those of RMS.

5 Conclusions

This study defines a new scoring rule, i.e., the relative model score (RMS), which balances the trade-off between the ensemble mean accuracy, precision, and reliability. The numerical example of soil respiration modeling of this study shows that single-criterion metrics are inadequate in assessing the overall predictive performance of different prediction ensembles obtained using individual models and model averaging based on different likelihood functions. The RMS-based ranking is reasonable on the following aspects: (1) model averaging outperforms individual models, (2) high fidelity model 6C outperforms low fidelity model 5C, and (3) the SEP likelihood function outperforms the SLS likelihood functions. These results are justified not only by physical and statistical characteristics of the models and the likelihood functions, but also by the visual inspection of the six prediction ensembles. The numerical example also compares RMS with the other three scoring rules, and the comparison reveals the following: (1) log-score cannot be used in this numerical example because it is infinite due to rounding error, (2) CRPS gives inaccurate model ranking; and (3) BME cannot be used for the prediction ensembles of model averaging and for individual models with different likelihood functions. In comparison with log-score, CRPS, and BME, RMS is applicable for a wide spectrum of multi-model analysis.

In summary, the study shows that (1) RMS is a new scoring rule that reasonably ranks candidate models accounting for the trade-off between prediction ensemble accuracy, precision, and reliability; (2) RMS can evaluate any ensemble prediction such as comparing single model with model average predictions; (3) RMS gives more accurate model ranking than the other three scoring rules, log-score, pseudo Bayes factor, and continuous ranked probability score. However, the RMS-based model ranking should not be accepted blindly without carefully examining prediction ensemble using other means. It is found in this study that the single-criterion metrics can be used to explain the ranking of RMS. The single-criterion metrics of accuracy, precision, and reliability can help understand which aspect of performance is decisive for a prediction ensemble to achieve the highest score. We thus recommend the joint use of RMS and single-criterion metrics for the general evaluation of ensemble predictive performance of multiple candidate models.

Acknowledgements This work was supported by the Department of Energy Early Career Award DE-SC0008272 and NSF-EAR Grant 1552329.

References

Ajami NK, Duan Q, Sorooshian S (2007) An integrated hydrologic Bayesian multimodel combination framework: confronting

input, parameter, and model structural uncertainty in hydrologic prediction. Water Resour Res 43:W01403. https://doi.org/10.1029/2005wr004745
Allison SD, Goulden ML (2017) Consequences of drought tolerance traits for microbial decomposition in the DEMENT model. Soil Biol Biochem 107:104–113
Anderson MP, Woessner WW (1992) Applied groundwater modeling: simulation of flow and advective transport, 2nd edn. Academic, London
Annan JD, Hargreaves JC (2010) Reliability of the CMIP3 ensemble. Geophys Res Lett. https://doi.org/10.1029/2009GL041994
Annan JD, Hargreaves JC, Tachiiri K (2011) On the observational assessment of climate model performance. Geophys Res Lett 38(24):L24702
Bulygina N, Gupta H (2011) Correcting the mathematical structure of a hydrological model via Bayesian data assimilation. Water Resour Res 47(5):W05514
Dawid AP (1984) Statistical theory: the prequential approach. J R Stat Soc Ser A 147:278–292
Diks CGH, Vrugt JA (2010) Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch Environ Res Risk Assess 24(6):809–820
Elshall AS, Tsai FTC (2014) Constructive epistemic modeling of groundwater flow with geological structure and boundary condition uncertainty under the Bayesian paradigm. J Hydrol 517:105–119
Evin G, Thyer M, Kavetski D, McInerney D, Kuczera G (2014) Comparison of joint versus postprocessor approaches for hydrological uncertainty estimation accounting for error autocorrelation and heteroscedasticity. Water Resour Res 50(3):2350–2375
Exbrayat JF, Viney NR, Frede HG, Breuer L (2013) Using multimodel averaging to improve the reliability of catchment scale nitrogen predictions. Geoscientific Model Development 6(1):117–125
Foglia L, Mehl SW, Hill MC, Burlando P (2013) Evaluating model structure adequacy: the case of the Maggia Valley groundwater system, southern Switzerland. Water Resour Res. https://doi.org/10.1029/2011wr011779
Gneiting T, Raftery AE (2007) Strictly proper scoring rules, prediction and estimation. J Am Stat Assoc 102(447):359–378
Good IJ (1952) Decisions. J R Stat Soc Ser B 14(1):107–114
Gulden LE, Rosero E, Yang ZL, Wagener T, Niu GY (2008) Model performance, model robustness, and model fitness scores: a new method for identifying good land-surface. Geophys Res Lett 35(11):L11404
Hargreaves JC, Annan JD, Yoshimori M, Abe-Ouchi A (2012) Can the Last Glacial Maximum constrain climate sensitivity? Geophys Res Lett 39(24):L24702
Heath MT (1997) Scientific computing: an introductory survey. McGraw-Hill, Boston
Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. Weather Forecast 15:559–570
Hill MC, Tiedeman CR (2007) Effective calibration of ground water models, with analysis of data, sensitivities, predictions, and uncertainty. Wiley, New York, p 480
Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. Stat Sci 14(4):382–401
Kumar P (2011) Typology of hydrologic predictability. Water Resourc Res 47(3):W00H05
Laloy E, Vrugt JA (2012) High-dimensional posterior exploration of hydrologic models using multiple-try DREAM(ZS) and high-performance computing. Water Resour Res 48(1):W01526. https://doi.org/10.1029/2011WR010608
Lartillot N, Philippe H (2006) Computing Bayes factors using thermodynamic integration. Syst Biol 55(2):195–207
Lawrence CR, Neff JC, Schimel JP (2009) Does adding microbial mechanisms of decomposition improve soil organic matter models? A comparison of four models using data from a pulsed rewetting experiment. Soil Biol Biochem 41(9):1923–1934
Liu PG, Elshall AS, Ye M, Beerli P, Zeng XK, Lu D, Tao YZ (2016) Evaluating marginal likelihood with thermodynamic integration method and comparison with several other numerical methods. Water Resour Res 52(2):734–758
Lu D, Ye M, Meyer PD, Curtis GP, Shi X, Niu X-F, Yabusaki SB (2013) Effects of error covariance structure on estimation of model averaging weights and predictive performance. Water Resour Res. https://doi.org/10.1002/wrcr.20441
Lu D, Ye M, Curtis GP (2015) Maximum likelihood Bayesian model averaging and its predictive analysis for groundwater reactive transport models. J Hydrol 529:1859–1873
Manzoni S, Moyano F, Kätterer T, Schimel J (2016) Modeling coupled enzymatic and solute transport controls on decomposition in drying soils. Soil Biol Biochem 95:275–287
Matheson JE, Winkler RL (1976) Scoring rules for continuous probability distributions. Manag Sci 22(10):1087–1096
Nash JE, Sutcliffe JV (1970) River flow forecasting through conceptual models part I—a discussion of principles. J Hydrol 10(3):282–290
Neuman SP (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. Stoch Environ Res Risk Assess 17(5):291–305
Nowak W, Rubin Y, de Barros FPJ (2012) A hypothesis-driven approach to optimize field campaigns. Water Resour Res 48(6):W06509
Oldenborgh GJ, Reyes FJD, Drijfhout SS, Hawkins E (2013) Reliability of regional climate model trends. Environ Res Lett 8(1):014055
Poeter EP, Anderson DA (2005) Multimodel ranking and inference in ground water modeling. Ground Water 43(4):597–605. https://doi.org/10.1111/j.1745-6584.2005.0061.x
Poeter EP, Hill MC, Banta ER, Mehl SW, Christensen S (2005) UCODE_2005 and six other computer codes for universal sensitivity analysis, inverse modeling, and uncertainty evaluation. U.S. Geological Survey Techniques and Methods, 6-A11
Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Mon Weather Rev 133(5):1155–1174
Renard B, Kavetski D, Kuczera G, Thyer M, Franks SW (2010) Understanding predictive uncertainty in hydrologic modeling: the challenge of identifying input and structural errors. Water Resour Res 46(5):W05521
Ricciuto DM, King AW, Dragoni D, Post WM (2011) Parameter and prediction uncertainty in an optimized terrestrial carbon cycle model: effects of constraining variables and data record length. J Geophys Res Biogeosci 116(G1):G01033
Sadegh M, Vrugt JA (2013) Bridging the gap between GLUE and formal statistical approaches: approximate Bayesian computation. Hydrol Earth Syst Sci 17(12):4831–4850
Schöniger A, Wöhling T, Samaniego L, Nowak W (2014) Model selection on solid ground: rigorous comparison of nine ways to evaluate Bayesian model evidence. Water Resour Res 50(12):9484–9513
Schoups G, Vrugt JA (2010) A formal likelihood function for parameter and predictive inference of hydrologic models with correlated, heteroscedastic, and non-Gaussian errors. Water Resour Res 46(10):W10531
Shi X, Ye M, Finsterle S, Wu J (2012) Comparing nonlinear regression and Markov chain Monte Carlo methods for assessment of predictive uncertainty in vadose zone modeling. Vadose Zone J 11(4):83–97

Shrestha DL (2014) Continuous rank probability score, MathWorks File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/47807-continuous-rank-probability-score/content/crps.m. Last checked 8 Feb 2017

Silverman BW (1998) Density estimation for statistics and data analysis. Chapman & Hall, Boca Raton, p 176

Smith RC (2014) Uncertainty quantification: theory, implementation, and applications. Computational science and engineering series, vol XVIII. Society for Industrial and Applied Mathematics, Philadelphia, p 382 s

Smith MW, Bracken LJ, Cox NJ (2010) Toward a dynamic representation of hydrological connectivity at the hillslope scale in semiarid areas. Water Resour Res 46(12):W12540

Thyer M, Renard B, Kavetski D, Kuczera G, Franks SW, Srikanthan S (2009) Critical evaluation of parameter consistency and predictive uncertainty in hydrological modeling: a case study using Bayesian total error analysis. Water Resourc Res 45(12):W00B14

Tsai FTC, Elshall AS (2013) Hierarchical Bayesian model averaging for hydrostratigraphic modeling: uncertainty segregation and comparative evaluation. Water Resour Res 49(9):5520–5536

Vrugt JA, ter Braak CJF, Gupta HV, Robinson BA (2009) Equifinality of formal (DREAM) and informal (GLUE) Bayesian approaches in hydrologic modeling? Stoch Environ Res Risk Assess 23(7):1011–1026

Wenger SJ, Som NA, Dauwalter DC, Isaak DJ, Neville HM, Luce CH, Dunham JB, Young MK, Fausch KD, Rieman BE (2013) Probabilistic accounting of uncertainty in forecasts of species distributions under climate change. Glob Change Biol 19(11):3343–3354

Winter CL (2010) Normalized Mahalanobis distance for comparing process-based stochastic models. Stoch Environ Res Risk Assess 24(6):917–923

Winter CL, Nychka D (2010) Forecasting skill of model averages. Stoch Environ Res Risk Assess 24(5):633–638

Wöhling T, Schöniger A, Gayler S, Nowak W (2015) Bayesian model averaging to explore the worth of data for soil-plant model selection and prediction. Water Resour Res 51(4):2825–2846

Xue L, Zhang D (2014) A multi-model data assimilation framework via the ensemble Kalman filter. Water Resour Res 50(5):4197–4219

Ye M, Neuman SP, Meyer PD (2004) Maximum likelihood Bayesian averaging of spatial variability models in unsaturated fractured tuff. Water Resour Res 40(5):W05113

Ye M, Meyer PD, Lin Y-F, Neuman SP (2010) Quantification of model uncertainty in environmental modeling. Environ Res Risk Assess, Stoch. https://doi.org/10.1007/s00477-010-0377-0

Yokohata T, Annan JD, Collins M, Jackson CS, Tobis M, Webb MJ, Hargreaves JC (2012) Reliability of multi-model and structurally different single-model ensembles. Clim Dyn 39(3–4):599–616

Zeng X, Ye M, Wu J, Wang D, Zhu X (2018) Improved nested sampling and surrogate-enabled comparison with other marginal likelihood estimators. Water Resour Res 54:797–826. https://doi.org/10.1002/2017WR020782

Zhang X, Niu G-Y, Elshall AS, Ye M, Barron-Gafford GA, Pavao-Zuckerman M (2014) Assessing five evolving microbial enzyme models against field measurements from a semiarid savannah-What are the mechanisms of soil respiration pulses? Geophys Res Lett 41(18):6428–6434