

# Comparison of point forecast accuracy of model averaging methods in hydrologic applications

Cees G. H. Diks · Jasper A. Vrugt

Published online: 30 March 2010

© The Author(s) 2010. This article is published with open access at Springerlink.com

**Abstract** Multi-model averaging is currently receiving a surge of attention in the atmospheric, hydrologic, and statistical literature to explicitly handle conceptual model uncertainty in the analysis of environmental systems and derive predictive distributions of model output. Such density forecasts are necessary to help analyze which parts of the model are well resolved, and which parts are subject to considerable uncertainty. Yet, accurate point predictors are still desired in many practical applications. In this paper, we compare a suite of different model averaging techniques by their ability to improve forecast accuracy of environmental systems. We compare equal weights averaging (EWA), Bates-Granger model averaging (BGA), averaging using Akaike's information criterion (AICA), and Bayes' Information Criterion (BICA), Bayesian model averaging (BMA), Mallows model averaging (MMA), and

Granger-Ramanathan averaging (GRA) for two different hydrologic systems involving water flow through a 1950 km<sup>2</sup> watershed and 5 m deep vadose zone. Averaging methods with weights restricted to the multi-dimensional simplex (positive weights summing up to one) are shown to have considerably larger forecast errors than approaches with unconstrained weights. Whereas various sophisticated model averaging approaches have recently emerged in the literature, our results convincingly demonstrate the advantages of GRA for hydrologic applications. This method achieves similar performance as MMA and BMA, but is much simpler to implement and use, and computationally much less demanding.

**Keywords** Bates-Granger weights · Bayesian model averaging · Granger-Ramanathan weights · Mallows model averaging · Streamflow forecasting · Tensiometric pressure head

---

C. G. H. Diks (✉)  
Center for Nonlinear Dynamics in Economics and Finance  
(CenDEF), Faculty of Economics and Business, University of  
Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, NL, The  
Netherlands  
e-mail: C.G.H.Diks@uva.nl

J. A. Vrugt  
Center for Nonlinear Studies (CNLS), Los Alamos National  
Laboratory, Mail Stop B258, Los Alamos, NM 87545, USA  
e-mail: jasper@uci.edu

J. A. Vrugt  
Institute for Biodiversity and Ecosystem Dynamics (IBED),  
Nieuwe Achtergracht 166, 1018 WV Amsterdam, The  
Netherlands

J. A. Vrugt  
Department of Civil and Environmental Engineering, The Henry  
Samueli School of Engineering, University of California, Irvine,  
CA 92697, USA

## 1 Introduction

The motivating idea behind model averaging is that, with various competing models at hand, each having its own strengths and weaknesses, it should be possible to combine the individual model forecasts into a single new forecast that, up to one's favorite standard, is at least as good as any of the individual forecasts. As usual in statistical model building, the aim is to use the available information efficiently, and to construct a predictive model with the right balance between model flexibility and over-fitting. Viewed as such, model averaging is a natural generalization of the more traditional aim of model selection. Indeed, the model averaging literature has its roots in the model selection literature, which continues to be a very active research area

with many applications in hydrology (see e.g., Wagener and Gupta 2005 and Ye et al. 2008). As a result of the steady increase in computer power, model averaging has gradually gained popularity as an alternative to model selection. For examples of recent applications of model averaging in hydrology, see, for instance, the contributions by Vrugt and Robinson (2007); Rojas et al. (2008) and Wöhling and Vrugt (2008).

Some model averaging techniques, such as equal weights averaging focus on point predictors, while others, such as Bayesian model averaging, are concerned with density forecasts. While the recent developments on density forecasts will undoubtedly prove extremely useful, and will find many practical applications, there will always be cases where accurate point predictors are desired. Since any density forecast has an associated point predictor, being the predictive mean of the density forecast, the question arises naturally which of the point predictors of the available averaging methods is most accurate. Because the recently developed sophisticated density forecast methods aim at obtaining accurate predictive densities, their point forecasts do not necessarily have to perform better than more traditional point forecast methods. For instance, Stockdale (2000) notes that simple averaging seems to have been the more successful approach, but also argues that more research is required.

With this in mind, the aim of this paper is to investigate the accuracy of a wide range of point predictors empirically, using hydrologic data from different case studies. Unknown model parameters are estimated using data from the calibration period, while the predictive performance is assessed using data from the subsequent evaluation period. The point predictors are obtained from a variety of model averaging techniques, and evaluated in terms of out-of-sample root mean squared prediction error (RMSE). Although other measures of predictive accuracy exist, the RMSE is natural in this context, since it is one of the objective functions that is being minimized (via maximization of the likelihood) during the calibration period for most of the density forecast methods in the literature.

Two hydrological case studies are considered. The first case study considers a classical forecasting problem in surface water hydrology, and involves rainfall-runoff modeling of the Leaf River watershed in Mississippi, USA, using a 36-year historical record of daily streamflow data. A set of 8 commonly used conceptual hydrologic models is used to predict and study streamflow dynamics. The second case study involves soil hydrology, and focuses on prediction of soil water flow through the vadose zone using data from a layered vadose zone of volcanic origin in New Zealand. A 7-member ensemble of soil hydraulic models is used to predict tensiometric pressure heads at multiple depths. The data sets, models, and ensembles used in the

two case studies have been described elsewhere (Vrugt and Robinson 2007; Wöhling and Vrugt 2008), and details can be found there.

The point forecasts considered are based on the following model averaging techniques: equal weights averaging (EWA) where each of the available models is weighted equally, Bates-Granger averaging (BGA) (Bates and Granger 1969), AIC and BIC-based model averaging (AICA and BICA, respectively) (Buckland et al. 1997; Burnham and Anderson 2002; Hansen 2008), Bayesian model averaging (BMA) (Raftery et al. 1997; Hoeting et al. 1999; Raftery et al. 2005), Mallows model averaging (MMA) (Hansen 2007; Hansen 2008) and weights equal to the ordinary least squares (OLS) estimates of the coefficients of a multiple linear regression model, as first suggested in the forecasting context by Granger and Ramanathan (1984), and referred to here as Granger-Ramanathan averaging (GRA). Note that some of these model averaging techniques only allow positive weights, summing to one (weights on the simplex). For comparison, whenever feasible, models for which the weights are usually restricted to the simplex were also estimated without this restriction.

The paper is organized as follows. Section 2 introduces the concept of model averaging, considers bias removal and provides a concise description of the various model averaging strategies considered here. Some statistical properties of these methods, in particular convergence to the optimal predictor, are considered briefly in Sect. 3. In Sect. 4 we describe the two hydrologic case studies, as well as the data and models used, and summarize the results for the different averaging procedures. The RMSE and optimized weights of the individual models of the ensemble are the primary focus of that section. Section 5 provides a discussion of our results against the background of the statistical properties of the model averaging methods.

## 2 Model averaging

Let us denote by  $\{Y_t\}$  a sequence of measurements of a hydrological quantity of interest, such as a sequence of daily discharge data. Further assume that there is an ensemble of  $k$  competing forecasts available, which constitutes the predictions of  $k$  competing models, with associated point forecasts  $X_{i,t}$ ,  $i = 1, \dots, k$ , for each variable  $Y_t$ , in terms of the information available (i.e., all variables available to construct a forecast of  $Y_t$ ). We focus on predicting  $Y_t$  (e.g., streamflow (case study 1) or soil water pressure head (case study 2) at time  $t$ ) based on the information available at the start of that period.

For hydrologic applications, the models that make up the ensemble could range from simple water balance models to parsimonious conceptual watershed models, and

increasingly complex fully integrated three-dimensional physically based models. Prior to our analysis, the various hydrologic models of the ensemble were calibrated through global optimization using a least squares objective function containing the error residuals between measured and predicted daily streamflow (case study 1), and tensiometric pressure head (case study 2). These variables are also subject of interest in all our averaging studies herein. The predictions of the various ensemble members can therefore be considered optimal in the least squares sense, and no further reductions in RMSE are possible by parameter tuning.

A bias correction step of the individual forecasts is performed prior to the construction of the weights, as follows. Each of the individual forecasts is adjusted by applying a linear transformation of the form  $X'_{i,t} = a_i + b_i X_{i,t}$ . The coefficients  $a_i$  and  $b_i$ ,  $i = 1, \dots, k$ , for each of the models are found by OLS estimation using only the observations in the calibration set. Typically this bias correction leads to a small improvement of the predictive performance of the individual models, with  $a_i$  close to zero and  $b_i$  close to 1. Only when the calibration set is very small, the OLS estimates underlying GRA becomes noisy, and bias correction may destabilize the ensemble. This has been demonstrated in the context of streamflow forecasting in Vrugt and Robinson (2007). In the case studies we used a bias correction, although for comparison we also provide RMSE values obtained without bias correction. Subsequently, for notational simplicity, the symbol  $X_{i,t}$  is used to indicate either an uncorrected or a bias-corrected predictor of  $Y_t$ , based on model  $i$ .

A popular way to combine point forecasts is to consider the following linear model combining the individual predictions:

$$Y_t = \mathbf{X}_t^T \boldsymbol{\beta} + \varepsilon_t = \sum_{i=1}^k \beta_i X_{i,t} + \varepsilon_t, \quad (1)$$

where  $\{\varepsilon_t\}$  is a white noise sequence, which will be assumed to have a normal distribution with zero mean and unknown variance. In hydrological case studies one typically uses a split sample test to divide the available data in a calibration and evaluation time series. The parameter vector  $\boldsymbol{\beta}$  is estimated from the calibration set, and the evaluation data set is used to determine the consistency and robustness of the calibrated model. We will follow the same approach in this paper, and use the RMSE during the evaluation period as a measure of out-of-sample predictive performance.

As indicated above, this paper focuses on comparing point predictors in terms of the out-of-sample prediction error. The point forecasts associated with model (1) are

$$\tilde{Y}_t^{\boldsymbol{\beta}} = \mathbf{X}_t^T \boldsymbol{\beta} = \sum_{i=1}^k \beta_i X_{i,t}. \quad (2)$$

We will compare a wide range of competing model averaging techniques proposed in the literature so far. These include simple methods such as equal weights averages and BGA weights, as well as more complex methods such as AICA, BICA, GRA, BMA, and MMA. We next describe each of these methods in some detail, and give expressions for the value of  $\boldsymbol{\beta}$  used by the methods. These are denoted by  $\hat{\boldsymbol{\beta}}_{\bullet}$ , where the subscript indicates the averaging method used. Table 1 summarizes the main properties of each of the model averaging techniques considered.

## 2.1 Equal weights averaging

Under equal weights averaging, the combined forecast is simply obtained by giving each of the models equal weight, i.e.,  $\hat{\boldsymbol{\beta}}_{\text{EWA}} = (\frac{1}{k}, \dots, \frac{1}{k})$ , which has associated predictor  $\tilde{Y}_t^{\hat{\boldsymbol{\beta}}_{\text{EWA}}} = \frac{1}{k} \sum_{i=1}^k X_{i,t}$ . Note that if no bias correction is used, this predictor does not even depend on the data in the calibration set (the bias correction does depend on the data in the calibration set).

## 2.2 Bates-Granger averaging

A well-known choice, proposed by Bates and Granger (1969), is to weight each model by  $1/\sigma_i^2$ , where  $\sigma_i^2$  is its forecast variance. If the models' forecasts are unbiased and their errors uncorrelated, these weights are optimal in the sense that they produce predictors with the smallest possible RMSE. In practice the forecast variance is unknown and needs to be estimated. This leads to the choice

$$\hat{\beta}_{\text{BGA},i} = \frac{1/\hat{\sigma}_i^2}{\sum_{j=1}^k 1/\hat{\sigma}_j^2},$$

where  $\hat{\sigma}_i^2$  denotes the forecast variance of model  $i$ , which we estimated as the sample variance of the forecast error  $e_{i,t} = X_{i,t} - Y_t$  within the calibration period.

## 2.3 Information criterion averaging

Buckland et al. (1997) and Burnham and Anderson (2002) proposed using weights of the form

$$\hat{\beta}_i = \frac{\exp(-I_i/2)}{\sum_{j=1}^k \exp(-I_j/2)}, \quad (3)$$

where  $I_i$  is an information criterion describing the fit of the model, of the form  $I_i = -2 \log(L_i) + q(p_i)$ , where  $L_i$  is the (maximized) likelihood of model  $i$ , and  $q(p_i)$  is a penalty increasing in the number of parameters,  $p_i$ , that need to be

**Table 1** Main characteristics of the model averaging methods considered

Method	Model type	Associated density forecast	$\beta$ restricted to $\Delta$	Estimation procedure
BMA <sub>mix</sub>	Finite mixture	Yes	Yes	DREAM
BMA <sub>lin</sub> <sup>Δ</sup>	Linear regression	Yes	Yes	DREAM
BMA <sub>lin</sub>	Linear regression	Yes	No	DREAM
MMA <sup>Δ</sup>	Linear regression	No	Yes	DREAM
MMA	Linear regression	Yes	No	DREAM
AICA	Combined point forecast	No	Yes	Analytic
BICA	Combined point forecast	No	Yes	Analytic
GRA	Combined point forecast	No	No	Analytic
BGA	Combined point forecast	No	Yes	Analytic
EWA	Combined point forecast	No	Yes	Data independent

estimated for model  $i$ . The cases considered here are Akaike's information criterion (AIC), with penalty  $q(p) = 2p$ , and the Bayes information criterion (BIC), with penalty  $q(p) = p \log(n)$ , where  $n$  is the calibration sample size. We refer to the model averaging scheme (3) based on AIC and BIC as AICA and BICA, respectively, and to their respective  $\beta$ -values as  $\hat{\beta}_{\text{AICA}}$  and  $\hat{\beta}_{\text{BICA}}$ . In the literature these methods are sometimes referred to as smooth AIC and smooth BIC, respectively. To evaluate the information criteria numerically, it is convenient to assume, as we do here, that the errors of the individual models are normally distributed. In that case the likelihood for model  $i$  is related to its estimated forecast error  $\hat{\sigma}_i^2$ , via  $-2 \log(L_i) = n \log \hat{\sigma}_i^2 + n$ .

## 2.4 Granger-Ramanathan averaging

The weighting schemes described above do not exploit covariance structure that may be present in the forecast errors. The predictors described above are weighted averages of the individual forecasts, with weights determined by a measure of fit for the individual models. A natural way to exploit the presence of covariances is by using OLS estimators within the linear regression model.

This OLS approach to forecast combination was suggested by Granger and Ramanathan (1984). They suggested using OLS to estimate the unknown parameters in the linear regression model. The OLS estimator of the parameter vector  $\beta$  of the linear regression model (1) is

$$\hat{\beta}_{\text{GRA}} = (X^T X)^{-1} X^T Y, \quad (4)$$

where  $X$  and  $Y$  stand for the matrix of  $X$ -values, and the vector of  $Y$ -values in the calibration set, respectively. Under the standard assumptions underlying the classical linear regression model the OLS estimator can be shown to be the best linear unbiased estimator of  $\beta$ . Even if some of these assumptions do not hold (e.g., if there are missing  $X$ -variables) the OLS estimates may still be shown to

converge to the pseudo-true parameter  $\beta^*$ , corresponding to the best linear model of  $Y_t$  (in the RMSE sense) in terms of  $X_{1,t}, \dots, X_{k,t}$ . Since GRA is based on these OLS weights, it should be expected to be a serious competitor for the other model averaging techniques considered here.

## 2.5 Bayesian model averaging

Hoeting et al. (1999) provide an excellent overview of the various BMA techniques that have been proposed in the literature. Applications of BMA in hydrology and meteorology have been described by Neuman (2003); Ye et al. (2004); Raftery et al. (2005); Gneiting et al. (2005); Vrugt and Robinson (2007); and Vrugt et al. (2008). See Bishop and Shanley (2008) for a recent contribution towards improving BMA's performance when confronted with data coming from extreme weather conditions.

Depending on the type of application one has in mind, different flavors of BMA are more suitable. For instance, it makes a crucial difference whether one would like to combine point forecasts (in which case some forecasts may be assigned negative weights) or density forecasts (in which case allowing for negative weights could lead to the undesired effect of negative density forecasts). The different BMA methods considered in this paper, being BMA in the finite mixture model and BMA in the linear regression model, correspond to averaging density forecasts and making linear combinations of forecasts, respectively.

### 2.5.1 BMA in the finite mixture model

In case one wishes to combine density forecasts  $f_{i,t}(y)$ ,  $i = 1, \dots, k$ , for  $Y_t$ , it is common to consider the combined forecast density  $g_t(y) = \sum_{i=1}^k \beta_i f_{i,t}(y)$ , known as a finite mixture model. To ensure that  $g_t(y)$  represents a density, the BMA weights  $\beta_i$  are assumed to be non-negative and to add up to one: i.e., the weights are constrained to the simplex  $\Delta^{k-1} = \{\beta | \beta_i \geq 0, i = 1, \dots, k \text{ and } \sum_{i=1}^k \beta_i = 1\}$ .

The BMA predictive density of  $Y_t$  thus consists of a mixture of normals, located at the individual point forecasts  $X_{i,t}$ , with weight  $\beta_i$  and variance  $\sigma_i^2$ . The point predictors associated with this forecast are again given by (2), where  $X_{i,t}$  coincides with the predictive mean  $\int_{-\infty}^{\infty} f_{i,t}(x)dx$ .

For this finite mixture model, Raftery et al. (2005) define the BMA weights to be those weights which maximize the likelihood based on the data available for estimation (the calibration data). Here, as is often done in the literature, the mixture densities are assumed to be centered around  $X_{i,t}$  with a normally distributed noise with a fixed unknown variance  $\sigma_i^2$ ,  $i = 1, \dots, k$ , i.e.,  $f_{i,t}(Y_t) = (2\pi\sigma_i^2)^{-\frac{1}{2}} \exp(-(Y_t - X_{i,t})^2/(2\sigma_i^2))$ .

We use the short-hand notation  $\text{BMA}_{\text{mix}}$  to indicate the BMA method in the context of the finite mixture model. The joint estimators of  $\beta$  and  $\sigma = (\sigma_1, \dots, \sigma_k)'$  are given by

$$\begin{aligned} (\hat{\beta}_{\text{BMA}_{\text{mix}}}, \hat{\sigma}) &= \arg \max_{\beta \in \Delta^{k-1}, \sigma \in \mathbb{R}_+^k} \sum_{t=1}^n \log g_t(Y_t) \\ &= \arg \max_{\beta \in \Delta^{k-1}, \sigma \in \mathbb{R}_+^k} \sum_{t=1}^n \log \left( \sum_{i=1}^k f_{i,t}(Y_t) \right), \end{aligned}$$

where  $n$  denotes the calibration sample size.

The estimation of the model weights and model variances  $\sigma_i^2$  is a moderately complex nonlinear optimization problem. One way to approach this is by using the expectation-maximization algorithm, as suggested by Raftery et al. (2005). In fact, any other reliable optimization routine can be used for the purpose of the present paper. We follow Vrugt et al. (2008) and use the Differential Evolution Adaptive Metropolis (DREAM) adaptive Markov chain Monte Carlo (MCMC) algorithm, proposed in Vrugt et al. (2009), for obtaining the model weights and the individual model variances. For a given calibration data set, the likelihood function was optimized numerically by sampling from the Bayesian posterior distribution with a flat prior (i.e., from a density in parameter space which is a normalized version of the likelihood function), and identifying the point in the MCMC sample for which the likelihood of the finite mixture model was maximized. We used the standard algorithmic settings of DREAM, sampling a total of  $2 \times 10^5$  values of  $\beta$  for each case, since we found that the solutions obtained with this number of iterations were sufficiently accurate as judged from comparing independent runs of the algorithm.

### 2.5.2 BMA in the linear regression model

Bayesian model averaging in the context of the linear regression model (1) has been discussed by Raftery et al. (1997). This particular BMA approach does allow for the model weights to become negative. The associated point

predictor in the case of a flat prior reduces to the point predictor of the maximum likelihood estimator, which happens to be the OLS estimator (4). Although this theoretical argument shows that formally the BMA version of the linear model is redundant in our set of model averaging approaches, we have included it nevertheless, to provide an extra check on our numerical procedures. We refer to the BMA method in the context of the linear regression model by using the subscript ‘lin’ (for ‘linear’), and an additional superscript ‘ $\Delta$ ’ in case the weights are restricted to the simplex  $\Delta^{k-1}$ . For instance,  $\text{BMA}_{\text{lin}}^{\Delta}$  refers to BMA in the linear regression model (1), with  $\beta$  on the simplex.

The BMA weights in the linear regression model are obtained by maximizing the Bayesian posterior density for a flat prior, that is

$$(\hat{\beta}_{\text{BMA}_{\text{lin}}^{\Delta}}, \hat{\sigma}) = \arg \max_{\beta \in \mathcal{H}, \sigma \in \mathbb{R}_+} \sum_{t=1}^n \log h_t(Y_t),$$

where  $h_t(Y_t) = (2\pi\sigma^2)^{-\frac{k}{2}} \exp(-(Y_t - \sum_{i=1}^k \beta_i X_{i,t})^2/(2\sigma^2))$ , and  $\mathcal{H}$  is the set of  $\beta$ -values under consideration, which is either  $\mathbb{R}^k$  (empty superscript  $s_{\mathcal{H}}$ ) or  $\Delta^{k-1}$  ( $s_{\mathcal{H}} = \Delta$ ) in this paper. The maximization was again performed using the DREAM algorithm.

### 2.6 Mallows model averaging

On the other side of the spectrum considerable effort has been made to come up with a frequentist solution to the problem of model averaging. Hjort and Claeskens (2003); Claeskens and Hjort (2003) made noticeable contributions in this direction. The general tendency in this literature is that there is no unique best model, but that the best model is a subjective notion that depends on one’s objective. For each objective, a different model may be optimal. Hansen (2007, 2008) has continued this line of research with his proof that model combination based on Mallows criterion, asymptotically leads to forecasts with the smallest possible mean squared error. For a more generally valid proof, which covers the context of the linear regression model (1) considered here, see Wan et al. (2010).

The MMA criterion is the penalized sum of squared residuals (across the calibration data)

$$C_n(\beta) = \sum_{t=1}^n (Y_t - \beta'X_t)^2 + 2 \sum_{j=1}^k \beta_j p_j S^2 \quad (5)$$

where, as before,  $p_j$  is the number of parameters of model  $j$  (i.e., a measure of the complexity of model  $j$ ), and  $S^2$  is an estimate of the variance  $\sigma^2$  of  $\varepsilon_t$  in (1) based on the most complex model considered. In this study  $S^2$  was taken to be the smallest observed RMSE for any individual model, among the set of models.



The Mallows criterion is

$$\hat{\beta}_{\text{MMA}^{\mathcal{H}}} = \arg \min_{\beta \in \mathcal{H}} C_n(\beta),$$

where  $\mathcal{H}$  is the set of parameters under consideration ( $\mathbb{R}^k$  or  $\Delta^{k-1}$ ). The value of  $\hat{\beta}_{\text{MMA}}$ , i.e., the value of  $\beta$  for which (5) is minimized, is found again using the DREAM algorithm. With the DREAM algorithm we generate a sample of size  $2 \times 10^5$ , as in the case of the BMA algorithm, from the probability density function  $h(\beta)$  in  $\beta$ -space, where

$$h(\beta) \propto e^{-\frac{1}{2}C_n(\beta)}.$$

Like in the BMA implementation, the maximum of this function is found by identifying the point  $\beta_i$  in the MCMC sample for which  $h(\beta_i)$  was largest.

### 3 Consistency and asymptotic RMSE optimality

This section briefly summarizes some of the statistical properties of the model averaging methods described above. The proofs are omitted as they are standard results in statistics. The aim of this section is to anticipate the types of results to expect in the case studies presented in the next section, and to facilitate their discussion afterwards.

In Sect. 3.1 we consider asymptotics with a fixed number,  $k$ , of ensemble members, and an increasing calibration size  $n$ . Asymptotics where the number of competing models increases with the calibration sample size are discussed in Sect. 3.2.

#### 3.1 Asymptotics with $k$ finite and fixed

We assume the vector-valued time series  $\{Z_t \equiv (Y_t, X_t')'\}$ ,  $t \in \mathbb{Z}$  to be strictly stationary, with finite variances. The best linear predictor of  $Y_t$  is then defined as

$$\beta^* = \arg \min_{\beta \in \mathbb{R}^k} \mathbb{E}[(Y_t - X_t'\beta)^2],$$

where  $\mathbb{E}$  denotes the mathematical expectation operator.

An estimator  $\hat{\beta}_n$  of  $\beta^*$  is consistent if it converges to  $\beta^*$  in probability as the calibration sample size  $n$  increases:  $\mathbb{P}(\|\hat{\beta}_n - \beta^*\| < \varepsilon) \xrightarrow{n \rightarrow \infty} 1$ , for any  $\varepsilon > 0$ , where  $\|\cdot\|$  denotes the Euclidean norm (or an equivalent norm) in  $\mathbb{R}^k$ . A model averaging method based on a consistent estimator of  $\beta^*$  is asymptotically RMSE optimal in that the expected out-of-sample RMSE converges to the lowest possible value,  $\sigma^2$ ; the variance of the part of  $Y_t$  that cannot be explained by any linear combination of the  $X$ -es.

Three of the model averaging methods described above are asymptotically RMSE optimal: GRA,  $\text{BMA}_{\text{lin}}$ , and MMA. GRA is exploiting a consistent (OLS) estimator and

hence is asymptotically RMSE optimal. Because the OLS estimators coincide with maximum likelihood estimators in linear models with normal errors,  $\text{BMA}_{\text{lin}}$  with a flat prior is formally equivalent to GRA. This means that apart from numerical inaccuracy associated with the optimization routine used,  $\text{BMA}_{\text{lin}}$  as we implemented it, with a flat prior, is equivalent to GRA. The consistency of Bayesian point estimators guarantees that  $\text{BMA}_{\text{lin}}$  with a prior distribution with  $\beta^*$  in its support is asymptotically RMSE optimal. Finally, the objective function of MMA and GRA differ only by a model dependent finite penalty, independent of sample size. Since the penalty plays no role asymptotically, MMA is asymptotically equivalent to GRA, and therefore also asymptotically RMSE optimal.

Equal weights averaging is only asymptotically RMSE optimal in the very special case that the optimal weight vector happens to be  $(\frac{1}{k}, \dots, \frac{1}{k})$ . Likewise, methods that are confined to the simplex  $\Delta^{k-1}$  are by definition not asymptotically RMSE optimal if  $\beta^*$  lies outside the simplex.  $\text{BMA}_{\text{lin}}^\Delta$  and  $\text{MMA}^\Delta$  inherit the asymptotic RMSE optimality from  $\text{BMA}_{\text{lin}}$  and MMA, respectively, if  $\beta^* \in \Delta^{k-1}$ .

$\text{BMA}_{\text{mix}}$ , AICA, and BICA are not asymptotically RMSE optimal, even if  $\beta^* \in \Delta^{k-1}$ . Because the likelihood functions of the finite mixture model and in the linear regression model are different, the weights underlying  $\text{BMA}_{\text{mix}}$  converge in probability to a value which is generally different from  $\beta^*$ . AICA and BICA are not asymptotically RMSE optimal, because the penalties  $q(p_i)$  are fixed while  $\log(L)$  grows, on average, linearly with sample size. Assuming that no two ensemble members have exactly the same average likelihood, the weight vector  $w$  converges (in probability) to the single model in the ensemble that exhibits the highest average log likelihood per sample point (the best performing individual model).

#### 3.2 Asymptotics with $k$ depending on $n$

The result that GRA and  $\text{BMA}_{\text{lin}}$  were found to be asymptotically RMSE optimal in Sect. 3.1, depends crucially on the assumption that  $k$  was fixed as the sample size increases. Hansen (2007, 2008) introduced MMA in a context where the number of competing models,  $k = k(n)$ , is allowed to grow with the sample size, and showed that MMA converges to the optimal model asymptotically. Under these asymptotics, GRA and  $\text{BMA}_{\text{lin}}$  would typically fail to converge to the optimal model, due to the lack of a penalty for model complexity.

For the case studies one can envisage three possible scenarios a priori: either the fixed  $k$  asymptotic theory applies, or Hansen's asymptotics, or neither. If the number of model parameters  $k$  is sufficiently small, and the calibration sample size large, one should expect fixed  $k$  asymptotics to apply. If  $k$  is too large compared to the

sample size Hansen's asymptotics apply. If the calibration sample set is too small neither asymptotics will work. Unfortunately, theory alone does not dictate which asymptotic theory applies in practice. This also depends on the type of data at hand, and the sample sizes used for determination of the weights of the individual models.

Note that fixed  $k$  asymptotic theory predicts that MMA, GRA, and  $BMA_{lin}$  perform comparably well for large calibration samples, while Hansen's asymptotic theory predicts that MMA is superior for large calibration samples. The main purpose of the case studies presented in the next section is to find out, by analyzing of out-of-sample prediction error, whether fixed  $k$  asymptotics, Hansen's asymptotics, or neither, apply in practice.

## 4 Case studies

In this section we compare the different model averaging strategies in terms of their ability to improve the forecast error of two hydrologic systems. The first case study considers daily streamflow forecasting of the Leaf River watershed in Mississippi using an 8-member ensemble of conceptual watershed models. This study deals with a variable that is highly skewed and perhaps not well described by a normal distribution. In the second case study we use seven different soil hydraulic models for forecasting of tensiometric pressure heads in a layered vadose zone in New Zealand.

### 4.1 Streamflow data

In this study, we apply the different model averaging approaches to probabilistic ensemble streamflow forecasting using historical data from the Leaf River watershed (1950 km<sup>2</sup>) located north of Collins, Mississippi. In a previous study Vrugt and Robinson (2007) generated a 36-year ensemble of daily streamflow forecasts using eight different conceptual watershed models involving the ABC (3) (Fiering 1967; Kuczera and Parent 1998), GR4J (4) (Perrin et al. 2003), HYMOD (5) (Boyle et al. 2001; Vrugt et al. 2002; Vrugt et al. 2005), TOPMO (8) (Oudin et al. 2005), AWBM (8) (Boughton 1993; Marshall et al. 2005), NAM (9) (Nielsen and Hansen 1973), HBV (9) (Bergström 1995), and SAC-SMA (13) (Burnash et al. 1973). These eight models are listed in order of increasing complexity, and the number of user-specified parameters is indicated in parentheses. Inputs to the models include mean areal precipitation (MAP), and potential evapotranspiration (PET). The output is estimated channel streamflow. In this study, the first 3000 data points on record (roughly corresponding to the first 8 years of data, WY 1953–1960) are used for

calibration of the parameters in the various hydrologic models.

The  $Y$ -variable,  $Y_t$ , is taken to be the observed river discharge, while the vector of  $X$ -variables comprises an ensemble of one-day-ahead predictors for  $Y_t$  constructed with eight different models, with respective number of parameters 3, 4, 5, 8, 8, 9, 9, and 13. Previous studies have indicated that a calibration data set of approximately 8–11 years of data, representing a range of hydrologic phenomena (wet, medium, and dry years), is desirable to achieve deterministic model calibrations that are consistent and generate good verification and forecasting performance, see e.g., Yapo et al. (1996) and Vrugt et al. (2006b). In this study we systematically vary the length of the calibration data set by using an increasing number of data before the end of the full calibration data set of 3000 observations. The remaining 28 years on record (WY 1961–1988, 10,500 observations) are subsequently used to evaluate the forecast performance.

Table 2 shows the estimates of the weights, as well as the RMSE measured across the evaluation period for each of the forecast methods, based on the full calibration data set (3000 data points). Using OLS for the evaluation period, it is possible to find the weights of the linear regression model that are, ex post, optimal for the specific evaluation data at hand. The corresponding model weights for this best model are given in the row indicated by ' $\beta_{opt}$ '. For comparison, the performance of the forecasts of the individual ensemble members is shown in the last eight rows.

By comparing the RMSE values displayed in the before-last column of Table 2, it can be seen that not all model averaging methods produce forecasts that are more accurate than the best performing individual model of the ensemble. Specifically, the predictive performance of  $BMA_{mix}$ ,  $MMA^{\Delta}$ , BGA, and EWA is worse than that of the best ensemble member, SAC-SMA, indicated in the table by  $X_8$ . AICA and BICA, on the contrary, put all their weight on the best ensemble member, and therefore exhibit similar predictive capabilities as this best member.  $BMA_{lin}^{\Delta}$  with weights restricted to the simplex beats the best individual ensemble member, but performs worse than its unrestricted counterpart  $BMA_{lin}$ . Unrestricted  $BMA_{lin}$  and MMA, and GRA each perform very well, with practically identical values of the weights. Note that these three best performing methods all have unrestricted weights, the estimations of which have the same signs as the optimal weights, given in row ' $\beta_{opt}$ '. The last column, denoted 'RMSE\*', shows the RMSE values obtained without performing the bias correction step prior to model averaging. The RMSE values are slightly larger without a bias correction, but the pattern is very similar.

To assess the effect of the length of the calibration data set, the length of the calibration data set was progressively

**Table 2** Streamflow data results

Method	$\Delta$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	$\hat{\beta}_8$	RMSE	RMSE*
Finite mixture model											
BMA <sub>mix</sub>	1	0.018	0.193	0.098	0.069	0.033	0.050	0.043	0.495	21.89	22.36
Linear regression model											
BMA <sub>lin</sub> <sup><math>\Delta</math></sup>	1	0.000	0.164	0.001	0.316	0.000	0.000	0.007	0.536	21.62	21.98
BMA <sub>lin</sub>	0	−0.072	0.092	0.094	0.590	−0.109	−0.230	−0.045	0.653	21.41	21.45
MMA <sup><math>\Delta</math></sup>	1	0.000	0.144	0.002	0.333	0.000	0.000	0.002	0.540	21.88	21.94
MMA	0	−0.109	0.088	0.128	0.603	−0.139	−0.214	−0.062	0.671	21.43	21.48
Combined point forecasts											
AICA	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	21.73	21.96
BICA	1	0.000	0.000	0.000	0.000	0.000	0.000	0.000	1.000	21.73	21.96
GRA	0	−0.074	0.090	0.094	0.582	−0.104	−0.237	−0.046	0.667	21.38	21.44
BGA	1	0.051	0.137	0.139	0.159	0.072	0.122	0.135	0.185	24.72	24.97
EWA	1	0.125	0.125	0.125	0.125	0.125	0.125	0.125	0.125	26.38	26.79
Best forecast											
$\beta_{\text{opt}}$	0	−0.066	0.254	0.115	0.212	−0.122	−0.170	−0.021	0.731	20.29	20.33
Individual forecasts											
$X_1$	−	1	0	0	0	0	0	0	0	49.00	50.39
$X_2$	−	0	1	0	0	0	0	0	0	25.03	25.26
$X_3$	−	0	0	1	0	0	0	0	0	28.78	28.72
$X_4$	−	0	0	0	1	0	0	0	0	27.55	27.48
$X_5$	−	0	0	0	0	1	0	0	0	41.65	42.18
$X_6$	−	0	0	0	0	0	1	0	0	32.84	32.87
$X_7$	−	0	0	0	0	0	0	1	0	30.36	31.01
$X_8$	−	0	0	0	0	0	0	0	1	21.73	21.96

The first column indicates the averaging method used. The second column, labeled  $\Delta$ , has an entry ‘1’ in case the weights were restricted to the simplex, and ‘0’ otherwise. Columns 3–10 show the weights found for each method. The column labeled RMSE provides the RMSE obtained in the evaluation data for the weights given. For comparison the column labeled RMSE\* gives the RMSE obtained without bias correction step

increased to include more past data in the calibration period, while keeping the evaluation period fixed. To obtain a convenient scale for plotting the results, we focus on the difference between the out-of-sample RMSE and the minimal RMSE (achieved for  $\beta_{\text{opt}}$ ), denoted by  $\Delta\text{RMSE}$ . We refer to  $\Delta\text{RMSE}$  as the ‘excess prediction error’.

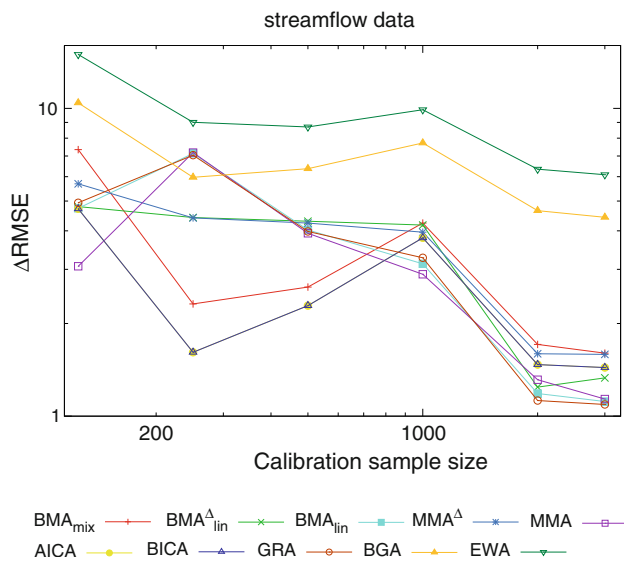
Figure 1 displays  $\Delta\text{RMSE}$  for increasing lengths of the calibration data set. The graph clearly indicates that EWA and BGA are performing poorly relative to the other methods, regardless of the length of the calibration period. The performance of AICA and BICA in this case study is equal for all calibration sample sizes; a result of the fact that these methods essentially selected the same weight vector throughout, putting all weight on the model that performed best in the calibration period. For smaller calibration sample sizes it is particularly difficult to find a single method that is consistently superior to the other averaging strategies. However, for larger calibration data sets ( $n = 3,000$ ) BMA<sub>lin</sub>, MMA, and GRA perform best, exactly as predicted by fixed- $k$  asymptotic theory.

## 4.2 Soil water pressure head data

The second case study concerns tensiometric pressure head data in a layered vadose zone of volcanic origin. These field data from the Spydia experimental site in the northern Lake Taup catchment, New Zealand, consist of water pressure measurements taken at various depths. These data have been described in detail by Wöhling et al. (2008). The dataset originally contained tensiometric pressure measurements at five different depths. In this case study we focus on the pressure data at the smallest depth (0.4 m), for which the data are the most dynamic.

Ensemble forecast combinations for these data have been studied by Wöhling and Vrugt (2008). They considered hourly ensemble forecasts based on seven different soil hydraulic models, listed here with their respective number of parameters in brackets: modified Mualem-van Genuchten (MVG), nonhysteric (15), MVG, hysteric (24), Brooks and Corey (21), Kosugi (24), Durner (15), Šimunek et al. (15), and non-hysteric MVG model with four soil





**Fig. 1** Streamflow data. Excess prediction error  $\Delta RMSE$  for increasing calibration sample size (logarithmic scales)

horizons (20). This latter model is an extension of the first MVG soil hydraulic model used in this study, but uses four instead of three soil horizons to further improve the vertical

description of moisture flow and uptake in the unsaturated zone. This model contains five additional parameters that define the hydraulic functions for this fourth horizon. These seven soil hydraulic models encompass not only different formulations of the same physical relationships but also different conceptual models.

A detailed description of each of these seven models has been given in Wöhling and Vrugt (2008) and so will not be repeated here. Here, the same evaluation period as in Wöhling and Vrugt (2008) is used (the last 2,301 observations). The length of the calibration data set is varied by using calibration data sets of various lengths at the end of the period of the first 6,769 observations, which were used as the calibration data set by Wöhling and Vrugt (2008).

Table 3 displays the estimated weights, as well as the RMSE measured across the evaluation period for each of the forecast methods, based on the full calibration set (6,769 observations). The layout is similar to that of Table 2, with the main difference that there are now seven ensemble members instead of eight.

Inspection of the RMSE values shows that EWA and BGA again perform worse than the best performing individual model (MVG, hysteric, denoted by  $X_2$  in Table 2). However, unlike for the streamflow data, now  $BMA_{mix}$  and

**Table 3** Tensiometric pressure head data, depth 0.4 m

Method	$\Delta$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$	$\hat{\beta}_5$	$\hat{\beta}_6$	$\hat{\beta}_7$	RMSE	RMSE*
Finite mixture model										
$BMA_{mix}$	1	0.000	0.294	0.009	0.065	0.001	0.000	0.632	0.104	0.106
Linear regression model										
$BMA_{lin}^{\Delta}$	1	0.000	0.152	0.000	0.256	0.000	0.000	0.595	0.106	0.104
$BMA_{lin}$	0	-0.189	0.481	-0.363	0.551	-0.633	0.369	0.785	0.103	0.103
$MMA^{\Delta}$	1	0.000	0.148	0.000	0.259	0.000	0.000	0.597	0.106	0.104
$MMA$	0	-0.187	0.484	-0.367	0.548	-0.633	0.369	0.786	0.103	0.103
Combined point forecasts										
AICA	1	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.107	0.112
BICA	1	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.107	0.112
GRA	0	-0.194	0.483	-0.364	0.550	-0.635	0.373	0.786	0.103	0.103
BGA	1	0.086	0.155	0.099	0.183	0.132	0.097	0.247	0.110	0.110
EWA	1	0.143	0.143	0.143	0.143	0.143	0.143	0.143	0.115	0.113
Best forecast										
$\beta_{opt}$	0	1.230	0.437	0.218	0.338	0.967	-2.142	-0.047	0.058	0.069
Individual forecasts										
$X_1$	-	1	0	0	0	0	0	0	0.137	0.124
$X_2$	-	0	1	0	0	0	0	0	0.111	0.110
$X_3$	-	0	0	1	0	0	0	0	0.139	0.175
$X_4$	-	0	0	0	1	0	0	0	0.114	0.116
$X_5$	-	0	0	0	0	1	0	0	0.119	0.144
$X_6$	-	0	0	0	0	0	1	0	0.131	0.110
$X_7$	-	0	0	0	0	0	0	1	0.107	0.112

The entries are as described for Table 2, with the difference that there are now 7 competing models instead of 8

MMA<sup>Δ</sup> perform slightly better than this simple benchmark. AICA and BICA are borderline cases, effectively putting all weight on the best performing individual model. Within the groups BMA<sub>lin</sub> and MMA, the versions with unrestricted weights again perform better than the restricted version. Unrestricted BMA<sub>lin</sub> and MMA, and GRA all perform best, with practically identical weight vectors. An important difference with the previous case study is that these three best performing methods now do not have weights with the same signs as the optimal weights (row ' $\beta_{\text{opt}}$ '). This indicates that the estimated weights are not optimal, which suggests that there is a structural difference between the calibration and the evaluation periods. This can easily happen, even for stationary time series data, if the calibration sample is too small. This is not unlikely to occur in hydrological applications. Note however that the excess prediction errors for BMA<sub>lin</sub>, MMA and GRA are smaller than that of the best performing single model, and hence model averaging still is able to reduce forecast errors in such a case. Judging from the RMSE\*-values reported in the last column, similar results would have been obtained if no bias correction would have been used.

The effect of the length of the calibration data set was again assessed by progressively increasing the length of the calibration data set, while keeping the evaluation period fixed. Figure 2 displays the excess prediction error ( $\Delta\text{RMSE}$ ) for increasing lengths of the calibration data set. The results are in line with the previous case study. It can be observed that EWA and BGA perform poorly, particularly for larger calibration period. Except for the smaller calibration sample sizes, the performances of AICA and BICA are equal (beyond  $n = 2,000$ ). Again, for the smaller calibration sample sizes no obvious winning method can be

distinguished, while for the largest calibration sample size considered ( $n = 6,769$ ) BMA<sub>lin</sub>, MMA and GRA perform best, and almost identically, as predicted by fixed  $k$  asymptotic theory.

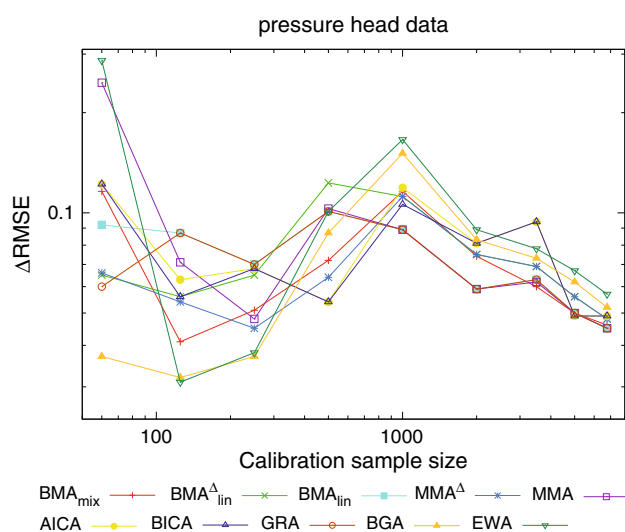
## 5 Discussion

Our case studies clearly indicate a couple of important conclusions for model averaging. Perhaps the most striking observation is that the methods that imposed the restriction that the model weights should be positive were performing worst. None of the three best methods had this restriction. We conclude that not only there are theoretical arguments for allowing negative weights in general, but the empirical results suggest that in surface and subsurface hydrologic case studies the best forecasts may have weights outside the simplex.

Another main conclusion is that, for the particular data sets considered, a relatively simple model averaging method, GRA, based on OLS regression, is shown to perform just as well as very sophisticated competing model averaging methods, such as BMA and MMA. The gain obtained from using GRA rather than BMA or MMA can be significant; analytical solutions exist to determine the weights for GRA, whereas computationally more demanding, iterative procedures, such as expectation-maximization or MCMC sampling with DREAM are required to find the optimal values of the weights for BMA and MMA.

The result that these methods were found to perform well here arguably depends on choosing the RMSE as a measure of performance. As indicated in the introduction the RMSE is a widely used measure for evaluating the accuracy of point forecasts. Moreover, the various model averaging strategies used herein are developed within the context of minimizing the squared prediction error. The only exception is BMA<sub>mix</sub>, whose weights are calibrated by posing the calibration problem in a maximum likelihood context. One might use different evaluation criteria, such as mean absolute error, yet to make a fair comparison of the different model averaging methods would then require the explicit use of these criteria within the calibration set. In that case we would be introducing new model averaging methods altogether, which is beyond the scope of this paper. Another possibility might be to use various calibration criteria simultaneously and interpret the Pareto solution set using multi-criteria optimization methods. Such an approach within the context of BMA<sub>mix</sub> has previously been presented in Vrugt et al. (2006a).

For increasing calibration sample size, it could be observed that the three asymptotically RMSE optimal model averaging methods, BMA, MMA, and GRA, quickly



**Fig. 2** Tensiometric pressure head data. Excess prediction error  $\Delta\text{RMSE}$  for increasing calibration sample size (logarithmic scales)

dominated the other methods, while converging towards the same common small forecast error, just as asymptotic theory with fixed finite  $k$  predicts. That this would be the case could not be known a priori because the finite sample size  $n$  for which asymptotics work well depends on the type of data under consideration. We can therefore conclude that the calibration sample sizes used typically in hydrology, are large enough, relative to  $k$ , to justify the use of predictions from asymptotic theory with a fixed number of ensemble members,  $k$ , and large sample size,  $n$ .

By using the ex post optimal model parameter  $\beta_{\text{opt}}$  for the evaluation sample it could be observed that the errors were not evenly distributed across the data. For the streamflow data, for instance, larger flows had larger prediction errors. Although we have not performed a quantitative study since we have focused on batch-optimization in the case studies, from graphical inspection of the data and the predictors it appeared that the best performing GRA method was specifically outperforming the other models during the high-flow periods. For the pressure head data similar observations could be made.

A final conclusion that can be drawn from our results is that BMA<sub>mix</sub>, which is increasingly being used in hydrology for obtaining density forecasts while taking into account model uncertainty, has point predictors that are not asymptotically optimal. Unfortunately, the asymptotically optimal GRA weights do not easily allow for the construction of a density forecast, because the weights are allowed to be negative. It therefore seems desirable to develop density forecasts that represent model uncertainty accurately, while at the same time having associated point predictors achieving a close to minimal prediction error. We leave the development of such a model averaging method for future research.

**Acknowledgments** The second author is supported by a J. Robert Oppenheimer Fellowship from the Los Alamos National Laboratory Postdoctoral Program. The authors gratefully acknowledge Thomas Wöhling, Lincoln Ventures Ltd., Ruakura Research Center, Hamilton, New Zealand, for kindly providing the tensiometric pressure head data.

**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

- Bates JM, Granger CWJ (1969) The combination of forecasts. *Oper Res Q* 20:451–468
- Bergström S (1995) The HBV model. In: Singh VP (ed) *Computer models of watershed hydrology*. Water Resources Publications, Highlands Ranch, Colorado, pp 443–476
- Bishop CH, Shanley KT (2008) Bayesian modeling averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon Weather Rev* 136:4641–4652
- Boughton WC (1993) A hydrograph based model for estimating the water yield of ungauged catchments. *J Irrigat Drain Eng* 116:83–98
- Boyle DP, Gupta HV, Sorooshian S, Koren V, Zhang Z, Smith M (2001) Toward improved streamflow forecast: value of semidistributed modeling. *Water Resour Res* 37(11):2749–2759
- Buckland ST, Burnham KP, Augustin NH (1997) Model selection: an integral part of inference. *Biometrics* 53:603–618
- Burnash RJ, Ferral RL, McQuir RA (1973) A generalized streamflow simulation system. Technical Report. Joint Federal-State River Forecast
- Burnham KP, Anderson DR (2002) *Model Selection and Multimodel Inference: a practical information-theoretic approach*, 2nd edn. Springer, New York
- Claeskens G, Hjort NL (2003) The focused information criterion. *J Am Stat Assoc* 98:900–916
- Fiering MB (1967) *Streamflow synthetic*. MacMillan, London
- Gneiting T, Raftery AE, Westveld AH, Goldman T (2005) Calibrated probabilistic forecasting using ensemble model output statistics and CRPS estimation. *Mon Weather Rev* 133:1098–1118
- Granger CWJ, Ramanathan R (1984) Improved methods of combining forecasts. *J Forecast* 3:197–204
- Hansen BE (2007) Least-squares model averaging. *Econometrica* 75:1175–1189
- Hansen BE (2008) Least-squares forecast averaging. *J Econom* 146:342–350
- Hjort NL, Claeskens G (2003) Frequentist model average estimators. *J Am Stat Assoc* 98:879–899
- Hoeting JA, Madigan D, Raftery AE, Volinsky CT (1999) Bayesian model averaging: a tutorial. *Stat Sci* 14:382–417
- Kuczera G, Parent E (1998) Monte Carlo assessment of parameter uncertainty in conceptual catchment models: the metropolis algorithm. *J Hydrol* 211:69–85
- Marshall L, Nott D, Sharma A (2005) Hydrological model selection: a bayesian alternative. *Water Resour Res* 41. doi:10.1029/2004WR003719
- Neuman SP (2003) Maximum likelihood Bayesian averaging of uncertain model predictions. *Stoch Environ Res Risk Assess* 17:291–305
- Nielsen SA, Hansen E (1973) Numerical simulation of the rainfall runoff processes on a daily basis. *Nordic Hydrol* 4:171–190
- Oudin L, Perrin C, Mathevet T, Andréassian V, Michel C (2005) Impact of biased and randomly corrupted inputs on the efficiency and the parameters of watershed models. *J Hydrol* 320:62–83. doi:10.1016/j.jhydrol.2005.07.016
- Perrin C, Michel C, Andréassian V (2003) Improvement of a parsimonious model for streamflow simulation. *J Hydrol* 279:275–289
- Raftery AE, Madigan D, Hoeting JA (1997) Bayesian model averaging for linear regression models. *J Am Stat Assoc* 92:179–191
- Raftery AE, Gneiting T, Balabdaoui F, Polakowski M (2005) Using Bayesian model averaging to calibrate forecast ensembles. *Mon Weather Rev* 133:1155–1174
- Rojas R, Feyen L, Dassargues A (2008) Conceptual model uncertainty in groundwater modeling: combining generalized likelihood uncertainty estimation and Bayesian model averaging. *Water Resour Res* 44:1–16
- Stockdale TN (2000) An overview of techniques for seasonal forecasting. *Stoch Environ Res Risk Assess* 14:305–318
- Vrugt JA, Robinson BA (2007) Treatment of uncertainty using ensemble methods: comparison of sequential data assimilation and Bayesian model averaging. *Water Resour Res* 43:1–15

- Vrugt JA, Bouten W, Gupta HV, Sorooshian S (2002) Toward improved identifiability of hydrologic model parameters: the information content of experimental data. *Water Resour Res* 38(12). doi:[10.1029/2001WR001118](https://doi.org/10.1029/2001WR001118)
- Vrugt JA, Diks CGH, Gupta HV, Bouten W, Verstraten JM (2005) Improved treatment of uncertainty in hydrologic modeling: combining the strengths of global optimization and data assimilation. *Water Resour Res* 41(1). doi:[10.1029/2004WR003059](https://doi.org/10.1029/2004WR003059)
- Vrugt JA, Clark MP, Diks CGH, Duan Q, Robinson BA (2006a) Multi-objective calibration of forecast ensembles using bayesian model averaging. *Geophys Res Lett* 33. doi:[10.1029/2006GL027126](https://doi.org/10.1029/2006GL027126)
- Vrugt JA, Gupta HV, Dekker SC, Sorooshian S, Wagener T, Bouten W (2006b) Application of stochastic parameter optimization to the Sacramento soil moisture accounting model. *J Hydrol* 325: 288–307. doi:[10.1016/j.hydrol.2005.10.041](https://doi.org/10.1016/j.hydrol.2005.10.041)
- Vrugt JA, Diks CGH, Clark MP (2008) Ensemble bayesian model averaging using Markov chain Monte Carlo sampling. *Environ Fluid Dyn* 8:579–595
- Vrugt JA, ter Braak CJF, Diks CGH, Robinson BA, Hyman JM, Higdon D (2009) Accelerating Markov chain Monte Carlo simulation by self-adaptive differential evolution with randomized subspace sampling. *Int J Nonlinear Sci Numer Simul* 10:273–290
- Wagener T, Gupta HV (2005) Model identification for hydrological forecasting under uncertainty. *Stoch Environ Res Risk Assess* 19:378–387
- Wan ATK, Zhang X, Zou G (2010) Least squares model averaging by Mallows criterion. *J Econom* (forthcoming)
- Wöhling T, Vrugt JA (2008) Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models. *Water Resour Res* 44:1–18
- Wöhling T, Vrugt JA, Barkle GF (2008) Comparison of three multiobjective algorithms for inverse modeling of vadose zone hydrologic properties. *Soil Sci Soc Am J* 72:305–319
- Yapo P, Gupta HV, Sorooshian S (1996) Automatic calibration of conceptual rainfall-runoff models: sensitivity to calibration data. *J Hydrol* 181:23–48
- Ye M, Neuman SP, Meyer PD (2004) Maximum likelihood bayesian averaging of spatial variability models in unsaturated fractured tuff. *Water Resour Res* 40. doi:[10.1029/2003WR002557](https://doi.org/10.1029/2003WR002557)
- Ye M, Meyer PD, Neuman SP (2008) On model selection criteria in multimodel analysis. *Water Resour Res* 44:1–12