

MODELAVG: A MATLAB Toolbox for Postprocessing of Model Ensembles

Jasper A. Vrugt^{a,b}

^a*Department of Civil and Environmental Engineering, University of California Irvine, 4130 Engineering Gateway, Irvine, CA 92697-2175*

^b*Department of Earth System Science, University of California Irvine, Irvine, CA*

Abstract

Model averaging is statistical method that is widely used to quantify the conceptual uncertainty of environmental system models and to improve the sharpness and skill of forecast ensembles of multi-model prediction systems. Here, I present a MATLAB toolbox for postprocessing of forecast ensembles. This toolbox, called MODELAVG implements many different model averaging techniques, including methods that provide point forecasts only, and methods that produce a forecast distribution of the variable(s) of interest. MCMC simulation with the DREAM_(zs) algorithm is used for model averaging methods that require an iterative solution for their weights and/or shape parameters. The toolbox returns to the user (among others) a vector (or matrix with posterior samples) of weights and (if appropriate) variance(s) and other shape parameters of the members' forecast distribution, a vector of averaged forecasts (and performance metrics thereof), and (if appropriate) estimates of the width (spread) and coverage of the forecast distribution. We also monitor convergence of the DREAM_(zs) algorithm to the target distribution. Three case studies illustrate the capabilities of the MODELAVG toolbox.

Keywords: Model averaging, Information criterion averaging, Bayes information criterion, Equal weights averaging, Granger-Ramanathan averaging, Bates-Granger averaging, Mallows model averaging, Bayesian model averaging, Markov chain Monte Carlo simulation, DREAM

Email address: jasper@uci.edu (Jasper A. Vrugt)
URL: <http://faculty.sites.uci.edu/jasper> (Jasper A. Vrugt),
<http://scholar.google.com/citations?user=zkNXecUAAAJ&hl=en> (Jasper A. Vrugt)

1. Introduction and Scope

Multi-model ensemble prediction systems are used by many agencies in the world to forecast the behavior of complex systems. Such systems much better capture the uncertainty of the initial states, boundary conditions, and model physics, and therefore produce more skillful predictions than forecasts derived from a single model run. Yet, as most ensemble prediction systems do not account perfectly for all sources of uncertainty, some postprocessing is necessary to provide accurate forecasts.

Model averaging is a statistical methodology that can be used to improve the skill of a multi-model ensemble. This methodology can also be used to quantify conceptual model uncertainty as predictions generated by a single model are prone to statistical bias (by reliance on an invalid model) and underestimation of uncertainty (by under-sampling the feasible model space) (*Raftery et al.*, 1999; *Hoeting et al.*, 1999; *Neuman*, 2003; *Raftery et al.*, 2005; *Vrugt et al.*, 2006). Figure 1 illustrates the concept of model averaging. Consider that at a given time we have available the output of multiple different models. These models do not necessarily have to be calibrated. Now the goal is to weight the different models in such a way that the weighted estimate (model) is a better (point) predictor of the observed system behavior (data) than any of the individual models of the ensemble. Moreover, the density of the averaged model characterizes adequately the total predictive uncertainty.

To formalize the various model averaging strategies considered herein, let me denote by $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^\top$ a $n \times 1$ vector of measurements of some quantity of interest, where the symbol \top signifies transpose. These observations can be made at different locations in space. Without loss of generality, I conveniently ignore the spatial coordinates and restrict attention to a time series of measurements of system behavior. We assume that K different models simulate and/or predict the observed data. The point forecasts of each model are stored in a $n \times 1$ vector $\mathbf{y} = (y_1, \dots, y_n)^\top$ and these column vectors are combined into a $n \times K$ matrix \mathbf{D} , where d_{tk} stores the prediction of the k^{th} model $k = (1, \dots, K)$ at time $t = (1, \dots, n)$. The t^{th} observation \tilde{y}_t may be written as weighted average prediction of the point forecasts of the K models as follows

$$\tilde{y}_t = \mathbf{d}_t \boldsymbol{\beta} + \epsilon_t = \sum_{k=1}^K \beta_k d_{tk} + \epsilon_t, \quad (1)$$

where the $1 \times K$ vector $\mathbf{d}_t = (d_{t1}, \dots, d_{tK})$ is equal to the t^{th} row of matrix \mathbf{D} , $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$ signifies the $K \times 1$ vector of weights and the residual $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ is assumed to be normally distributed with zero mean and constant variance, σ_ϵ^2 . In the remainder of this manual, the index t is used to mean "for all $t \in (1, \dots, n)$ ".

A bias correction step of the individual forecasts is performed before the weights of the individual models are estimated. For instance, if we assume a linear transformation of the raw point forecasts of the k^{th} model

$$d_{tk}^b = a_k + b_k d_{tk}, \quad (2)$$

then the intercept a_k and dimensionless slope b_k of each model $k = (1, \dots, K)$ can be calculated by ordinary least squares using the $n \times 1$ vector of training data observations $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^\top$ to yield

$$\begin{bmatrix} \hat{a}_k \\ \hat{b}_k \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \tilde{\mathbf{y}}, \quad (3)$$

MODEL AVG MANUAL

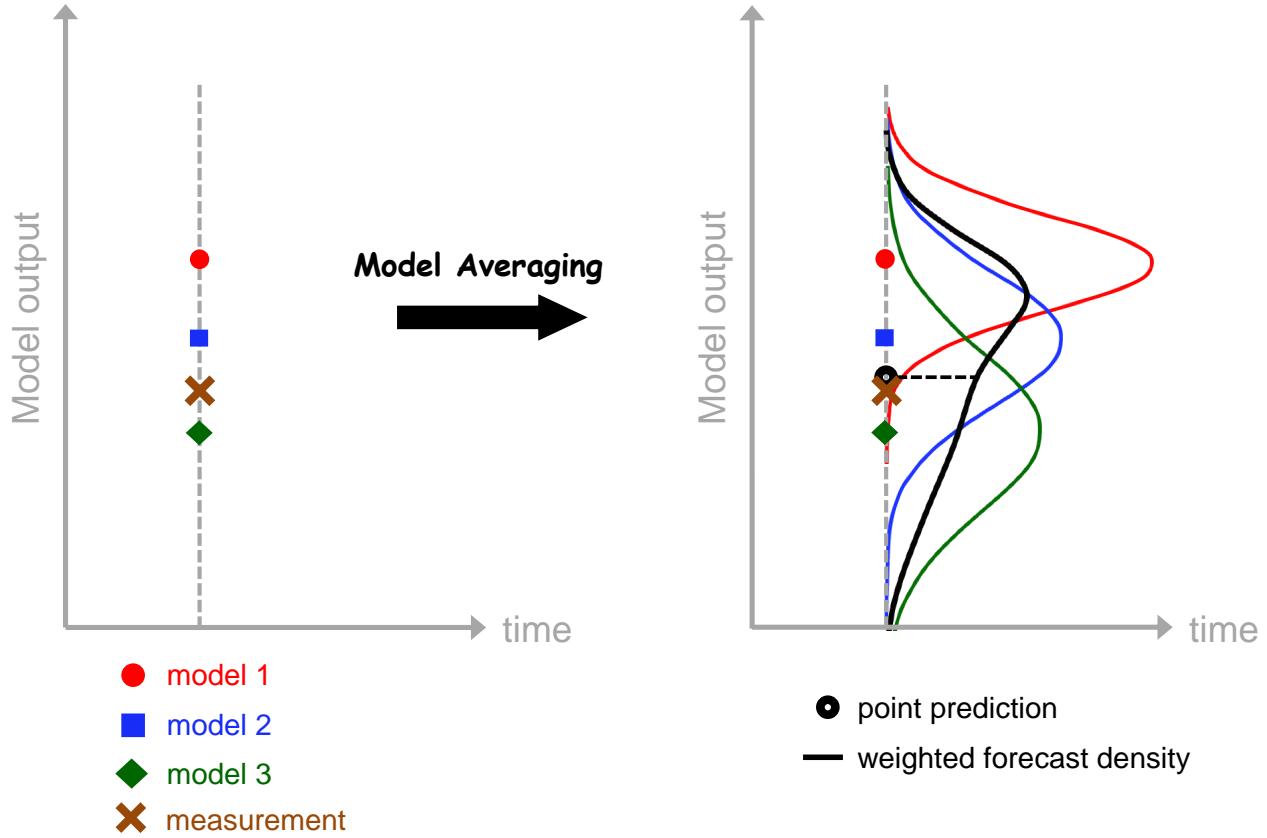


Figure 1: Schematic illustration of model averaging using a three member ensemble and single prediction of interest. The forecast of each models are displayed with the solid red circle, blue square and green diamond, respectively, and the verifying observation is indicated separately with the brown "X" symbol. The dotted black line connected with the symbol "○" denotes the weighted average of the forecasts of the three different models. This point predictor satisfies the underlying premise of model averaging as it is in better agreement with the data (smaller distance) than any of the three models of the ensemble. Some model averaging methods also construct a predictive density of this averaged forecast. This overall forecast pdf (solid black line) can be used for probabilistic forecasting and uncertainty analysis, and is simply a weighted average of each ensemble members predictive distribution (indicated with solid red, blue and green lines) centered at their respective forecasts.

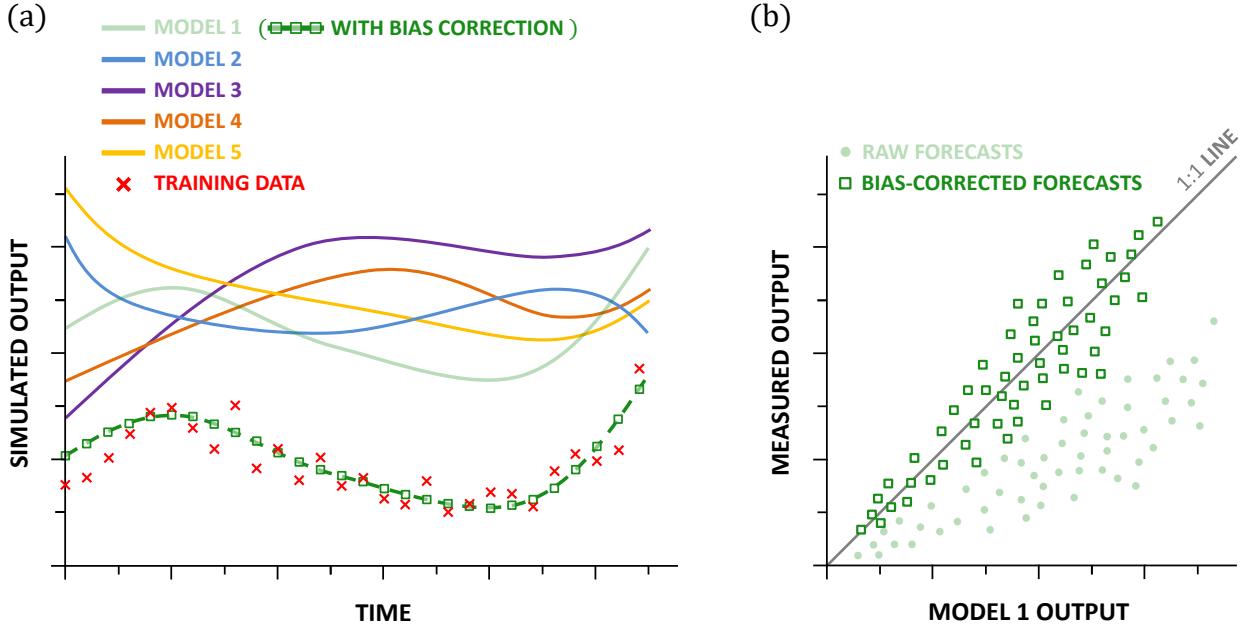


Figure 2: Schematic illustration of bias correction using a hypothetical ensemble of five different models: (a) time series (solid lines) of the simulated outputs of the quantity of interest of the different models. The $n \times 1$ vector of verifying observations $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^\top$ is presented with the X symbol, (b) two-dimensional scatterplot of the simulated output of the first model and the verifying observations of the training data set. The light-green dots correspond to the raw point forecasts of model 1, whereas the green squares signify the bias-corrected forecasts derived from linear regression of the raw forecasts using the training data observations.

where the $n \times 2$ design matrix \mathbf{X} has unit entries in the first column and the raw forecasts of the k^{th} model $(d_{1k}, \dots, d_{nk})^\top$ in the second column. This bias correction step is rather simplistic but deemed necessary for removing systematic bias between the models' point forecasts and the verifying observations (see Figure 2). For an unbiased model, the a_k 's should be near zero and the b_k 's should attain values close to one. If the training data set is small, the least squares estimates of the intercepts and slopes may become unstable, and bias correction may actually degrade the predictive capabilities of the ensemble (Vrugt and Robinson, 2007). Linear bias correction is performed for each of the constituent models of the ensemble but this step is not made explicit in the subsequent notation. In this manual, I simply use d_{tk} rather than d_{tk}^b for the bias corrected forecasts of \tilde{y}_t , where $t = (1, \dots, n)$.

According to Equation (1), the weighted average point forecast of the ensemble may be written as follows

$$\dot{y}_t = \mathbf{d}_t \boldsymbol{\beta} = \sum_{k=1}^K \beta_k d_{tk}, \quad (4)$$

where the bullet \bullet in superscript is used to indicate the expected value of the averaged model.

In this manual, I introduce a MATLAB toolbox for postprocessing of forecast ensembles. This toolbox, called MODELAVG implements a large number of model averaging techniques, including methods that provide only a point forecast, and methods that produce a forecast distribution of the variable(s) of interest. MCMC simulation with the DREAM_(ZS) algorithm is used for averaging methods without a direct closed-

form solution of their (optimal) point forecasts. The toolbox returns to the user (among others) a vector (or matrix with posterior samples) of weights and (if appropriate) standard deviation(s) of the members' forecast distribution, a vector of averaged forecasts (and performance metrics thereof), and (if appropriate) estimates of the width and coverage of the forecast distribution, and convergence diagnostics of the DREAM_(ZS) algorithm. The various options of the toolbox are illustrated using three different case studies involving ensemble forecasts of river discharge, sea surface temperature and sea level pressure. These studies serve as templates for other data sets.

The remainder of this manual is organized as follows. Section 2 summarizes the theory of each of the model averaging methods that is available to the user. This is followed in section 3 with a detailed description of the MODELAVG toolbox. In this section I discuss each of the input and output arguments of the toolbox, and discuss the various options available to the user. Section 4 illustrates the different functionalities of the toolbox by application to three different forecast ensembles. Section 5 highlights recent research efforts aimed at further improving the sharpness and coverage of the ensemble. Finally, section 6 provides a summary of the content of this manual.

2. Model Averaging methods

The MATLAB toolbox MODELAVG implements seven different model averaging techniques. These methods will be described in this section. Some of these methods restrict the weights $\beta = (\beta_1, \dots, \beta_K)^\top$ of the ensemble members to the unit simplex \mathcal{S}^K also called standard simplex or probability simplex

$$\mathcal{S}^K = \{\beta \in \mathbb{R}_+^K : \beta_1 + \dots + \beta_K = 1, \beta_k \geq 0 \text{ for } k = (1, \dots, K)\}, \quad (5)$$

Other methods in the toolbox relax this assumption and allow for any values (positive and negative) of the weights, $\beta = (\beta_1, \dots, \beta_K)^\top \in \mathbb{R}^K$.

2.1. Equal weights averaging

Equal weights averaging (EWA) assumes that each member of the ensemble has an equal weight

$$\beta_{\text{EWA}} = \left(\frac{1}{K}, \dots, \frac{1}{K} \right), \quad (6)$$

independent of each model's performance for the training data set, $\tilde{\mathbf{y}}$. The weighted forecast $\dot{y}_t = \frac{1}{K} \sum_{k=1}^K d_{tk}$ at time $t = (1, \dots, n)$ is simply equal to the ensemble mean.

2.2. Bates-Granger averaging

A well-known choice proposed by *Bates and Granger* (1969) is to set the weights of each model proportional to the reciprocal of their respective forecast variance, $\beta_k = 1/\hat{\sigma}_k^2$. The forecast variance of the k^{th} model is determined from its residuals of the training data period

$$\hat{\sigma}_k^2 = \frac{1}{n} \sum_{t=1}^n (\tilde{y}_t - d_{tk})^2. \quad (7)$$

If the models' forecasts are unbiased and their errors uncorrelated, these weights are optimal in the sense of producing point forecasts with the smallest possible Root Mean Square Error (RMSE). To enforce the weights to lie on \mathcal{S}^K they are normalized as follows

$$\beta_{\text{BGA},k} = \frac{1/\hat{\sigma}_k^2}{\sum_{k=1}^K 1/\hat{\sigma}_k^2} \quad (8)$$

so that they add up to one. In the remainder of this manual, I use the acronym BGA for Bates-Granger averaging.

2.3. Information criterion averaging

Information criterion averaging (ICA) was proposed by *Buckland et al.* (1997) and *Burnham and Anderson* (2002) and computes the weights as follows

$$\beta_{\text{ICA},k} = \frac{\exp(-\frac{1}{2}I_k)}{\sum_{k=1}^K \exp(-\frac{1}{2}I_k)}, \quad (9)$$

where I_k is an information criterion that depends on the complexity and goodness-of-fit of each model

$$I_k = -2 \log(L_k) + q(p_k), \quad (10)$$

where L_k is the maximum likelihood of model k , and $q(p_k)$ signifies a penalty term which corrects for the number of model parameters. I consider herein Akaike's information criterion (AIC), for which $q(p) = 2p$, and Bayes information criterion (BIC), for which $q(p) = p \log(n)$. I refer to the model averaging method of Equation (9) as AICA (IC) and BICA (BIC), respectively, and to their respective weights as β_{AICA} and β_{BICA} . In the literature these methods are sometimes referred to as smooth AIC and smooth BIC, respectively. I assume that the number of parameters of each model are stored in the $K \times 1$ vector \mathbf{p} , and thus $\mathbf{p} = (p_1, \dots, p_K)^\top$.

To evaluate the information criteria numerically, it is convenient to assume, as I do herein, that the errors of the individual models are normally distributed. In this case, the log-likelihood of the k^{th} model of the ensemble, $\log(L_k)$, can be calculated using

$$-2 \log(L_k) = n \log(\hat{\sigma}_k^2) + n \quad (11)$$

2.4. Granger-Ramanathan averaging

The weighting schemes described above do not exploit the covariance structure that may be present in the forecast errors of the individual models. A natural way to exploit the presence of covariances is to implement ordinary least squares (OLS) using the regression model of Equation (4).

Granger and Ramanathan (1984) suggests the following OLS estimates of the weights

$$\beta_{\text{GRA}} = (\mathbf{D}^\top \mathbf{D})^{-1} \mathbf{D}^\top \tilde{\mathbf{y}}, \quad (12)$$

where \mathbf{D} is the $n \times K$ matrix of bias-corrected ensemble forecasts and $\tilde{\mathbf{y}}$ signifies the $n \times 1$ vector of training data observations. The OLS estimator can be shown to be the best linear unbiased estimator of β . I

conveniently refer to this model averaging method as GRA.

2.5. Bayesian model averaging

Hoeting et al. (1999) provide an excellent overview of the different variants of Bayesian Model Averaging (BMA) proposed in the literature. BMA differs from the other model averaging methods in this toolbox as it considers explicitly the uncertainty of each model's forecasts - and uses this uncertainty to construct a predictive distribution instead of only a weighted-average, deterministic, forecast. The BMA method offers an alternative to the selection of a single model from a number of candidate models, by weighting each candidate model according to its statistical evidence. Applications of BMA in hydrology and meteorology have been described by *Raftery et al.* (2005), *Gneiting et al.* (2005), *Vrugt and Robinson* (2007), *Vrugt et al.* (2008b) and *Bishop and Shanley* (2008).

The BMA method has several desirable properties, one of which is that it cannot only provides users with a deterministic (averaged) forecast but also with an associated forecast distribution. This forecast distribution summarizes all our knowledge about the target variable of interest, and can be used for probabilistic analysis and/or construction of 90 or 95% intervals. The BMA forecast density imposes one important constraint for the weights however, and that is that they must lie on the unit simplex, $\beta \in \mathcal{S}^K$. Without this restriction the weights can otherwise produce forecast distributions with negative densities. Negative weights are tolerated if the goal of model averaging is point prediction, but cannot be sustained for density forecasts. I now describe an implementation of the BMA method that has found widespread application and use for postprocessing forecast ensembles of dynamic simulation models.

To start, lets assume that the forecasts of each model are subject to uncertainty. We can describe this uncertainty, with an (unknown) forecast distribution, $f_k(\cdot)$. This distribution expresses the prediction uncertainty of each model, $k = (1, \dots, K)$ and its parameters can be inferred from a training data set. For now, I conveniently assume that the forecast distribution is centered at the forecasts of each individual model of the ensemble. The probability density function or PDF $g(y)$ of the BMA model at time $t = (1, \dots, n)$ can be written as follows

$$g_t(y) = \sum_{k=1}^K \beta_k f_{tk}(y) \quad (13)$$

The black line in Figure 1 provides an example of how the BMA density is computed from the individual models' forecast distribution, $f_k(\cdot)$. The cumulative distribution function (CDF) of the BMA model is simply equal to

$$G_t(y) = \sum_{k=1}^K \beta_k F_{tk}(y) \quad (14)$$

where the $F_{tk}(\cdot)$'s are the conditional CDFs at time t of the different members of the ensemble. This CDF will play an important role in determining the prediction quantiles of the BMA model for different significance (confidence) levels. Suppose that the $f_{tk}(\cdot)$'s are assumed equal to a normal distribution with mean equivalent to the model's point forecast d_{tk} and variance σ_k^2 . Then, the probability density function (PDF) of the k^{th} model at time t is equal to

$$f_{tk}(y|d_{tk}, \sigma_k^2) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{1}{2}\sigma_k^{-2}(y - d_{tk})^2\right), \quad (15)$$

and the BMA predictive density $g_t(y)$ is equal to a Gaussian mixture distribution made up of K normal PDFs, each centered at their respective point forecast d_{tk} and with variance σ_k^2 (see Figure 3).

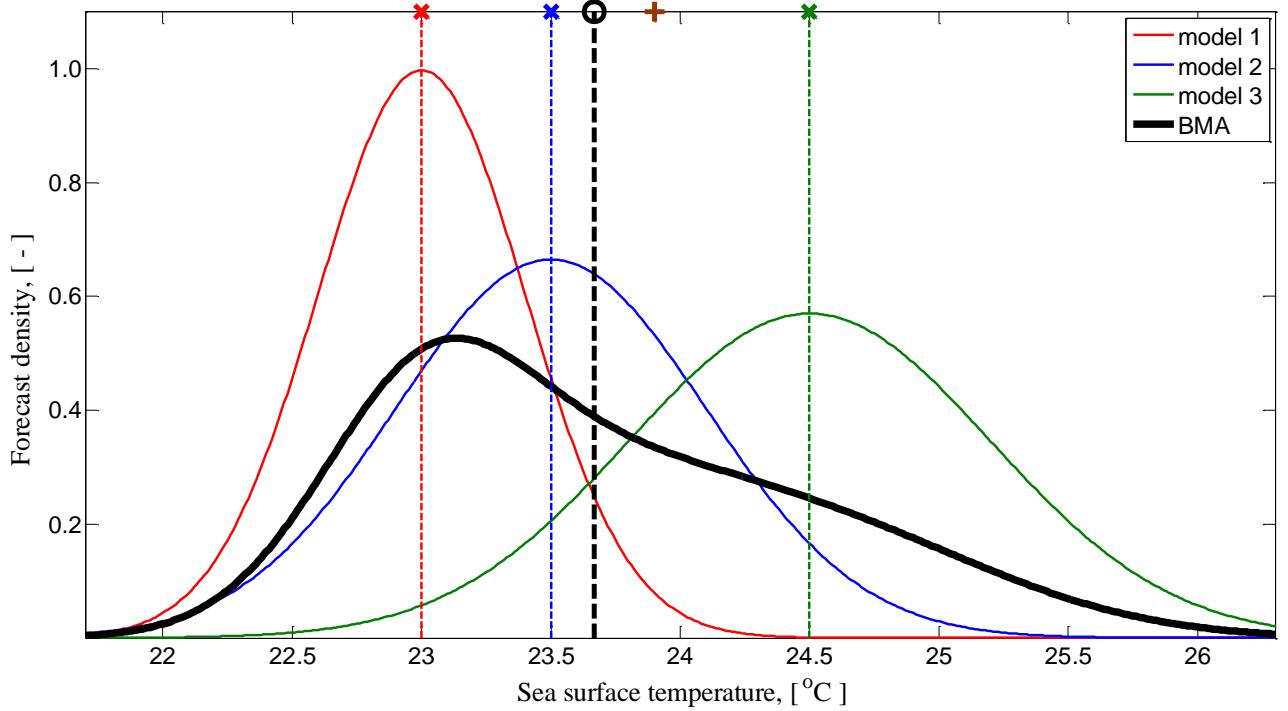


Figure 3: Schematic illustration of Bayesian model averaging using a $K = 3$ member ensemble for the sea surface temperature in degrees Celsius. The BMA predictive pdf $g(y)$ is indicated with the solid black line and equivalent to a weighted average of the conditional forecast distributions, $f_k(\cdot)$, of the members, $k = (1, \dots, K)$ of the ensemble (displayed with solid red, blue and green lines). The forecast density of BMA can be used to compute prediction uncertainty ranges of the quantity of interest (sea surface temperature) at any desired confidence interval, $\alpha = 0.9$, 0.95 or 0.99. Also shown are the individual model forecasts (' \times ' symbol), the BMA deterministic point forecast (' \circ ' symbol), and the verifying observation (' $+$ ' symbol). The deterministic point forecast of BMA can be compared to the ensemble mean and/or point predictors of other model averaging methods.

To ensure that $g(y)$ is a proper density (integrates to one), the BMA weights must lie on the unit simplex, $\mathcal{S}^K \in \mathbb{R}_+^K$ and thus the weights must be strictly positive and add up to one. The BMA weight of each ensemble member can then be viewed as each model's relative contribution to predictive skill over the training (calibration) period. The BMA weights can thus be used to assess the usefulness of ensemble members, and this can be used as a basis for selecting ensemble members given the CPU-cost of running large ensembles (*Raftery et al.*, 2005).

The mean of the BMA forecast density is simply equal to the weighted average of the bias-corrected model forecasts

$$\dot{g}_t = \sum_{k=1}^K \beta_k d_{tk}, \quad (16)$$

and is a deterministic forecast in its own right, whose predictive performance can be compared with the individual models of the ensemble or with the ensemble mean or median. If we assume a normal conditional

PDF for each model of the ensemble, then the BMA forecast variance of Equation (16) can be computed directly using (*Raftery et al.*, 2005)

$$\text{Var}(g_t^\bullet | d_{t1}, \dots, d_{tK}) = \sum_{k=1}^K \beta_k \left(d_{tk} - \sum_{l=1}^K \beta_l d_{tl} \right)^2 + \sum_{k=1}^K \beta_k \sigma_k^2 \quad (17)$$

This variance consists of two terms, the first representing the ensemble spread, and the second representing the within-ensemble forecast variance.

2.5.1. Inference of BMA weights and variances

Successful implementation of the BMA method requires estimates of the weights $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$ and variances $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)^\top$ of the normal conditional pdfs of the ensemble members. Their values can be estimated by maximum likelihood from the vector of training data observations. This estimator has several desirable statistical properties and involves finding the optimum of the likelihood function (*Raftery et al.*, 2005)

$$L(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \mathbf{D}, \tilde{\mathbf{y}}) = \prod_{t=1}^n \sum_{k=1}^K \beta_k f_{tk}(\tilde{y}_t | d_{tk}, \sigma_k^2). \quad (18)$$

where $\boldsymbol{\beta} \in \mathcal{S}^K$ and $\boldsymbol{\sigma}^2 \in \mathbb{R}_+^K$. The likelihood of the BMA parameter values, $\boldsymbol{\Phi} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}^2\}$ is thus computed as follows. First, the density of the BMA mixture distribution is evaluated at each observation of the training data set $\tilde{\mathbf{y}} = (\tilde{y}_1, \dots, \tilde{y}_n)^\top$ using Equation (13). This BMA density is simply a weighted average of the PDFs of the individual ensemble members at the respective observations. These n values are then multiplied and this product is equivalent to the likelihood on the right-hand-side of Equation (18).

Figure 3 presents an example of the density function of the BMA mixture distribution for a three model ensemble with equal weights. If the observation ('+' symbol) were part of the training data set then the density of the BMA mixture distribution (black line) at this observation would equal about 0.3.

The use of the product operator in Equation (18) can pose numerical issues due to rounding errors introduced by the floating-point arithmetic of digital computers. Indeed, if n is sufficiently large, the likelihood of Equation (18) will eventually go to zero. For algebraic simplicity and numerical stability we therefore work with the logarithm of the likelihood instead. For the normal conditional forecast PDF of Equation (15) the log-likelihood function, $\mathcal{L}(\cdot)$, is equivalent to

$$\mathcal{L}(\boldsymbol{\beta}_{\text{BMA}}, \boldsymbol{\sigma}_{\text{BMA}} | \mathbf{D}, \tilde{\mathbf{y}}) = \sum_{t=1}^n \log \left\{ \sum_{k=1}^K \beta_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp \left[-\frac{1}{2} \sigma_k^{-2} (\tilde{y}_t - d_{tk})^2 \right] \right\}. \quad (19)$$

where the summation is over all models, $k = (1, \dots, K)$, and observations, $t = \{1, \dots, n\}$, of the training data set. This thus involves the inference of $d = 2K$ parameters, namely the weight β_k and variance σ_k^2 of each model's conditional normal forecast distribution.

Unfortunately, there are no closed-form solutions that conveniently maximize Equation (19). We therefore have to resort to an iterative solution method to find the optimum weights and variances of the BMA

model with normal conditional PDFs

$$\left(\hat{\beta}_{\text{BMA}}, \hat{\sigma}_{\text{BMA}}^2\right) = \arg \max_{\beta \in \mathcal{S}^K, \sigma^2 \in \mathbb{R}_+^K} \prod_{t=1}^n \sum_{k=1}^K \beta_k f_{tk}(\tilde{y}_t | d_{tk}, \sigma_k^2). \quad (20)$$

In their seminal paper, *Raftery et al.* (2005) recommends the Expectation-Maximization (EM) algorithm for BMA model training. This method alternates between an expectation (E) step, which calculates the expected value of the log-likelihood of Equation (19) at the current estimate of the BMA parameters, and a maximization (M) step, which computes new values of the parameters that maximize the expected log-likelihood value of the E step. A detailed description of the EM method appears in Appendix A of this manual. There, I also present an algorithm that can be used for BMA model training if the forecast pdf of each member is described with a normal distribution.

The EM method exhibits many desirable properties as it is relatively easy to implement, computationally efficient, and the maximization step of Equation (A.2) is designed such that the weights are always positive and add up to one. Nonetheless, global convergence of this algorithm cannot be guaranteed as a single starting point is used in the BMA model space. Of course, multiple different initial starting points can be used, but as each trial operates independently, this is rather CPU-inefficient (*Duan et al.*, 1992). What is more, the mathematical formulations of the E and M step in the EM algorithm depend on the forecast distribution, $f_k(\cdot); k = (1, \dots, K)$ that is used for the members of the ensemble. Indeed, the function `EM_NORMAL` in Appendix A can be used only for variables such as temperature and pressure whose conditional pdf is well described with a normal distribution (see left two plots in Figure 4). The histograms of the other three variables (C: wind speed, D: rainfall, and E: discharge) are truncated by zero and exhibit much more skew to the right. Their conditional pdf is much better approximated by a gamma distribution (*Vrugt and Robinson*, 2007; *Sloughter et al.*, 2010), yet this requires modifications to the E and M step in the EM algorithm (more later).

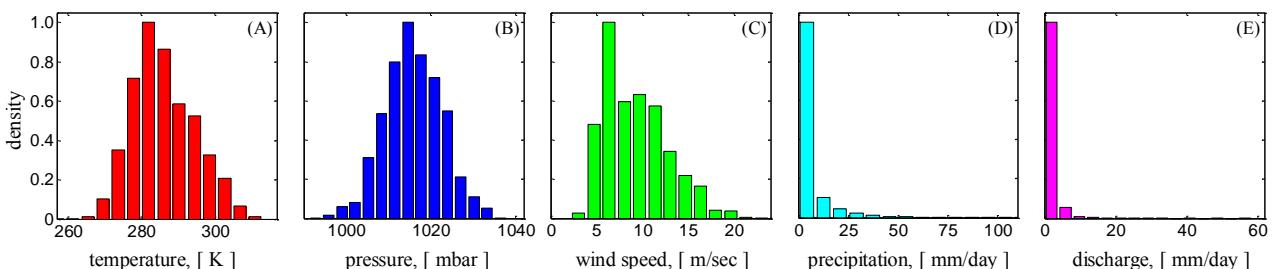


Figure 4: Histograms of daily measurement records of five different variables, including (A) outside temperature [K], (B) atmospheric pressure [mbar], (C) wind speed [m/sec], (D) precipitation [mm/day], and (E) river discharge [mm/day]. Whereas the first two variables (temperature and pressure) follow a normal distribution, the last three variables are truncated at zero and much better described with a gamma (skewed) distribution.

Another limitation of the EM method is that it returns only the maximum likelihood values of the BMA model parameters without recourse to their underlying posterior uncertainty. This information is helpful to assess the usefulness of individual ensemble members (*Vrugt and Robinson*, 2007). A small ensemble has important computational advantages as fewer models need to be setup and executed. These imitations of the EM method motivated (*Vrugt et al.*, 2008b) to use Bayesian inference with the DREAM algorithm

for BMA model training. This multi-chain Markov chain Monte Carlo simulation method uses differential evolution as genetic algorithm for population evolution with a Metropolis selection rule to decide whether candidate points should replace their respective parents or not. This approach scales automatically the orientation and scale of the proposal distribution en route to the target distribution, and returns not only the maximum likelihood values of the BMA model parameters but also their posterior uncertainty. The use of multiple chains also offers a robust protection against premature convergence, and opens up the use of a wide arsenal of statistical measures to test whether DREAM has converged to the posterior distribution. What is more, the user does not have to adapt the log-likelihood function itself but rather instead can work with the default definition

$$\mathcal{L}(\boldsymbol{\phi}|\mathbf{D}, \tilde{\mathbf{y}}) = \sum_{t=1}^n \log \left\{ \sum_{k=1}^K \beta_k f_{tk}(\tilde{y}_t|\cdot) \right\}, \quad (21)$$

where $\boldsymbol{\phi}$ is the $d \times 1$ vector of BMA weights, variances (or proxies thereof) and/or additional shape parameters of the members' conditional forecast distributions. For the normal, lognormal, gamma and Weibull conditional distributions defined next we yield that $\boldsymbol{\phi} = \{\boldsymbol{\beta}, \boldsymbol{\sigma}\}$. The user only has to specify which parametric distribution to use for the $f_{tk}(\cdot)$'s of the ensemble. A detailed description of the DREAM_(ZS) algorithm appears in Appendix B of this manual along with a basic implementation of this algorithm in MATLAB. Interested readers are also referred to the MATLAB toolbox of DREAM Package presented in Vrugt (2016).

2.5.2. The BMA conditional distribution

The MATLAB toolbox presented herein allows the user to implement three different distributions for the conditional pdf, $f_{kt}(\cdot)$, of the ensemble members. This includes the normal, gamma, generalized normal, lognormal and Weibull distributions and allows for a proper characterization of variables with/without a skew (see Figure 4). In the next paragraphs, I will discuss the different forecast distributions.

The gamma distribution is given by

$$f_k(y|a, b) = \frac{1}{b^a \Gamma(a)} y^{(a-1)} \exp(-y/b), \quad (22)$$

where $a > 0$ and $b > 0$ are a shape and scale parameter, respectively and $f_k(y) = 0$ if $y \leq 0$. Figure 5 displays the gamma distribution for different values of the shape and scale parameter. The mean and variance of the gamma distribution are determined by the values of a and b , as follows, $\mu = ab$ and $\sigma^2 = ab^2$. Hence, the values of a and b cannot be chosen freely as their product should equate to d_{tk} and the mean of the gamma distribution centers around the actual forecast of the k^{th} member of the ensemble. I therefore follow Sloughter *et al.* (2007) and calculate the values of a and b at each time t as follows

$$a_{tk} = \frac{|d_{tk}|^2}{\sigma_k^2} \quad ; \quad b_{tk} = \frac{\sigma_k^2}{|d_{tk}|}, \quad (23)$$

and estimate the variance σ_k^2 of the gamma distribution using MCMC simulation with the DREAM_(ZS) algorithm. This implementation requires the estimation of $d = 2K$ parameters, namely the weight β_k and variance σ_k^2 of the gamma forecast distribution of each member $k = (1, \dots, K)$ of the ensemble.

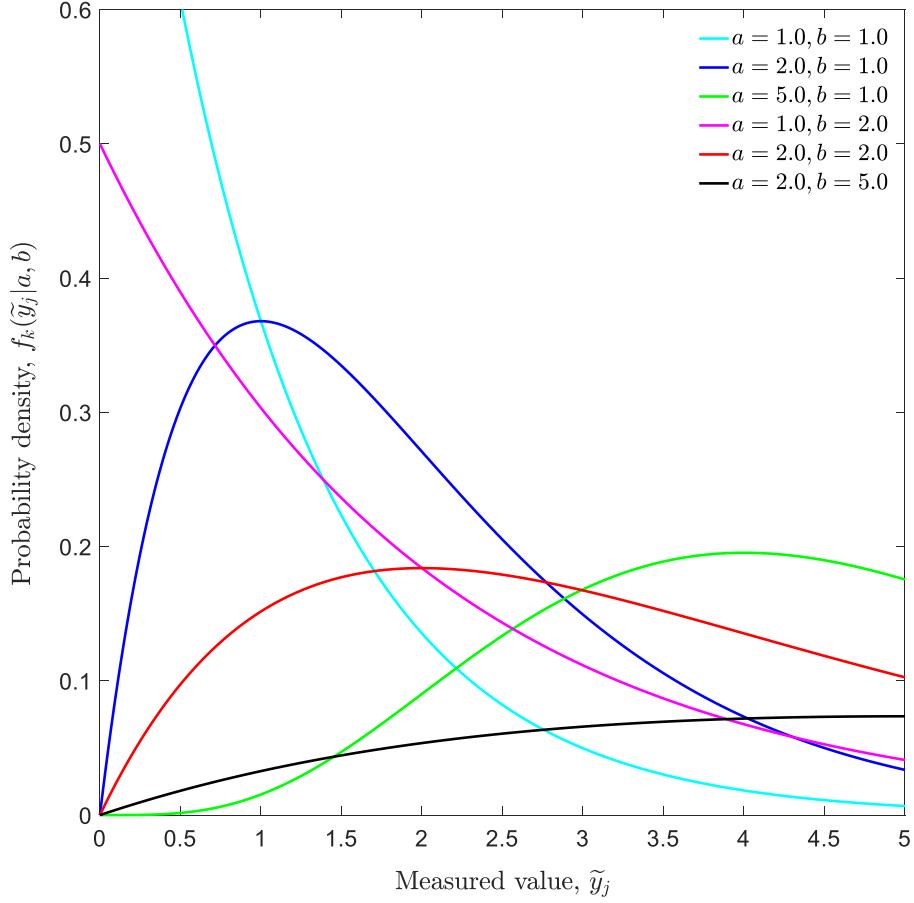


Figure 5: Probability density function, $f_k(\tilde{y}_j|a,b)$, of the gamma distribution for different values of a and b . The shape and scale parameters, a and b , respectively, change considerably the skewness and peakedness (and thus mean and standard deviation) of the distribution. In practice, the values of a and b should be defined according to Equation (23) so that the mean of the distribution is equivalent to the deterministic forecast, d_{tk} of the k^{th} ensemble member. The exponential decay function with $a = 1.0$ and $b = 1.0$ approximates the histograms of the precipitation and river discharge in Figure 4.

As third built-in option for each member's forecast density users can select the generalized normal distribution. The density of this distribution is given by

$$f_{tk}(y|d_{tk}, \sigma_k^2, \tau_k) = \frac{\tau_k}{2\sigma_k \Gamma(1/\tau_k)} \exp\left(-(|y - d_{tk}|/\sigma_k)^{\tau_k}\right), \quad (24)$$

where $\sigma_k^2 > 0$ signifies the forecast variance of the k^{th} ensemble member, and $\tau_k > 0$ is a shape parameter that determines the kurtosis of the distribution. A value of $\tau_k = 2$ results in a normal distribution (albeit with variance $\sigma^2/2$), a value of $\tau_k = 1$ equates to a Laplace (double-exponential) distribution, and $\tau_k \rightarrow \infty$ converges to a uniform density on the interval $(d_{tk} - \sigma_k, d_{tk} + \sigma_k)$ with zero density outside this range (see Figure 6). Thus, the larger the value of τ_k the more peaked the generalized normal density will be - and the more tight the associated prediction intervals around d_{tk} . The generalized normal density of Equation (24)

necessitates definition of the variance σ_k^2 and shape parameter τ_k for each member $k = (1, \dots, K)$ of the ensemble. This equates to a BMA mixture density with $d = 3K$ unknown parameter values. Alternatively, it is possible to assume a common value of τ for all ensemble members, that is, $\tau = \tau_1 = \dots = \tau_K$. This demands posterior inference of $2K + 1$ parameters with the DREAM_(zs) algorithm.

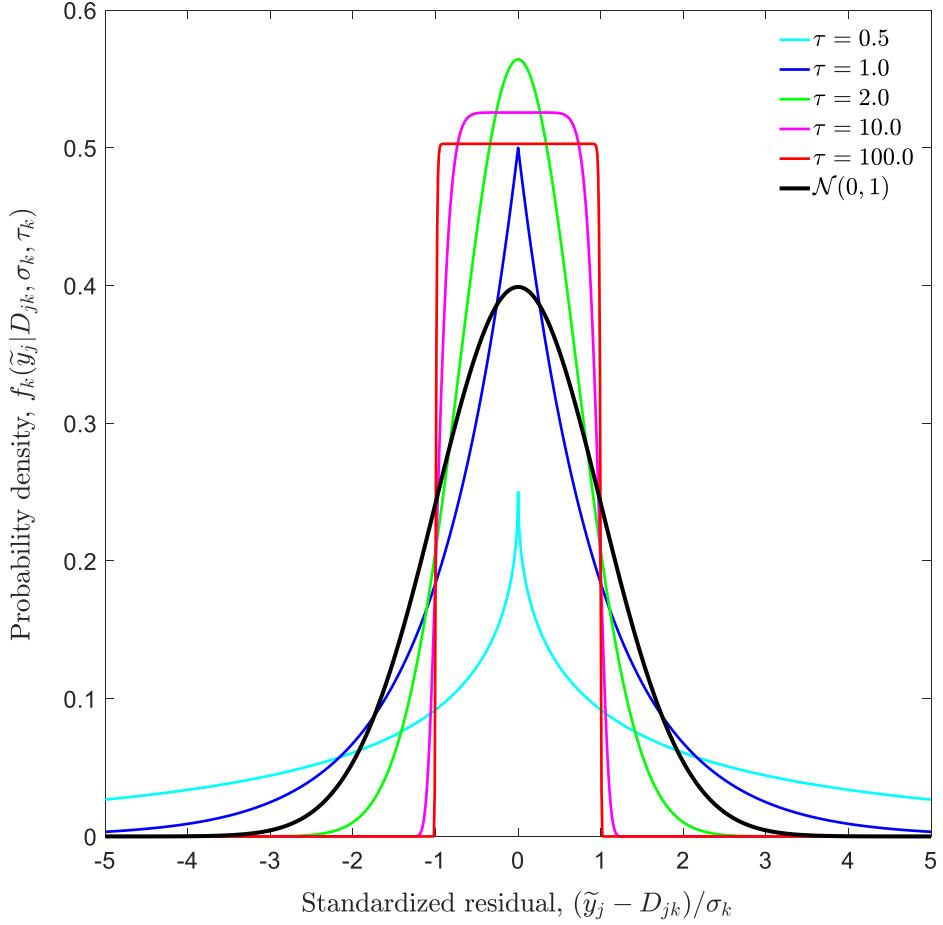


Figure 6: Probability density function, $f_{tk}(y|d_{tk}, 1, \tau_k)$, of the zero-mean generalized normal distribution with unit variance for different values of τ . The density $f_N(y|0, 1)$ of the standard normal distribution $N(0, 1)$ is separately indicated with the solid black line. The shape parameter τ changes drastically the kurtosis of the generalized normal distribution. This allows us to mimic a wide variety of symmetric densities ranging from a uniform to highly peaked (= Laplacian-type) distribution.

In V2.0 of the MODELAVG toolbox, I added several other parametric forms for the predictive distributions of the ensemble members. This includes the lognormal, truncated normal, Weibull, generalized Pareto and generalized extreme value distributions. I briefly review these distributions and discuss their implementation.

The PDF of the lognormal distribution is given by

$$f_{tk}(y|\mu_{tk}, v_{tk}^2) = \frac{1}{yv_{tk}\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{\log(y) - \mu_{tk}}{v_{tk}}\right)^2\right] \quad (25)$$

where $\mu_{tk} = \log(d_{tk}^2/(d_{tk}^2 + \sigma_k^2)^{1/2})$ centers the mean of the lognormal distribution on d_{tk} , the (bias-corrected) point forecast of the k^{th} ensemble member at time t . The variance of the lognormal distribution is equal to $v_{tk}^2 = \log(\sigma_k^2/d_{tk}^2 + 1)$. This implementation provides stable results for homoscedastic and heteroscedastic variances.

The PDF of the truncated normal distribution is given by

$$f_{tk}(y|\mu_{tk}, \sigma_k^2, a, b) = \frac{1}{\sigma_k} \frac{\phi\left(\frac{y - \mu_{tk}}{\sigma_k}\right)}{\Phi\left(\frac{b - \mu_{tk}}{\sigma_k}\right) - \Phi\left(\frac{a - \mu_{tk}}{\sigma_k}\right)}, \quad (26)$$

where $a \in \mathbb{R}$ and $b \in \mathbb{R}$ ($a < b$) define the support $a \leq y \leq b$, otherwise $f_{tk}(y|\mu_{tk}, \sigma_k^2, a, b) = 0$ and

$$\phi(\xi) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}\xi^2\right), \quad (27)$$

and

$$\Phi(\xi) = \frac{1}{2}(1 + \text{erf}(\xi/\sqrt{2})), \quad (28)$$

signify the probability density and cumulative distribution functions of the standard normal distribution, respectively. The truncated normal PDF should center on the point forecasts d_{tk} of each ensemble member, $t = (1, \dots, n)$ and $k = (1, \dots, K)$. This can be accomplished by defining μ_{tk} in Equation (26) as follows

$$\mu_{tk} = d_{tk} - \phi\left(\frac{a - \mu_{tk}}{\sigma_k}\right) \frac{\sigma_k}{\Phi\left(\frac{b - \mu_{tk}}{\sigma_k}\right) - \Phi\left(\frac{a - \mu_{tk}}{\sigma_k}\right)}. \quad (29)$$

Unfortunately, μ_{tk} , appears on both sides of the expression. This prohibits a direct solution for μ_{tk} . We instead set $\mu_{tk} = |d_{tk}|$ and affix the mode of the truncated normal distribution coincides to the prediction d_{tk} of the k^{th} model at time t . In the MODELAVG toolbox, we hardwired $a = 0$ and $b = \infty$, and, thus, limit the use of the truncated normal PDF to strictly positive variables such as streamflow, wind speed and rainfall. In subsequent releases, we can let the user determine the values of the lower and upper limits, a and b .

The PDF of the Weibull distribution is given by

$$f_{tk}(y|\lambda_{tk}, \sigma_k^2) = \frac{\sigma_k}{\lambda_{tk}} \left(\frac{y}{\lambda_{tk}}\right)^{\sigma_k-1} \exp\left[-\left(\frac{y}{\lambda_{tk}}\right)^{\sigma_k}\right]. \quad (30)$$

where λ_{tk} and σ_k^2 are so-called scale and shape parameters. The scale parameter $\lambda_{tk} = |d_{tk}|/\Gamma(1 + 1/\sigma_k)$ affixes the mean of the Weibull distribution to the point forecast d_{tk} of the k^{th} ensemble member. This implementation is not ideal; we should treat the scale parameter in a fashion similar to the variance σ_k^2 and then determine this shape parameter from the constraint that the mean of the Weibull distribution should equal the deterministic forecast, d_{tk} . This alternative or second approach is implemented in the MODELAVG toolbox but requires use of the inverse gamma function, $\Gamma^{-1}(x)$, which does not have an analytic form. There are ways to deal with this - that is for a future release of the toolbox.

The PDF of the generalized extreme value distribution is given by

$$f_{tk}(y|\mu_{tk}, \sigma_k^2, \xi_k) = \frac{1}{\sigma_k} \left[1 + \xi_k \left(\frac{y - \mu_{tk}}{\sigma_k} \right) \right]^{-1-\frac{1}{\xi_k}} \exp \left(- \left[1 + \xi_k \left(\frac{y - \mu_{tk}}{\sigma_k} \right) \right]^{-\frac{1}{\xi_k}} \right) \quad (31)$$

where μ_{tk} , σ_k^2 and ξ_k are the location (mean), scale (variance) and shape parameter. The location parameter $\mu_{tk} = d_{tk} + (1 - g_1(\xi_k))\sigma_k/\xi_k$ affixes the mean of the GEV distribution to the point forecast d_{tk} of the k^{th} ensemble member at time t , where the function $g_a(\xi) = \Gamma(1 - a\xi)$. The shape parameter ξ_k of the GEV distribution controls its tail behavior. For $\xi_k = 0$, $\xi_k > 0$ and $\xi_k < 0$ we yield the Gumbel (unbounded), Fréchet (lower tail bounded) and reversed Weibull (upper tail bounded) distributions, respectively. We limit $x_k \in [0, \frac{1}{2}]$, thus, assume that the variable of interest is lower tail bounded and that the skew cannot become negative. This avoids numerical instabilities of the GEV mixture distribution. Furthermore, an upper bound of $\xi_k = \frac{1}{2}$ guarantees an analytic expression for the variance of Equation (31), which is used to compute the coefficient of variation of the BMA mixture distribution.

The PDF of the generalized Pareto distribution is given by

$$f_{tk}(y|\mu_{tk}, \sigma_k^2, \xi_k) = \frac{1}{\sigma_k} \left[1 + \frac{\xi_k(y - \mu_{tk})}{\sigma_k} \right]^{-1-\frac{1}{\xi_k}} \quad (32)$$

where μ_{tk} , σ_k^2 and ξ_k are the location (mean), scale (variance) and shape parameter. The location parameter $\mu_{tk} = d_{tk} - \sigma_k/(1 - \xi_k)$ affixes the mean of the generalized Pareto distribution to the point forecast d_{tk} of the k^{th} ensemble member at time t . The shape parameter $\xi \in (-\infty, \frac{1}{2})$ of the generalized Pareto distribution controls its tail behavior. If ξ_k and μ_k are both zero, the generalized Pareto distribution simplifies to the exponential distribution. If $\xi_k > 0$ and $\mu_k = \sigma_k/\xi_k$, then the generalized Pareto distribution is equal to the Pareto distribution. We limit $x_k \in [-1, \frac{1}{2}]$. An upper bound of $\xi_k = \frac{1}{2}$ guarantees an analytic expression for the variance of Equation (31), which is used to compute the coefficient of variation of the BMA mixture distribution.

The generalized extreme value and Pareto distributions of Equations (31) and (32) demand definition of the variance σ_k^2 and shape parameter ξ_k for each member $k = (1, \dots, K)$ of the ensemble. This equates to a BMA mixture density with $d = 3K$ unknown parameter values. Alternatively, it is possible to assume a common value of ξ for all ensemble members, that is, $\xi = \xi_1 = \dots = \xi_K$. This involves posterior estimation of $2K + 1$ parameters with the DREAM_(ZS) algorithm.

The variances $\sigma_k^2; k = (1, \dots, K)$ of the lognormal, truncated normal, Weibull, generalized extreme value and generalized Pareto distributions are estimated jointly with the BMA weights $\beta_k; k = (1, \dots, K)$ of the ensemble members using MCMC simulation with the DREAM_(ZS) algorithm.

Thus far, I have made three important assumptions, (a) each member of the ensemble has the same parametric form of the forecast distribution (normal, lognormal, truncated normal, generalized normal, gamma or Weibull), (b) the variance of this distribution differs among the members of the ensemble, and (c) the variance is invariant (= constant with time). The first assumption will not be contested, as the shape of the forecast distribution is determined by the frequency distribution (histogram) of the target variable, yet the second and third assumption may be too restrictive. The MATLAB toolbox of MODELAVG therefore implements four different "models" for the variances of the forecast distributions.

- (1) common constant variance: all members of the ensemble have the same variance, that is $\sigma_1^2 = \dots = \sigma_K^2$. This simplifies somewhat the inference and involves $d = K + 1$ fitting parameters for the normal, lognormal, truncated normal, gamma and Weibull forecast distributions, and $d = K + 2$ (common τ for all ensemble members) or $d = 2K + 1$ (individual τ for each ensemble member) parameters for the generalized normal conditional pdf.
- (2) individual constant variance: all members of the ensemble have their own specific variance, and thus $\boldsymbol{\sigma}^2 = \{\sigma_1^2, \dots, \sigma_K^2\}$. This assumption was made in our notation thus far and results in a BMA mixture distribution with $d = 2K$ unknowns for the normal, lognormal, truncated normal, gamma and Weibull conditional pdf's. This equates to $d = 2K + 1$ (common τ) or $d = 3K$ (individual τ 's) parameters for the generalized normal conditional pdf.
- (3) common non-constant variance: the variances of the forecast distributions are dependent on the magnitude of the forecast. This approach is implemented using $\sigma_{tk}^2 = (cd_{tk})^2$, where the multiplier c holds for all models and forecasts of the ensemble. This approach involves an equal number of unknowns d as defined for each conditional PDF under (1) above.
- (4) individual non-constant variance: the variances of the forecast distributions are member and forecast dependent. This approach is implemented using $\sigma_{tk}^2 = (c_k d_{tk})^2$, where the K multipliers are subject to inference. The number of unknowns for the normal, lognormal, truncated normal, Weibull, gamma and generalized normal forecast distributions is equal to defined for (2) above.

The first two approaches assume a homoscedastic (constant) variance of the conditional distribution of each ensemble member. This approach may be appropriate for variables such as the temperature and pressure of the atmosphere that are known to have a constant measurement error (*Vrugt et al.*, 2006). The last two approaches assume a heteroscedastic (non-constant) variance of the conditional forecast distribution. The variance of this distribution increases with value of the forecast. This assumption is appropriate for variables such as rainfall (*Sloughter et al.*, 2007), discharge (*Vrugt and Robinson*, 2007) and wind speed (*Sloughter et al.*, 2010) whose measurement errors are known to increase with magnitude of the observation. Note, it is rather easy to formulate other models for the variance of the forecast distribution, for instance, one can augment the heteroscedastic variance with a constant value and write $\sigma_{tk}^2 = (c_k d_{tk} + c_0)^2$ instead so that the variance does not become zero if $d_{tk} = 0$. The unknown coefficient c_0 can augment the vector $\boldsymbol{\phi}$ of BMA parameters and their values are estimated using MCMC simulation with the DREAM_(ZS) algorithm.

The MATLAB function BMA_CALC calculates the value of the log-likelihood, $\mathcal{L}(\boldsymbol{\phi}|\mathbf{D}, \tilde{\mathbf{y}})$ of the BMA model in Equation (21) for the $d \times 1$ vector $\boldsymbol{\phi}$ of BMA weights, variances (or proxies thereof) and/or shape parameters of the ensemble members' conditional PDFs, $n \times K$ matrix of (bias-corrected) ensemble forecasts \mathbf{D} and $n \times 1$ vector of training data observations $\tilde{\mathbf{y}}$. The fifth and last input argument, options, stores the desired forecast distribution, and is defined by the user in the input file of the MODELAVG toolbox (see section 3).

MATLAB code of `BMA_calc`: This function computes the value of the log-likelihood (return argument) for a vector ϕ with BMA weights, standard deviations (or proxies thereof) and/or shape parameters (for generalized normal distribution) of the members' conditional PDFs. The second and third input arguments, D and y , are the $n \times K$ matrix of ensemble forecasts and $n \times 1$ vector of verifying observations, respectively. The last input argument `options` is a structure that conveys the choice of forecast distribution and its variance treatment. Notation is consistent with main text. Built-in functions are highlighted with a low dash. The fields `PDF`, `VAR` and `TAU` of `options` specify the conditional PDF, variance option and treatment of τ parameter, respectively. The `switch` function cycles among the cases listed in the code. The built-in function `pdf` computes the density at the observed values, y , according to each member's conditional PDF. `lik` is a $n \times 1$ vector with likelihoods of the BMA mixture density and `loglik` returns the sum of the natural logarithm of the n likelihood values.

Thus, the code first computes the value of the standard deviation for each forecast and member of the ensemble. This results in the $n \times K$ matrix `sigma` which has the same numbers of rows and columns as matrix `D` with ensemble forecasts. Then, the likelihood of the BMA mixture distribution is evaluated at each observation of the training data set by taking the sum of the weighted densities of the K different conditional distributions evaluated at \tilde{y} . Then, the log-likelihood of the BMA model is computed by taking the sum of the natural log values of the n different densities. It is not too difficult to add other conditional PDFs to the `BMA_calc` function. But this does require making changes to one other function of the MODELAVG toolbox as well. As final remark, the built-in `pdf` function is not always CPU efficient. The MODELAVG toolbox implements more efficient procedures to evaluating the densities of the different conditional forecast PDFs.

Lower and upper endpoints of the $100\alpha\%$ prediction intervals of the BMA mixture density can be derived from its CDF in Equation (14) so that $G_t(l) = (1 - \alpha)/2$ and $G_t(u) = (1 + \alpha)/2$. For each observation of the training data set, we solve for the lower and upper predictive quantiles of each respective α value using an iterative root finding method. The MATLAB function `BMA_quantile` computes the prediction quantiles of the BMA model and returns as well $n \times 1$ vectors with the CDF and PDF of the BMA mixture distribution evaluated at the observations of the training data record.

MATLAB code of BMA_quantile: This function returns the quantiles of the prediction intervals of the BMA mixture distribution (return argument) using as input arguments the vector phi of BMA weights, standard deviations (or proxies thereof) and/or shape parameters (for generalized normal distribution) of the members' conditional PDFs, the $n \times K$ matrix D of ensemble forecasts, $n \times 1$ vector y of verifying observations and structure options with choice of forecast distribution and its variance treatment. Built-in functions are highlighted with a low dash. To shorten the length of the code, I only present the implementation for the normal, lognormal, gamma and Weibull conditional PDFs. pred is a $n \times 2n_\alpha$ matrix with prediction quantiles of the BMA mixture distribution. pdf_y and cdf_y are $n \times 1$ vectors with the PDF and CDF of the BMA mixture distribution evaluated at the verifying observations of the training data record.

```

function [ pred , pdf_y , cdf_y ] = BMA_quantile ( phi, D , y , options)
% MODELAVG toolbox, V2.0: Function computes quantiles, PDF and CDF at y of BMA mixture

[n,K] = size(D);                                % Number of forecasts and number of ensemble members
beta = phi(1:K);                                % Unpack weights of each member's conditional PDF
switch options.VAR      % VARIANCE OPTION
    case '1', sigma = phi(K+1) * ones(n,K);          % 1: common constant variance
    case '2', sigma = repmat(phi(K+1:2*K),n,1);        % 2: sole constant variance
    case '3', c = phi(K+1); sigma = c * D;            % 3: common nonconstant variance
    case '4', c = phi(K+1:2*K); sigma = bsxfun(@times,c,D); % 4: sole nonconstant variance
    otherwise, error('MODELAVG:BMA_quantile: Unknown variance option');
end
sigma = max(sigma,eps);                          % sigma > 2.22e-16
[pdf_y,cdf_y] = deal(nan(n,1));                % BMA PDF/CDF at y
alfa = sort([(1-options.alpha)/2,(1-options.alpha)/2+options.alpha]); % Quantiles
na = numel(alfa); pred = nan(n,na);             % Init. pred qnts
switch options.PDF      % CONDITIONAL DISTRIBUTION
    case 'normal', A = D; B = sigma; f = 'normcdf';
    case 'lognormal', S2 = log(sigma.^2./D.^2 + 1); A = log(abs(D)) - S2/2;
        B = sqrt(S2); f = 'logncdf';
    case 'gamma', A = (abs(D)./sigma).^2; B = sigma.^2./abs(D); f = 'gamcdf';
    case 'weibull', A = abs(D)./.gamma(1+1./sigma); B = sigma; f = 'wblcdf';
end
for k = 1:K, nk = num2str(k);                  % Write out BMA mixture CDF as a string
    switch k
        case 1, Fx = ['Fx = @(x,t,alfa) beta('nk,')*',f,'(x,A(t,1),B(t,1))'];
        otherwise, Fx = [Fx,' + beta('nk,')*',f,'(x,A(t,',nk,'),B(t,',nk,'))'];
    end
end
Fx = strcat(Fx,'- alfa;'); fx = Fx; eval(Fx);           % CDF function handle
fx = strrep(strrep(fx,'cdf','pdf'),'Fx','fx'); eval(fx); % PDF function handle
for t = 1:n
    pdf_y(t) = fx(y(t),t,0); cdf_y(t) = Fx(y(t),t,0); % PDF/CDF at y(t)
    for z = 1:na                                % Prediction quantiles using root finding
        pred(t,z) = fzero(@(x) Fx(x,t,alfa(z)),abs(mean(D(t,1:K))));
    end
end

```

In the latest release of the MODELAVG toolbox (V2.0) the user can admit different confidence levels for α .

Then, in V2.0 of the toolbox, I added various new functionalities related to the evaluation of the BMA mixture distribution. Besides all the original performance metrics that were returned in structure `output` the toolbox now also computes scoring rules of the BMA mixture distribution. This includes performance metrics such as the reliability (from PIT) and coefficient of variation, and *strictly proper* quadratic, logarithmic, spherical and continuous ranked probability scoring rules of the BMA distribution forecast and of Equation (13). We also return the 1-norm and 2-norm of the BMA mixture PDF. The scoring rules and norms are computed at each $t = (1, \dots, n)$ and returned to the user in fields `QS`, `LS`, `SS`, `CRPS` and `mix_norm` of structure array `output`, details of which later. *Vrugt et al.* (2006) used the CRPS in multi-criteria calibration of the BMA model, whereas *Vrugt* (2023) demonstrates the application of scoring rules to BMA model training.

2.6. Mallows model averaging

Mallows model averaging (MMA) is a Frequentist solution to the problem of model averaging. The MMA method uses the following penalized sum of squared residuals objective function

$$C_n(\boldsymbol{\beta} | \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}_\epsilon^2, \mathbf{p}) = \sum_{t=1}^n (\tilde{y}_t - \mathbf{d}_t \boldsymbol{\beta})^2 + 2\hat{\sigma}_\epsilon^2 \sum_{k=1}^K \beta_k p_k \quad (33)$$

where, as before, p_k is the number of "free" parameters of the k^{th} model of the ensemble and $\hat{\sigma}_\epsilon^2$ is an estimate of the residual variance of the ϵ_t 's in Equation (1). This value is often set equivalent to the variance of the forecast error of the most complex (= parameter rich) model of the ensemble.

We can now find the optimal values of the MMA weights by minimizing Mallows' criterion in Equation (33) or

$$\hat{\boldsymbol{\beta}}_{\text{MMA}} = \arg \min_{\boldsymbol{\beta} \in \mathbb{R}^K} C_n(\boldsymbol{\beta} | \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}^2, \mathbf{p}), \quad (34)$$

where the weights are allowed to vary freely in \mathbb{R}^K and are thus not restricted to the unit simplex \mathcal{S}^K . The value of $\boldsymbol{\beta}_{\text{MMA}}$ can also be estimated by maximizing the following log-likelihood function

$$\mathcal{L}(\boldsymbol{\beta} | \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}^2, \mathbf{p}) \simeq -\frac{1}{2} C_n(\boldsymbol{\beta} | \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}^2, \mathbf{p}), \quad (35)$$

using MCMC simulation with the DREAM_(ZS) (*Diks and Vrugt*, 2010; *Vrugt et al.*, 2008b, 2009; *Vrugt*, 2016) algorithm. Indeed, the maximum likelihood of the MMA weights is simply equivalent to the sample of DREAM_(ZS) with largest value of Equation (35). This sample is easy to find in MATLAB using the built-in `max` operator. The theory and MATLAB implementation of the DREAM_(ZS) algorithm is presented in Appendix B. A separate toolbox of this algorithm is available in MATLAB and described in detail by *Vrugt* (2016).

If so desired, we can also restrict the MMA weights to lie on the unit simplex, \mathcal{S}^K , in \mathbb{R}_+^K and thus to be positive and add up to one, $\beta_k \geq 0; \sum_{k=1}^K \beta_k = 1$. This alternative model averaging method is hereafter conveniently referred to as MMA ^{\mathcal{S}} .

The MATLAB function `MMA_calc` listed below calculates the log-likelihood of Equation (35) for a given input vector of weights, `beta`.

MATLAB code of `MMA_calc`: This function calculates the log-likelihood of the MMA deterministic point forecasts using as input arguments, the weights \mathbf{x} , ensemble forecasts, \mathbf{D} , verifying observations, \mathbf{y} , number of "free" parameters of each model of the ensemble, p , and variance of the forecast error, `var_err` of the most complex model of the ensemble. Notation is consistent with Equation (35) in main text. Matrix algebra is used to minimize the CPU-time. Built-in functions are highlighted with a low dash. The function `sum()` calculates the sum of the squared differences between the MMA point forecast and the verifying observations, and `size(D)` returns the number of rows and columns of the matrix \mathbf{D} .

```

function loglik = MMA_calc ( beta , D , y , var_err , p )
    % This function calculates the log likelihood of MMA
    % Function of MODELAVG toolbox, V2.0

    % B.C. Hansen, "Least Squares Model Averaging", Econometrica, vol. 75,
    % no. 4, pp. 1175-1189, 2007

    if nargin < 5
        error('MODELAVG:MMA_calc: Requires at least five input arguments.');
    end

    G = D*beta';                                % MMA deterministic point forecast
    Cn = sum((y - G).^2) + 2*var_err*beta*p;    % Mallows criterion: ref Equation (11)
    loglik = -1/2 * Cn + realmin;                % Log-likelihood of x = { MMA weights }

```

This concludes the theory of the different model averaging methods. I now describe the implementation of the theory and codes described above in the MODELAVG toolbox in MATLAB.

3. The MODELAVG toolbox

The MODELAVG toolbox implements each of the model averaging methods described in section 2 in MATLAB and returns to the user the values of the weights, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$ and members' standard deviation(s) (or proxies thereof) of the conditional distribution, $f_k(\cdot)$ (if BMA is used). You can download the MODELAVG toolbox from my website at the following link <http://faculty.sites.uci.edu/jasper/software/> (scroll down to appropriate link). Appendix C explains how to setup the MODELAVG toolbox in MATLAB.

3.1. MODELAVG: MATLAB implementation

The MODELAVG toolbox can be executed from the MATLAB prompt by typing the following statement in the command window

$$[\phi, \text{output}] = \text{MODELAVG}(\text{method}, \mathbf{D}, \mathbf{y}, \text{options}) \quad (36)$$

where `method` (string), \mathbf{D} ($n \times K$ matrix), \mathbf{y} ($n \times 1$ vector) and `options` (structure array) are input arguments defined by the user, and `phi` (vector) and `output` (structure array) are output arguments that are computed by the function MODELAVG and returned to the user. To minimize the number of input

Table 1: Acronym used by input argument method for each of the model averaging methods of section 2. The last column summarizes for each method whether the weights are restricted to the unit simplex, \mathcal{S}^K , or not.

Content method	Description	\mathcal{S}^K
'EWA'	Equal weight averaging	Yes
'BGA'	Bates-Granger averaging	Yes
'AICA'	Akaike information criterion averaging	Yes
'BICA'	Bayes information criterion averaging	Yes
'GRA'	Granger-Ramanathan averaging	No
'BMA'	Bayesian model averaging	Yes
'MMA'	Mallows' model averaging	No
'MMA-S'	Mallows' model averaging	Yes

and output arguments of MODELAVG, the structure options and output group related variables using data containers called fields, more of which later. The structure options is an optional input argument required only for information criterion averaging, Bayesian model averaging and Mallows model averaging.

A summary of the different functions of the MODELAVG toolbox appears in Appendix D. I will now discuss each of the input and output variables.

3.2. First input argument: *method*

The variable *method* defines with a string enclosed between quotes the name of the model averaging method that will be used by the function MODELAVG. The user can select among the eight different methods discussed of section 2 using the acronym (in quotes) listed in the first column of Table 1.

The function MODELAVG is case insensitive; thus, the user can input to *method* lower case (small) and upper case (capital) letters of the acronyms listed in Table 1. Almost all methods restrict the weights to the unit simplex, except Granger-Ramanathan averaging (*method* = 'GRA') and variant MMA^S of Mallows Model Averaging (*method* = 'MMA-S').

The first five model averaging methods listed in Table 1 ('EWA', 'BGA', 'AICA', 'BICA' and 'GRA') will execute rapidly as they have a direct closed-form solution for their weights. The last three methods, 'BMA', 'MMA', and 'MMA-S', require an iterative solution with DREAM_(zs) to locate the maximum likelihood values of their weights. If BMA is used, then the DREAM_(zs) algorithm returns as well estimates of the standard deviation of the members' conditional distribution, $f_k(\cdot); k = (1, \dots, K)$.

3.3. Second input argument: *D*

The second input argument *D* of the function MODELAVG is a $n \times K$ matrix with n forecasts of each member of the ensemble. This input argument is thus equivalent to **D** and uses a separate column for each of the K models of the ensemble. Bias correction is recommended for each ensemble member particularly for model averaging methods that restrict the weights to the unit simplex. The MODELAVG toolbox has a built-in utility for linear bias correction (see Equation (2)), yet more advanced bias-correction methods can be devised by the user - more of which later.

3.4. Third input argument: *y*

The third input argument *y* of the MODELAVG function stores the training data set, \tilde{y} with verifying observations. This $n \times 1$ vector is used to determine the weights of each model averaging method. If BMA is used, then this also includes estimates of the variance(s) of the members' forecast distribution. The number of rows of *y* should match exactly the number of rows of the forecast ensemble stored in *D* (second input argument of function MODELAVG).

3.5. Fourth input argument: *options*

The fourth and last input argument of the function MODELAVG is the structure *options*. This input argument is optional and used only by information criterion averaging (AICA or BICA), BMA, MMA, and MMA^S although the field *print* applies to all methods. Table 2 summarizes the different fields of structure options and their content.

Table 2: Overview of the different fields of input argument *options*

Field	Description	Options	Type	method
PDF	Forecast distribution	'normal' 'lognormal' 'gen_normal' 'tnormal' 'gamma' 'weibull' 'gev' 'gpareto'	string	BMA
VAR	Variance option	'1'/'2'/'3'/'4'	string	BMA
TAU	Shape parameter option	'1'/'2'	string	BMA
alpha	Prediction interval	e.g. 0.90/0.95/0.99	real	BMA [†]
p	Model complexity	> 0	K-vector	AICA/BICA/MMA/MMA ^S
print	Screen output	'yes' or 'no'	string	All

[†] The field *alpha* contains the desired confidence intervals; this can be a single value or a vector of confidence levels. The field *pred* of structure *output* will return the corresponding BMA prediction quantiles for each *t* (listed from low to high).

The first four fields PDF, VAR, TAU and alpha of structure options are necessary and/or required input for the BMA method. The field PDF lists with a string enclosed between quotes the name of the conditional distribution, $f_k(\cdot)$ that is used for the $k = (1, \dots, K)$ ensemble members. Built-in options include PDF = 'normal', PDF = 'lognormal', PDF = 'gen_normal', PDF = 'tnormal', PDF = 'gamma', PDF = 'weibull', PDF = 'gev', PDF = 'gpareto' for a normal, lognormal, generalized normal, truncated normal (strictly positive variable), gamma, Weibull, generalized extreme value (positive skew) and generalized Pareto conditional forecast distribution, respectively. The 'lognormal', 'tnormal', 'weibull', 'gev' and 'gpareto' predictive PDFs have been added in V2.0 of the MODELAVG toolbox.

The field VAR of structure options allows the user to select with a string (between quotes) the variance of the conditional distribution. The user can select among four different options,

1. (A) VAR = '1' : common constant variance; $\sigma_1^2 = \dots = \sigma_K^2$.

2. (B) $\text{VAR} = '2'$: individual constant variance; $\sigma^2 = \{\sigma_1^2, \dots, \sigma_K^2\}$.
3. (C) $\text{VAR} = '3'$: common non-constant variance; $\sigma_{tk}^2 = (c \cdot D_{tk})^2$.
4. (D) $\text{VAR} = '4'$: individual non-constant variance; $\sigma_{tk}^2 = (c_k D_{tk})^2$.

These options allow for a homoscedastic and heteroscedastic variance of the conditional distribution of the ensemble members, and have been discussed in section 2.5.2 of this manual.

The field **TAU** is required input for the generalized normal, generalized extreme value and generalized Pareto distributions. Then, **TAU** = '1' would assume use of a common value of τ for each ensemble member, and **TAU** = '2' uses a different τ value for each ensemble member.

The field **alpha** of structure options defines the desired prediction interval(s) of the BMA model. The latest release allows users to specify more than one value for $\alpha \in (0, 1)$. The BMA prediction quantiles are returned to the user in the field **pred** of the structure output of MODELAVG (see next section). Typical values of **alpha** are 0.90, 0.95 or 0.99 for a 90, 95 or 99% prediction interval of the BMA mixture distribution, respectively. If the user does not specify the field **alpha** or leaves empty its content then the toolbox will assume a default value of $\alpha = 0.95$. In that case, the field **pred** of structure output would consists of only two columns with the BMA mixture CDF at $(1 - \alpha)/2$ and $(1 + \alpha)/2$, respectively. Jointly, these two quantiles define the $100\alpha\%$ prediction interval.

The field **p** of structure options stores the number of parameters for each model of the ensemble. Thus, **p** should contain K values, and have dimensions $1 \times K$ (horizontal vector) or $K \times 1$ (vertical vector), respectively. This field is required input for information criterion averaging (AICA and BICA), MMA, and MMA^S.

Finally, the field **print** of structure options controls the output writing of the MODELAVG toolbox. If the content of **print** equates to 'yes' then the toolbox will visualize, in many different figures, the output of each model averaging method. This output writing is suppressed if field **print** is set to 'no'.

The user can employ upper case and lower case letters for each of the fields of structure options. The same holds for the content of each field of options. Thus, **options.pdf = 'NORMAL'** is equally valid as **options.PDF = 'normal'**.

3.6. Output arguments

The function MODELAVG returns to the user two output arguments, **phi** (vector or matrix) and **output** (structure array). The content of these two output arguments differs per model averaging method and their respective settings.

3.6.1. Return argument **phi**

The content of **phi** depends on the model averaging method that is being used (see Table 3).

MODELAVG MANUAL

Table 3: Content and number of rows and columns of output argument `phi` of the MODELAVG toolbox

method	VAR	TAU	Content of phi	Size of phi
'EWA'			$(\beta_1, \dots, \beta_K)^\top$	$1 \times K$
'BGA'			$(\beta_1, \dots, \beta_K)^\top$	$1 \times K$
'AICA'			$(\beta_1, \dots, \beta_K)^\top$	$1 \times K$
'BICA'			$(\beta_1, \dots, \beta_K)^\top$	$1 \times K$
'GRA'			$(\beta_1, \dots, \beta_K)^\top$	$1 \times K$
'MMA'			$M \times \{\beta_1, \dots, \beta_K, \mathcal{L}(\beta \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}^2, \mathbf{p})\}$	$M \times (K + 1)$
'MMA-S'			$M \times \{\beta_1, \dots, \beta_K, \mathcal{L}(\beta \mathbf{D}, \tilde{\mathbf{y}}, \hat{\sigma}^2, \mathbf{p})\}$	$M \times (K + 1)$
'BMA'	'1'	-	$M \times \{\beta_1, \dots, \beta_K, \sigma, \mathcal{L}(\beta, \sigma \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (K + 2)$
'BMA'	'2'	-	$M \times \{\beta_1, \dots, \beta_K, \sigma_1, \dots, \sigma_K, \mathcal{L}(\beta, \sigma \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 1)$
'BMA'	'3'	-	$M \times \{\beta_1, \dots, \beta_K, c, \mathcal{L}(\beta, c \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (K + 2)$
'BMA'	'4'	-	$M \times \{\beta_1, \dots, \beta_K, c_1, \dots, c_K, \mathcal{L}(\beta, c \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 1)$
'BMA'	'1'	'1'	$M \times \{\beta_1, \dots, \beta_K, \sigma, \tau, \mathcal{L}(\beta, \sigma, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (K + 3)$
'BMA'	'2'	'1'	$M \times \{\beta_1, \dots, \beta_K, \sigma_1, \dots, \sigma_K, \tau, \mathcal{L}(\beta, \sigma, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 2)$
'BMA'	'3'	'1'	$M \times \{\beta_1, \dots, \beta_K, c, \tau, \mathcal{L}(\beta, c, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (K + 3)$
'BMA'	'4'	'1'	$M \times \{\beta_1, \dots, \beta_K, c_1, \dots, c_K, \tau, \mathcal{L}(\beta, c, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 2)$
'BMA'	'1'	'2'	$M \times \{\beta_1, \dots, \beta_K, \sigma, \tau_1, \dots, \tau_K, \mathcal{L}(\beta, \sigma, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 2)$
'BMA'	'2'	'2'	$M \times \{\beta_1, \dots, \beta_K, \sigma_1, \dots, \sigma_K, \tau_1, \dots, \tau_K, \mathcal{L}(\beta, \sigma, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (3K + 1)$
'BMA'	'3'	'2'	$M \times \{\beta_1, \dots, \beta_K, c, \tau_1, \dots, \tau_K, \mathcal{L}(\beta, c, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (2K + 2)$
'BMA'	'4'	'2'	$M \times \{\beta_1, \dots, \beta_K, c_1, \dots, c_K, \tau_1, \dots, \tau_K, \mathcal{L}(\beta, c, \tau \mathbf{D}, \tilde{\mathbf{y}})\}$	$M \times (3K + 1)$

For EWA, BGA, AICA, BICA and GRA, the output argument `phi` is equivalent to a horizontal vector with K values of the weights, $(\beta_1, \dots, \beta_K)^\top$. The entries of `phi` correspond to the different columns of input argument \mathbf{D} . Thus, entry k of `phi` stores the weight of the k^{th} column (ensemble member) of matrix \mathbf{D} .

If MMA or MMA^S are used then output argument `phi` consist of a $M \times (K + 1)$ matrix with M posterior samples of the MMA or MMA^S weights and their corresponding values of the log-likelihood of Equation (35). These M solutions summarize the posterior of the MMA or MMA^S weights and can be used (among others) to analyze the uncertainty of the MMA weights and point forecasts.

The maximum likelihood values of the MMA and MMA^S weights, $\hat{\beta}_{\text{MMA}}$ or $\hat{\beta}_{\text{MMA}^S}$ are stored separately in the field `ML` of output argument `output` (more of which later). Their values can also be derived from `phi` by locating the row number of this matrix with largest value of the log-likelihood of Equation (35). This row of `phi` can be located with the MATLAB command

$$[\text{na} , \text{row_max}] = \max(\text{phi}(:, \text{K}+1)) \quad (37)$$

and thus the maximum likelihood values of the MMA or MMA^S weights are equivalent to

$$\text{beta_MMA} = \text{phi}(\text{row_max}(1), 1:\text{K}) \quad (38)$$

Finally, if BMA is used for postprocessing of the forecast ensemble, then output argument `phi` is equivalent to a $M \times (d + 1)$ matrix with M posterior samples of the d parameters of the BMA mixture distribution, and corresponding value of the log-likelihood of Equation (21). The first K columns of `phi` list the values of the BMA weights of each member of the ensemble. The content of columns $K + 1$ to d of `phi` depends on

the assumed properties of the members' forecast distribution defined by the user in field VAR and/or field AU of structure options. For instance, if `options.VAR = '1'` then all K members of the ensemble are assumed to have the same standard deviation of their forecast distribution, and column $K + 1$ of `phi` lists the posterior values of σ . If `options.VAR = '2'` then each members' forecast distribution has a different standard deviation, and columns $K + 1$ to $2K$ store the posterior values of $\{\sigma_1, \dots, \sigma_K\}$. If the user has selected the generalized normal distribution as conditional pdf of each ensemble member, then the content of `phi` depends not only on the variance option (homoscedastic or heteroscedastic) but also on the field TAU of structure options. For example, if `TAU = '2'` then each member will have its own τ value, otherwise with `TAU = '1'` a single common value of τ will be used for all members. Table 3 summarizes the content of each column of `phi` for the different settings of fields VAR and TAU of structure options.

The maximum likelihood values of the parameters of the BMA mixture distribution are stored in field ML of structure options. The user can also derive these values from output argument `phi` by locating the (posterior) sample of this matrix with largest value of the log-likelihood. The recipe of how to do this in MATLAB was given in Equations (37) and (38).

3.6.2. Return argument output

The second output argument of the function MODELAVG is a structure array called `output` and stores important information about the performance of each model averaging method, and (if appropriate) convergence properties of the DREAM_(zs) algorithm. Most of the fields of structure `output` are only defined if MMA, MMA^S or BMA are used (see Table 4).

MODELAVG MANUAL

Table 4: Content of the return argument `output` of the MODELAVG function

Field output	Description	Content
<code>Ye</code>	Deterministic point forecast, Equation (4)	$n \times 1$ vector
<code>RMSE</code>	Root mean square error point forecast	scalar
<code>R</code>	Pearson correlation coefficient of point forecasts and training data	scalar
<code>RMSE_mod</code>	Root mean square errors ensemble members	$1 \times K$ vector
<code>RunTime</code>	Elapsed time	scalar
<code>str_table</code>	Cell array with names of the estimated parameters	$1 \times d$ cell array
MMA and MMA ^S methods		
<code>ML</code>	Maximum likelihood values of weights	$1 \times K$ vector
<code>loglik</code>	Maximum value of log-likelihood Equation (35)	scalar
<code>std</code>	Posterior standard deviation weights	$1 \times K$ vector
<code>corr</code>	Posterior correlation coefficients of weights	$K \times K$ matrix
BMA method		
<code>ML</code>	Maximum likelihood BMA model parameters	$1 \times d$ vector
<code>loglik</code>	Maximum value of log-likelihood Equation (21)	scalar
<code>corr</code>	Posterior correlation coefficients BMA parameters	$d \times d$ matrix
<code>std</code>	Posterior standard deviation BMA parameters	$1 \times d$ vector
<code>pred</code>	Lower and upper prediction intervals BMA mixture distribution	$n \times 2n_\alpha$ matrix
<code>coverage</code>	Coverage of n_α prediction intervals	$1 \times n_\alpha$ vector
<code>spread</code>	Average width of n_α prediction intervals	$1 \times n_\alpha$ vector
<code>par_unc</code>	BMA prediction quantiles due to parameter uncertainty only	$n \times 2n_\alpha$ matrix
<code>pdf_y</code>	PDF of BMA mixture distribution for each training observation	$n \times 1$ vector
<code>cdf_y</code>	CDF of BMA mixture distribution for each training observation	$n \times 1$ vector
<code>CRPS</code>	Continuous ranked probability score for each training observation	$n \times 1$ vector
<code>QS</code>	Quadratic score for each training observation	$n \times 1$ vector
<code>LS</code>	Logarithmic score for each training observation	$n \times 1$ vector
<code>SS</code>	Spherical score for each training observation	$n \times 1$ vector
<code>mRLBL_anal</code>	Reliability of BMA distribution forecasts	scalar
<code>mCv_anal</code>	Time-averaged coefficient of variation BMA mixture [= exact]	scalar
<code>mQS</code>	Time-averaged quadratic score BMA forecast distribution	scalar
<code>mLS</code>	Time-averaged logarithmic score BMA forecast distribution	scalar
<code>mSS</code>	Time-averaged spherical score BMA forecast distribution	scalar
<code>mCRPS</code>	Time-averaged continuous ranked probability score	scalar
<code>mix_norm</code>	1-norm and 2-norm of BMA mixture PDF	$n \times 2$ matrix
<code>mu_mix</code>	Mean of BMA mixture forecast [= exact]	$n \times 1$ vector
<code>var_mix</code>	Variance of BMA mixture forecast [= exact]	$n \times 1$ vector
<code>KGE</code>	Kling-Gupta efficiency of weighted-average BMA forecast	scalar
BMA, MMA and MMA ^S (DREAM _(ZS) diagnostics)		
<code>MR_stat</code>	Multivariate scale-reduction factor	$N \times 2$ matrix
<code>R_stat</code>	Univariate scale-reduction factor	$N \times (d + 1)$ matrix
<code>AR</code>	Acceptance rate (%) of proposals	$N \times 2$ matrix

The field `Ye` of `output` stores the weighted forecast of each model averaging method. This $n \times 1$ vector is derived from Equation (4) using the (maximum likelihood) weights of each averaging method and ensemble forecasts of input argument `D`. The fields `RMSE` and `R` of `output` (both scalars) summarize the performance

of the averaged forecast using the root mean square error and Pearson's correlation coefficient, respectively. The field `RMSE_mod` is a $1 \times K$ vector with RMSEs of the individual members of the ensemble, and the field `RunTime` (scalar) of `output` stores the CPU-time of each method.

If MMA or MMA^S are used then the structure `output` contains several other fields. The fields `ML` (vector), and `loglik` (scalar) store the maximum likelihood values of the MMA or MMA^S weights and corresponding value of the log-likelihood function of Equation (35), respectively. The fields `std` (vector) and `corr` (matrix) list the posterior standard deviations and correlation coefficients of the weights.

If the BMA method is used then structure `options` returns to the user several more fields that summarize the performance of the BMA mixture distribution. The field `pred` (matrix) stores the lower and upper prediction quantiles of the BMA forecast distribution corresponding to the n_α different confidence level(s) specified in the field `alpha` of structure `options`. The matrix `pred` will have $2n_\alpha$ columns. Columns 1 and $2n_\alpha$ of matrix `pred` will make up the prediction interval of the largest value of `alpha`; columns 2 and $2n_\alpha - 1$ will define the lower and upper prediction quantiles of the second largest confidence level of field `alpha`, and so forth. The fields `coverage` and `spread` are $1 \times n_\alpha$ vectors with the coverage and spread of the prediction intervals of the different `alpha` values. These values are listed in order of increasing values of `alpha`.

Diagnostic information about the performance of the $\text{DREAM}_{(\text{ZS})}$ algorithm can be found in fields `MR_stat` (matrix), `R_stat` (matrix) and `AR` (matrix) of the return argument `output`. These fields are only defined if MMA, MMA^S , or BMA are used as these three methods use MCMC simulation with $\text{DREAM}_{(\text{ZS})}$ to find the maximum likelihood values of the weights (MMA and MMA^S) or weights and standard deviation(s) of the members forecast distribution (BMA). The fields `R_stat` and `MR_stat` list the values of the \hat{R} and \hat{R}^d convergence diagnostics of *Gelman and Rubin* (1992) and *Brooks and Gelman* (1998) at different iterations, respectively. Field `AR` of `output` stores the acceptance rate of $\text{DREAM}_{(\text{ZS})}$. The MATLAB command

```
plot (output.R_stat (:,1), output.R_stat (:,2:end))
```

 (39)

plots the evolution of the \hat{R} convergence diagnostic of the weights of each member of the ensemble. If BMA is used then this plot includes as well the standard deviation of the members' forecast distribution. The results in this graph can be used to judge when convergence of $\text{DREAM}_{(\text{ZS})}$ has been achieved and thus which samples to return in `output` argument `phi`. The latest release of the MODELAVG toolbox has a new function called `BMA_sample` which, when called after the main code has terminated will generate samples of the BMA forecast distribution and returns exact estimates of the mean and variance of the BMA mixture distribution at each time. The samples can be used to compute different (insufficient) performance metrics (e.g. reliability, coefficient of variation) and sufficient scoring rules of the BMA model (Vrugt, 2023). Please refer to example 12.

The MODELAVG toolbox not only returns to the user the variables `phi` and `output` but also generates graphical output. These figures display many of the input and output variables of the toolbox, and include (i) a time series plot of the forecast ensemble with averaged forecast and verifying observations, (ii) a plot of the predictions intervals of the BMA model and corresponding observations, (iii) histograms and bivariate scatter plots of the posterior samples of $\text{DREAM}_{(\text{ZS})}$, (iv) trace plots of the Markov chains sampled by $\text{DREAM}_{(\text{ZS})}$, (v) trace plots of the convergence diagnostics of $\text{DREAM}_{(\text{ZS})}$, and (vi) a quantile-quantile plot of the residuals of the weighted-average forecast. The main results of the toolbox are also written to the file "MODELAVG_output.txt" in the MATLAB editor. The case study section provides a screen shot of the

content of this file.

V2.0 of the MODELAVG toolbox has added the computation of *strictly proper* scoring rules of the BMA distribution forecasts. The quadratic, logarithmic, spherical and continuous ranked probability scoring rules are computed for each training observation (it is also possible to compute them for an evaluation data period, see next) and returned in $n \times 1$ vectors in fields QS, LS, SS and CRPS of structure output. We also return the mean values of these scoring rules in fields mQS, mLS, mSS and mCRPS. Furthermore, we also compute the 1 and 2-norm of the BMA mixture PDF at each time t and return the output in field norm_mix of output. Additionally, we return the mean and variance of the BMA distribution forecast at each time $t = (1, \dots, n)$ in $n \times 1$ vectors mu_mix' and var_mix, respectively. Lastly, time-averaged values of performance metrics such as the reliability and coefficient of variation of the BMA distribution forecasts are stored in mRLBL_anal and mCV_anal of structure array output.

3.7. Evaluation data set: structure val

The output variables returned by MODELAVG apply to the training data set only. The built-in script MODELAVG_EVAL allows the user to calculate performance statistics for the evaluation data set. This function can be executed as follows

```
[val] = MODELAVG_eval(method,D_eval,y_eval,options,phi,a,b,output) (40)
```

where method and options are defined by user prior to running the main function, MODELAVG of the MODELAVG toolbox, phi is the first output argument (matrix of weights and/or more) of the MODELAVG function, a and b are $1 \times K$ vectors with intercepts and slopes of the K ensemble members derived from linear bias correction of the training data set using Equation (2), output is the second output argument of the MODELAVG function, and D_val ($m \times K$ matrix) and y_val ($m \times 1$ vector) signify the ensemble forecasts and corresponding observations of the evaluation data set, respectively. The fields of the return argument val are listed in table 5.

MODELAVG MANUAL

Table 5: Content of the argument `val` computed by the function `MODELAVG_EVAL` after the MODELAVG toolbox has returned the output arguments `phi` and `output`

Field <code>val</code>	Description	Content
<code>Ye</code>	Deterministic point forecast, Equation (4)	$m \times 1$ vector
<code>R2</code>	Coefficient of determination of point forecasts	scalar
<code>RMSE</code>	Root mean square error point forecast	scalar
<code>R</code>	Pearson correlation coefficient of point forecasts and evaluation data	scalar
<code>RMSE_mod</code>	Root mean square error ensemble members	$1 \times K$ vector
BMA method		
<code>loglik</code>	Log-likelihood of Equation (21) for evaluation data record	scalar
<code>pred</code>	Lower and upper prediction intervals BMA mixture distribution	$m \times 2n_\alpha$ matrix
<code>coverage</code>	Coverage of n_α prediction intervals	$1 \times n_\alpha$ vector
<code>spread</code>	Average width of n_α prediction intervals	$1 \times n_\alpha$ vector
<code>par_unc</code>	BMA prediction quantiles due to parameter uncertainty only	$m \times 2n_\alpha$ matrix
<code>pdf_y</code>	PDF of BMA mixture distribution for each evaluation observation	$m \times 1$ vector
<code>cdf_y</code>	CDF of BMA mixture distribution for each evaluation observation	$m \times 1$ vector
<code>CRPS</code>	Continuous ranked probability score for each evaluation observation	$m \times 1$ vector
<code>QS</code>	Quadratic score for each evaluation observation	$m \times 1$ vector
<code>LS</code>	Logarithmic score for each evaluation observation	$m \times 1$ vector
<code>SS</code>	Spherical score for each evaluation observation	$m \times 1$ vector
<code>mRLBL_anal</code>	Reliability of BMA distribution forecasts	scalar
<code>mCv_anal</code>	Time-averaged coefficient of variation BMA mixture [= exact]	scalar
<code>mQS</code>	Time-averaged quadratic score BMA forecast distribution	scalar
<code>mLS</code>	Time-averaged logarithmic score BMA forecast distribution	scalar
<code>mSS</code>	Time-averaged spherical score BMA forecast distribution	scalar
<code>mCRPS</code>	Time-averaged continuous ranked probability score	scalar
<code>mix_norm</code>	1-norm and 2-norm of BMA mixture PDF	$m \times 2$ matrix
<code>mu_mix</code>	Mean of BMA mixture forecast [= exact]	$m \times 1$ vector
<code>var_mix</code>	Variance of BMA mixture forecast [= exact]	$m \times 1$ vector
<code>KGE</code>	Kling-Gupta efficiency of weighted-average BMA forecast	scalar

Thus, the structure `val` stores performance metrics, scoring rules, prediction quantiles and other variables for the evaluation data set. These metrics can be used to assess which conditional PDF and/or treatment of the variance of the predictive PDFs is most supported by an independent evaluation data set. The structure `val` is computed after the main function `MODELAVG` has returned its output (with/without screen output).

4. Numerical examples

I now illustrate the main functionalities of the MODEAVG package by application to multi-model forecast ensembles of river discharge, surface temperature and sea level pressure, respectively¹.

¹These examples do not use the lognormal, generalized normal and Weibull conditional forecast distributions as these three densities were added later to V2.0 of the MODELAVG toolbox.

4.1. Case Study 1: The rainfall-runoff transformation

The first case study involves an eight-member ensemble of calibrated watershed models of the Leaf River, near Collins, Mississippi. This discharge ensemble was created by *Vrugt and Robinson (2007)* and used to evaluate the sharpness and coverage of the BMA forecast distribution. Figure 7 provides a snapshot of the model ensemble for a portion of the training data set.

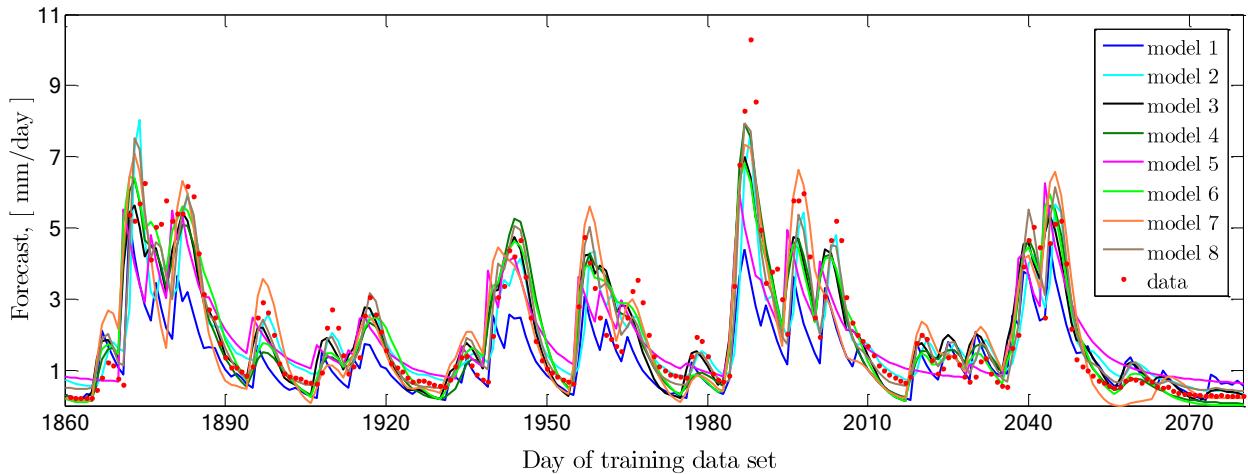


Figure 7: Streamflow predictions of the eight individual models of the ensemble for a representative portion of the calibration period. The red dots represent the verifying observations.

The spread of the ensemble is sufficient and generally brackets the observations (red dots). The calibrated models appear to provide somewhat different forecasts. This is a necessary requirement for an ensemble prediction system, otherwise model averaging cannot improve the forecast skill and distribution.

The file "example_1.m" summarizes the setup of case study 1 in the MODELAVG toolbox.

MATLAB input script "example_1.m", with data and setup of case study 1. The BMA method is used (`method = 'bma'`) with a gamma conditional pdf (`options.PDF = 'gamma'`) for the members' forecast distribution. As the field `VAR` of `options` is set to '3', the standard deviation of the gamma distribution is assumed to be heteroscedastic, or, $\sigma_{tk} = cD_{tk}$, where the value of the multiplier c applies to all models of the ensemble. The discharge forecasts of the ensemble members and verifying observations are stored in the ascii-file "discharge.txt". This data file is unpacked in the $n \times K$ matrix `D` and $n \times 1$ vector `y`. Bias correction is used of the forecasts of each member using the linear regression model of Equation (2).

```
% -----
%
% MM      MM    OOOOOOOO  DDDDDDDD  EEEEEEEE  LL          AAA        VV        VV  GGGGGGGG %
% MMM     MM    OOOOOOOOOO  DDDDDDDDD  EEEEEEEE  LL          AA AA      VV        VV  GG      GG %
% MMMMM   MMMM  OO     OO  DD     DD  EE          LL          AA AA      VV        VV  GG      GG %
% MM MM  MM MM  OO     OO  DD     DD  EEEE     LL          AA AA      VV        VV  GGGGGGGG %
% MM  MMM  MM  OO     OO  DD     DD  EEEE     LL          AAAAAAAA  VV  VV  GGGGGGGG %
% MM      MM  OO     OO  DD     DD  EE          LL          AA AA      VV  VV  GG      %
% MM      MM    OOOOOOOOOO  DDDDDDDDD  EEEEEEEE  LLLLLLLL  AA        AA      VV  VV  GGGGGGGG %
% MM      MM    OOOOOOOO  DDDDDDDD  EEEEEEEE  LLLLLLLL  AA        AA      VVV  GGGGGGGG %
%
% -----
%
%% CASE STUDY I: RAINFALL-RUNOFF TRANSFORMION - ENSEMBLE OF CALIBRATED WATERSHED MODELS
%% CHECK: J.A. VRUGT AND B.A. ROBINSON, WRR, 43, W01411, doi:10.1029/2005WR004838, 2007

%% DEFINE MODEL AVERAGING METHOD
method = 'bma'; % 'ewa'/'bga'/'aica'/'bica'/'gra'/'bma'/'mma'/'mma-s'

%% BMA -> CONDITIONAL DISTRIBUTION NEEDS TO BE DEFINED
options.PDF = 'gamma'; % normal conditional pdf
options.VAR = '1'; % common constant variance
options.alpha = 0.95; % prediction intervals of BMA model (0.90/0.95/0.99)
options.print = 'yes'; % print output (figures, tables) to screen

%% NOW LOAD DATA
S = load('discharge.txt'); % daily discharge forecasts (mm/day) of models and verifying data
T_idx = [ 1:1:3000 ]; % start/end training period

%% DEFINE ENSEMBLE AND VECTOR OF VERIFYING OBSERVATIONS
D = S(T_idx,1:8); y = S(T_idx,9);

%% APPLY LINEAR BIAS CORRECTION TO ENSEMBLE ( UP TO USER )
[ D , a , b ] = Bias_correction ( D , y );

%% NUMBER OF PARAMETERS OF EACH MODEL (ABC/GR4J/HYMOD/TOPMO/AWBM/NAM/HBV/SACSMA)
options.p = [ 3 4 5 8 8 9 9 13 ]; % ( only used for AICA, BICA, MMA, MMA-S)

%% Run MODELAVG toolbox with two outputs
[ phi , output ] = MODELAVG ( method , D , y , options );
```

The BMA method is used for postprocessing of the discharge forecast ensemble. A gamma forecast

distribution is used for each ensemble members' conditional pdf. The standard deviation of this distribution is assumed to be heteroscedastic with coefficient c that applies to all models of the ensemble. This thus requires the inference of $d = K + 1$ parameters with the DREAM_(ZS) algorithm, namely the values of the K weights of the watershed models, and multiplier c , or $\mathbf{x} = (\beta_1, \dots, \beta_K, c)^\top$ (see Table 3).

Figure 8 presents histograms of the marginal posterior distribution of the weights of each model of the ensemble. The maximum likelihood values are separately indicated with the "x" symbol.

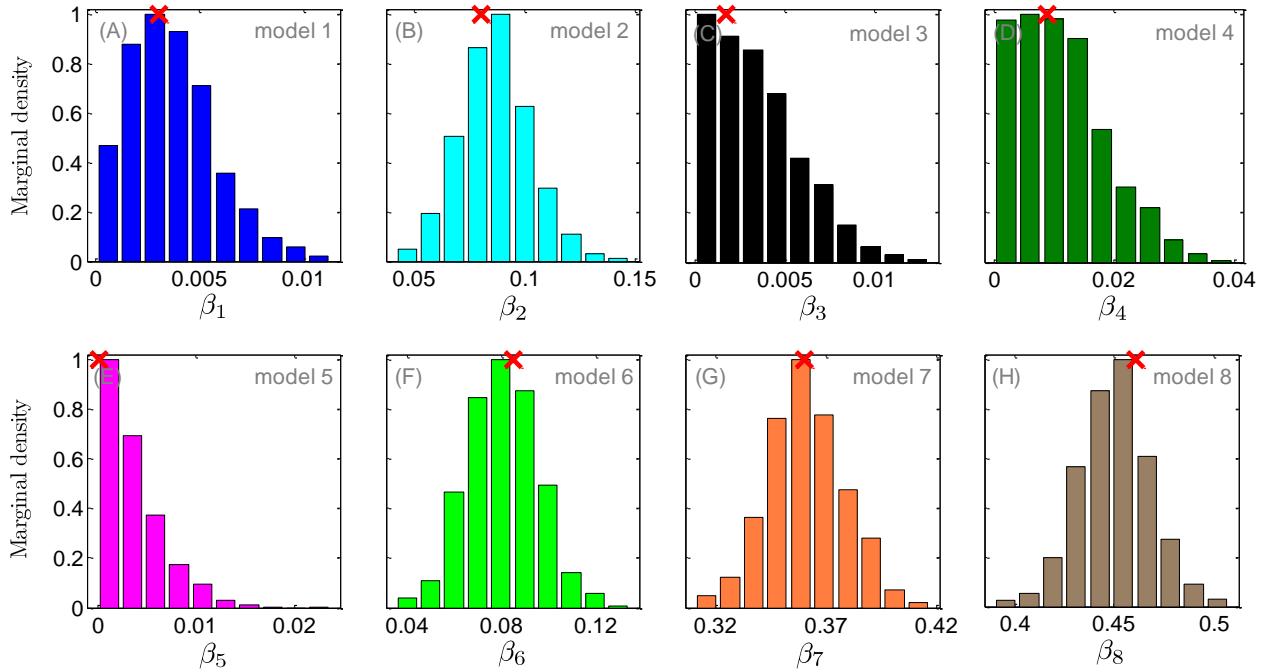


Figure 8: Histograms of the marginal posterior distribution of the weights of each watershed model of the discharge ensemble. The corresponding forecasts of each model can be found in Figure 7. The red cross symbol in each plot indicates the maximum likelihood solution.

Most histograms are well described with a normal distribution, except those of models with a very low weight truncated at zero to lie on the unit simplex. The posterior uncertainty of the models' weights appears rather small as most histograms are well defined. Models 1, 3, 4, and 5 receive negligible weights and can thus be removed from the ensemble without much loss of forecast skill. The marginal distributions of the weights are particularly useful to determine the importance of each members forecasts.

To understand how the BMA posterior parameter uncertainty translates into predictive uncertainty, please consider Figure 9 that presents the 95% hydrograph prediction uncertainty ranges (gray region) of the BMA mixture distribution for a portion of the training data period. The mean forecast of the BMA model is separately indicated with the black line, and the red dots denote the verifying observations.

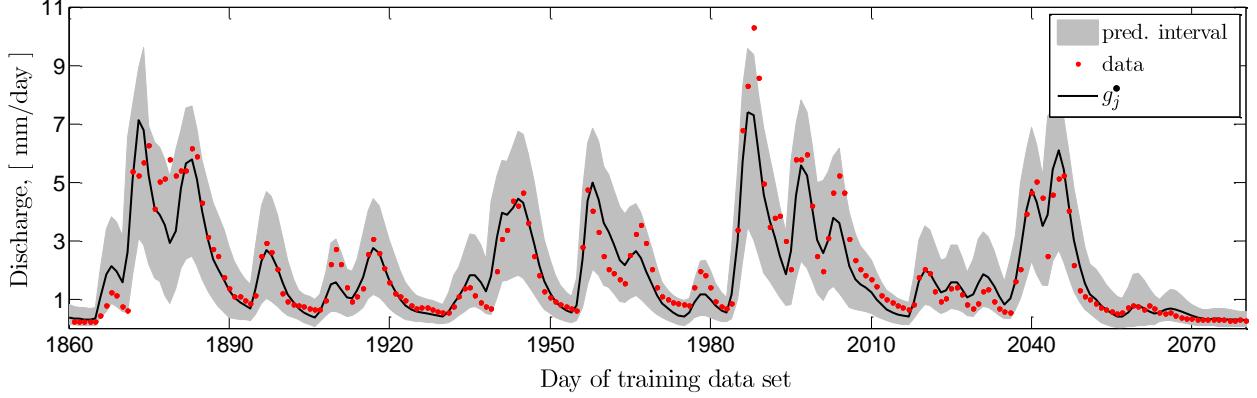


Figure 9: 95% prediction intervals (gray region) of the BMA predictive density for a representative portion of the 3000 days calibration period. The black line displays the mean point forecast of the BMA model derived from Equation (16), whereas the red dots signify the verifying observations.

The 95% prediction uncertainty ranges of the BMA model appear rather large, particularly at lower discharge values, yet envelop almost 95% observations (data appears in Table 6). The RMSE of the BMA point forecast (mean of the mixture density) is about 0.723 mm/day, and nearly similar as the value of 0.720 derived from the best model of the ensemble (= Sacramento soil moisture accounting model). Figure 10 makes a screen copy of the output (postprocessor) of the latest release (V2.0) of the MODELAVG toolbox using three different values for α , the prediction quantiles of the BMA model, `options.alpha = [0.5 0.9 0.95]`. The predictive PDF was changed from a gamma to a truncated normal distribution with heteroscedastic individual variance or `options.VAR = '4'`.

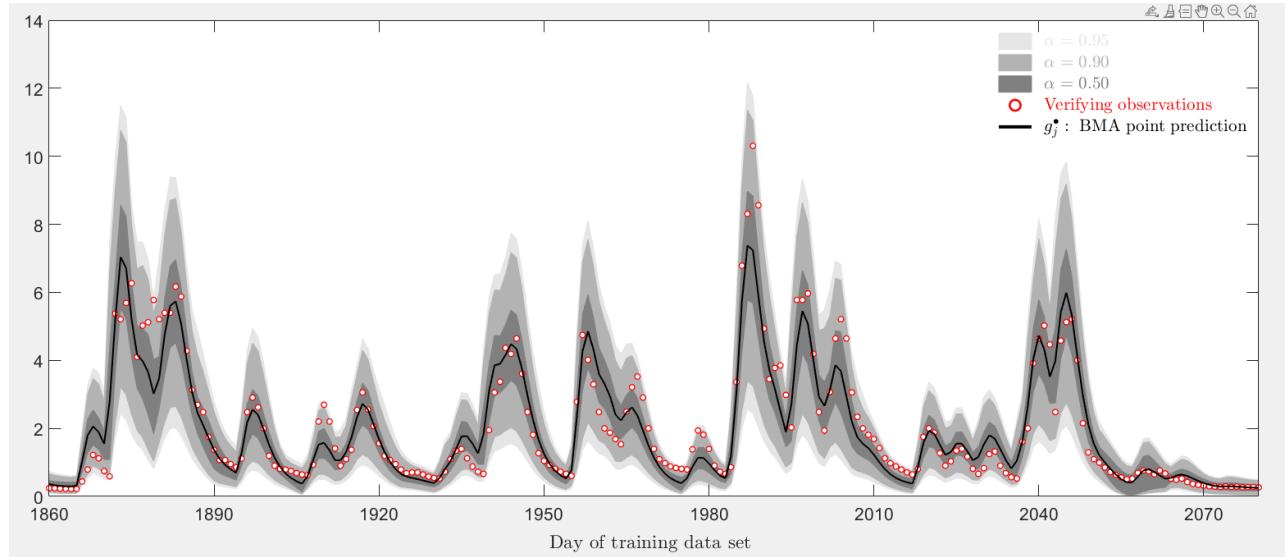


Figure 10: V2.0 of MODELAVG toolbox: 50, 90 and 95% prediction intervals (different gray tints) of the BMA predictive density for a representative portion of the 3000 days training data records. The black line displays the mean point forecast of the BMA model derived from Equation (16), whereas the red circles signify the verifying observations.

Table 6 summarizes the results of the BMA method for two different forecast distributions (`options.PDF = 'normal'` or `PDF = 'gamma'`) and four different treatments of the variance of this distribution (`options.VAR = '1'` or `'2'`, `'3'` or `'4'`). For convenience, I list only the maximum likelihood values of the weights of each implementation. Values listed in parentheses denote the posterior standard deviation derived from the DREAM_(ZS) sample. I also report the maximum value of the log-likelihood of Equation (21), the RMSE (mm/day) of the averaged forecast, coverage (%) and spread (mm/day) of the 95% prediction intervals of the BMA model.

Table 6: BMA results for the discharge ensemble of the Leaf River watershed using the normal and gamma forecast distribution and four different treatments of the variance of these distributions (see section 2.5.2). The different columns list the maximum likelihood values of the BMA model parameters, corresponding value of the log-likelihood function of Equation (21), RMSE (mm/day), coverage (%) and spread (mm/day) of the 95% prediction intervals of the BMA model.

BMA	Normal distribution				Gamma distribution			
	'1'	'2'	'3'	'4'	'1'	'2'	'3'	'4'
β_1	0.0169	0.0382	0.0022	0.0038	0.1087	0.0369	0.0026	0.0027
β_2	0.2005	0.1631	0.0882	0.0891	0.1859	0.1431	0.0852	0.0857
β_3	0.1041	0.0080	0.0059	0.0109	0.0180	0.1107	0.0025	0.0010
β_4	0.0701	0.1129	0.0271	0.0246	0.0022	0.0433	0.0054	0.0134
β_5	0.0330	0.0317	0.0002	0.0004	0.0633	0.0443	0.0007	0.0037
β_6	0.0444	0.1303	0.1228	0.1582	0.0076	0.0987	0.0818	0.0841
β_7	0.0423	0.1401	0.3226	0.3024	0.3342	0.2297	0.3663	0.3674
β_8	0.4887	0.3757	0.4309	0.4108	0.2801	0.2932	0.4555	0.4422
σ	0.4724				0.4013			
c			0.317				0.3548	
σ_1		0.5277				0.3348		
σ_2		0.1587				0.0854		
σ_3		5.4986				0.7936		
σ_4		1.0512				2.7103		
σ_5		0.1827				0.0489		
σ_6		0.0722				0.0470		
σ_7		0.0830				0.0882		
σ_8		0.1125				0.0959		
c_1			0.3023				0.3360	
c_2			0.2379				0.2910	
c_3			0.3737				0.9083	
c_4			0.1975				0.1234	
c_5			0.2974				0.5990	
c_6			0.4333				0.3868	
c_7			0.3096				0.3610	
c_8			0.2992				0.3495	
$\mathcal{L}(\Phi \mathbf{D}, \tilde{\mathbf{y}})$	-2416.3	-617.92	150.93	164.95	-2965.6	-831.63	244.75	249.28
RMSE	0.7072	0.7210	0.7234	0.7258	0.7499	0.7332	0.7237	0.7245
Coverage	96.133	94.767	95.600	95.667	95.867	94.633	96.033	95.967
Spread	2.2263	2.0809	1.4444	1.4589	2.2217	1.4665	1.4493	1.4465

The maximum likelihood values of the BMA weights depend somewhat on the assumed forecast distri-

bution of the deterministic prediction of each model. The HBV (model 7) and SAC-SMA (model 8) models almost always receive the highest weights of the ensemble and their forecasts are thus of crucial importance for the BMA model. The TOPMO model (model 6) receives particularly low weights, despite it having the second lowest RMSE for the 3000-day raining data period. The TOPMO forecasts might be redundant as they are correlated with other models of the ensemble. Note that the performance of the BMA model appears to be much more affected by the standard deviation of the forecast distribution, than the shape of this distribution (normal or gamma). The best results are obtained if a heteroscedastic standard deviation is assumed for the members forecast distribution. This is perhaps not surprising as the measurement error of the discharge data is known to increase with flow level. I refer interested readers to the paper of *Vrugt and Robinson (2007)* and *Rings et al. (2012)* for a more detailed analysis of the BMA results, and a comparison with data assimilation methods.

4.2. Case Study 2: 48 hour forecasting of sea level temperature

The second case study involves a five-member multianalysis ensemble of 48-h forecasts of surface temperature (in Kelvin) between January and June 2000 in the Pacific Northwest of the USA. The data set was created by the mesoscale short-range ensemble system of the University of Washington (UW) (*Grimit and Mass, 2002*) using different runs of the fifth-generation Pennsylvania State University - National Center for Atmospheric Research Mesoscale Model (MM5) and initial conditions from different operational centers.

I use a 25-day period between April 16 and 9 June 2000 for BMA model calibration. This involves a total of $n = 14668$ temperature forecasts at different locations in the Pacific Northwest. For some days the data were missing, so that the number of calendar days spanned by the training data set is larger than the number of days of training used.

Figure 11 displays the ensemble forecasts, \mathbf{D} and verifying observations, $\tilde{\mathbf{y}}$ for a small portion of the training data set.

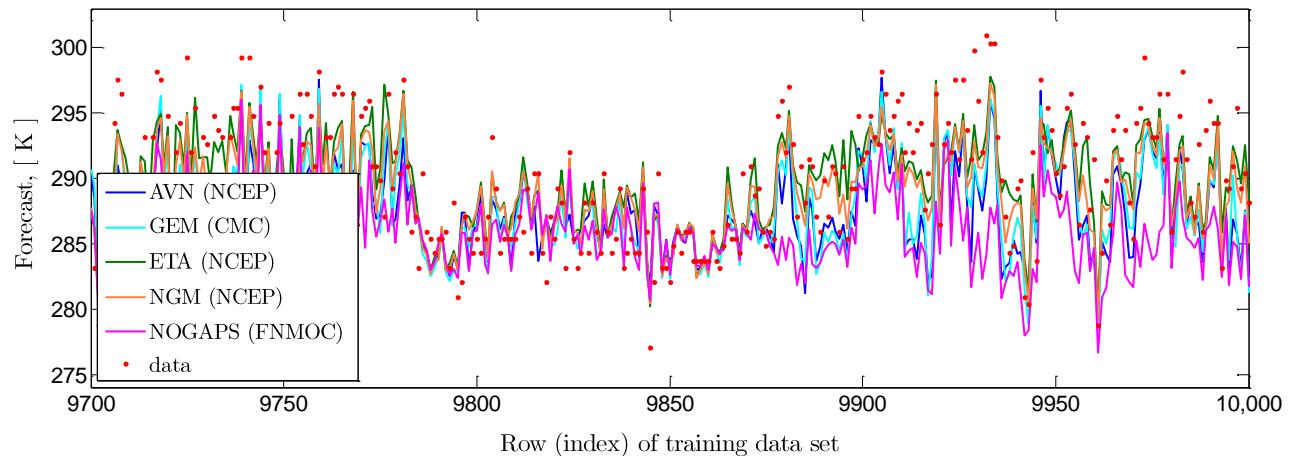


Figure 11: Temperature forecasts for the Pacific Northwest of the five-member UW ensemble for a short period of the training data set. The red dots represent the verifying observations.

The different members of the ensemble issue different 48-h temperature forecasts, yet the spread is not always sufficient to bracket the observations (red dots). Note, that *Raftery et al. (2005)* used the exact same

25-day data set to introduce and benchmark the BMA method.

The file "example_2.m" summarizes the setup of case study 2 in the MODELAVG toolbox.

MATLAB input script "example_2.m", with data and setup of case study 2. The BMA method is used (method = 'bma') with a normal conditional pdf, options.PDF = 'normal' for the members' forecast distribution. As the field VAR of options is set to '2', each model is assumed to have a different standard deviation of the forecast distribution, but this standard deviation is constant. The surface temperature forecasts and verifying observations are stored in the ascii-file "temp.txt". This data file is unpacked in the $n \times K$ matrix D and $n \times 1$ vector y. Bias correction is used of the forecasts of each member using the linear regression model of Equation (2).

```
% -----
%
% MM      MM    OOOOOOOO   DDDDDDDD   EEEEEEEE   LL          AAA     VV     VV   GGGGGGGG %
% MMM     MM    OOOOOOOOOO  DDDDDDDDDD  EEEEEEEE   LL          AA AA   VV     VV   GG   GG %
% MMMM    MMMM   OO     OO   DD     DD   EE       LL          AA AA   VV     VV   GG   GG %
% MM MM MM MM   OO     OO   DD     DD   EEEE   LL          AA AA   VV     VV   GGGGGGGG %
% MM   MMM   MM   OO     OO   DD     DD   EEEE   LL          AAAAAAAA   VV     VV   GGGGGGGG %
% MM     MM   OO     OO   DD     DD   EE       LL          AA     AA   VV     VV   GG %
% MM     MM    OOOOOOOOO  DDDDDDDDDD  EEEEEEEE   LLLLLLLL  AA     AA   VV     VV   GGGGGGGG %
% MM     MM    OOOOOOOO   DDDDDDDD   EEEEEEEE   LLLLLLLL  AA     AA   VVV    GGGGGGGG %
%
% -----
%
%% CASE STUDY II: 48-Forecasts of sea surface temperature
%% CHECK: A.E. RAFTERY ET AL., MWR, 133, pp. 1155-1174, 2005.

%% DEFINE MODEL AVERAGING METHOD
method = 'bma'; % 'ewa'/'bga'/'aica'/'bica'/'gra'/'bma'/'mma'/'mma-s'

%% BMA -> CONDITIONAL DISTRIBUTION NEEDS TO BE DEFINED
options.PDF = 'normal'; % pdf predictor: normal/heteroscedastic/gamma
options.VAR = '2'; % individual non-constant variance
options.alpha = 0.95; % prediction intervals of BMA model (0.90/0.95/0.99)
options.print = 'yes'; % print output (figures, tables) to screen

%% NOW LOAD DATA
T = load('temp.txt'); % 48-hour forecasts temperature (Kelvin) and verifying data

%% DEFINE ENSEMBLE AND VECTOR OF VERIFYING OBSERVATIONS ( APRIL 16 TO JUNE 9, 2000 )
idx = find(T(:,1) == 2000 & T(:,2) == 4 & T(:,3) == 16); start_idx = idx(1);
idx = find(T(:,1) == 2000 & T(:,2) == 6 & T(:,3) == 9); end_idx = idx(1);
D = T(start_idx:end_idx,5:9); y = T(start_idx:end_idx,4);

%% APPLY LINEAR BIAS CORRECTION TO ENSEMBLE ( UP TO USER )
[ D , a , b ] = Bias_correction ( D , y );

%% Run MODELAVG toolbox with two outputs
[ phi , output ] = MODELAVG ( method , D , y , options );
```

I implement the BMA method for the 25-day temperature ensemble and assume a normal distribution

for the members forecast distribution. The choice of this forecast distribution is supported by the frequency distribution of the temperature observations of the training data set (see also Figure 4A). As temperature observations have a constant measurement error, we assume that the forecast distribution has a fixed variance, but allow this variance to vary among the ensemble members. This thus involves the inference of $d = 2K = 10$ parameters with DREAM_(ZS), namely the weight and standard deviation of each models' forecast distribution. Appendix E presents the screen output of the main function of the MODELAVG toolbox for the data set and BMA model defined in the MATLAB input file above.

Figure 12 presents trace plots of the \hat{R} -statistic of *Gelman and Rubin* (1992) for the DREAM_(ZS) sampled weights and standard deviations of each members' conditional distribution.

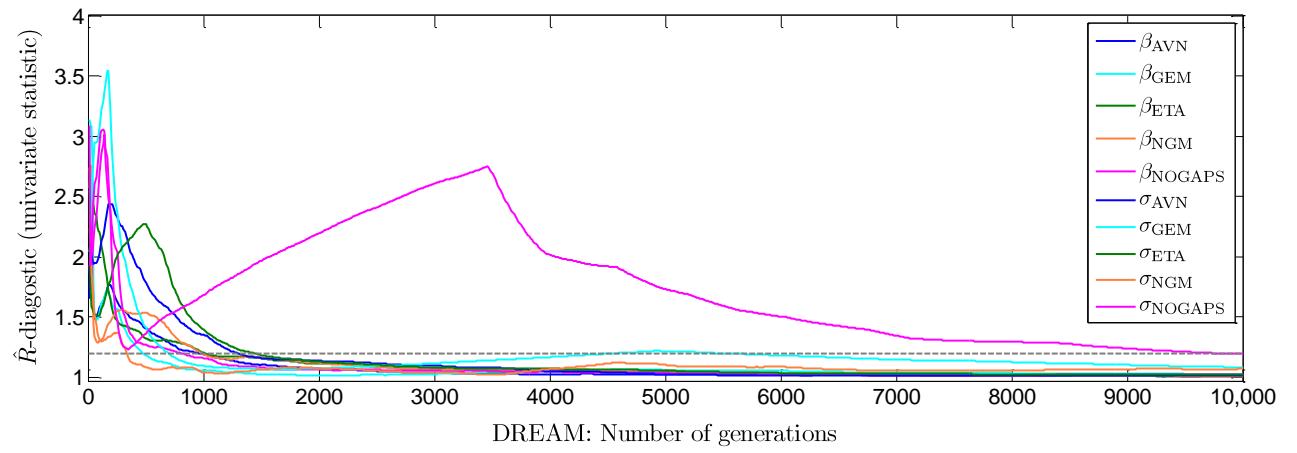


Figure 12: Evolution of the \hat{R} scale reduction factor of *Gelman and Rubin* (1992) used to diagnose convergence of the sampled Markov chains of DREAM_(ZS) to a limiting distribution. The dashed grey horizontal line signifies the commonly used threshold for convergence.

The \hat{R} diagnostic illustrates that about 10,000 generations are required with DREAM_(ZS) to reach convergence to a stationary distribution $\hat{R} \leq 1.2; \forall i = \{1, \dots, d\}$. This constitutes a relative large number of iterations and is explained by bimodality of the posterior distribution of the standard deviation of the forecast distribution of ETA. This is demonstrated graphically in Figures E17 and E18 in Appendix E.

Figure 13 presents marginal distributions of the posterior samples of the BMA weights (top panel) and standard deviations (bottom panel) of the forecast distribution. The maximum likelihood values are separately indicated with the red cross "x" symbol.

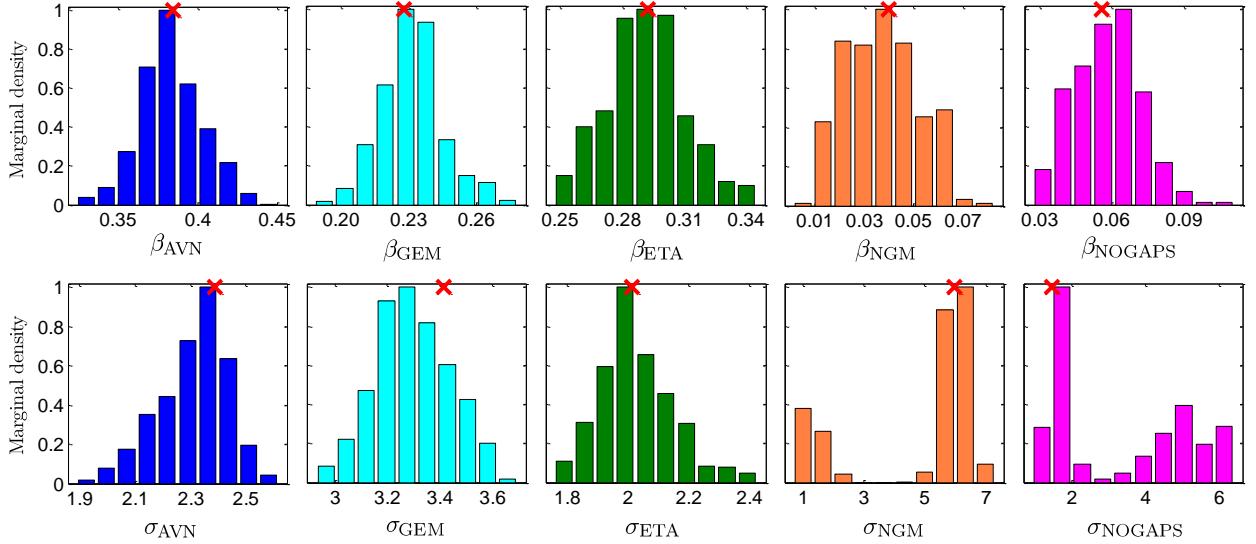


Figure 13: Histograms of the marginal posterior distribution of the weights and standard deviations of the five models of the ensemble. The maximum likelihood solution of each BMA model parameter is indicated with the red cross.

All histograms appear approximately Gaussian, and the median solution of each weight and standard deviation coincides almost perfectly with their respective maximum likelihood values. Note the presence of bimodality in the marginal distribution of the standard deviation of the forecast distribution of the NGM and NOGAPS models. This bimodality is particularly obvious for σ_{NGM} (two disconnected modes) and makes it much more difficult to MCMC methods to sample the target distribution (*Vrugt, 2016*). Hence, this explains why DREAM_(ZS) needs a relatively large number of function evaluations to convergence to the posterior distribution. In fact, NGM and NOGAPS, receive much lower weights than the order three models of the ensemble, and can perhaps be discarded without affecting too much the predictive skill and coverage of the BMA model.

Figure 14 presents the 95% prediction uncertainty of the maximum likelihood BMA mixture distribution. The deterministic point forecast of the BMA model is separately indicated with the solid black line. This point predictor is derived from Equation (16) using the forecast ensemble, \mathbf{D} and maximum likelihood values of the weights, $\hat{\beta}_{\text{BMA}}$.

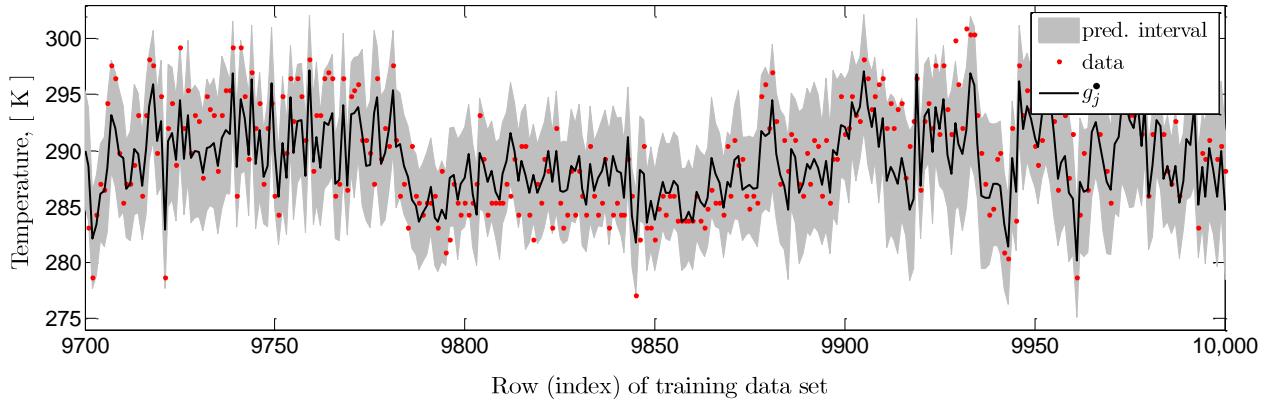


Figure 14: 95% prediction intervals (gray region) of the BMA predictive density for a representative portion of the 25-day training data set. The black line displays the mean point forecast of the BMA model derived from Equation (16), whereas the red dots signify the verifying observations.

The 95% prediction uncertainty ranges of the BMA model appear rather large with average spread of about 9.7 Kelvin and coverage of 90.7% of the observations. The RMSE of the BMA point forecast (mean of the mixture density) is approximately 2.96 K which is equivalent to the RMSE of the best model (AVN) of the ensemble (see Table 7).

Table 7 lists the (maximum likelihood) values of the weights for each of the model averaging methods of the MODELAVG toolbox.

Table 7: Results of different model averaging methods for the five-member multimodel ensemble of 48-h forecasts of surface temperature in the Pacific Northwest of the USA. The first two columns list the names of the different models of the ensemble and their corresponding RMSE values (in units of Kelvin), respectively. Subsequent columns document the (maximum likelihood) values of the weights of each model averaging method of the toolbox. The bottom part of the Table lists the RMSE of the deterministic point forecasts of each method. This forecast is calculated with Equation (4) using the ensemble forecasts of \mathbf{D} and maximum likelihood values of the weights.

Model	RMSE	EWA	BGA	AICA [†]	BICA [†]	GRA	BMA	MMA [†]	MMA ^{S †,‡}
AVN	3.0578	0.2000	0.2208	1.0000	1.0000	0.4836	0.3896	0.4798	0.4730
GEM	3.3425	0.2000	0.1848	0.0000	0.0000	0.2028	0.2277	0.2002	0.1984
ETA	3.1143	0.2000	0.2129	0.0000	0.0000	0.3602	0.2877	0.3596	0.3119
NGM	3.1881	0.2000	0.2031	0.0000	0.0000	-0.0692	0.0442	-0.0712	0.0000
NOGAPS	3.4020	0.2000	0.1784	0.0000	0.0000	0.0227	0.0508	0.0316	0.0166
Point	2.9941	2.9886	3.0578	3.0578	3.0578	2.9541	2.9587	2.9542	2.9548

†: I assume $p = 20 * \text{ones}(1, K)$ (twenty parameters assigned to each model)

‡: Presence of numerous local optima on likelihood surface

Information criterion averaging (AICA and BICA) assigns the AVN model a weight of unity, whereas all other models are given a zero weight. The BMA method distributes the weights more equally among the different ensemble members, yet assigns the NGM and NOGAPS members relatively low weights. Discarding these two models hardly affects the performance of each of the model averaging methods.

The point forecasts of the different model averaging methods exhibit a rather similar performance. The main advantage of the BMA method, however is that it provides a forecast distribution which can be used

for probabilistic analysis and prediction. Nevertheless, if point forecasting is of main concern, the Granger-Ramanathan averaging provides the lowest RMSE of the deterministic forecast at a negligible CPU-cost.

4.3. Case Study 3: 48 hour forecasting of sea surface pressure

I now do a similar analysis but using 48-h forecasts of sea surface pressure (in hPa) from the University of Washington mesoscale short-range ensemble prediction system (*Grimit and Mass*, 2002). The same 25-day training period as in case study 2 is used for BMA model calibration (April 16 - June 9, 2000). For some days the data were missing, so that the number of calendar days spanned by the training data set is larger than the number of days of training used. The training data set includes $n = 4013$ observations.

Figure 15 displays the ensemble forecasts of the sea level pressure, \mathbf{D} and verifying observations, $\tilde{\mathbf{y}}$ for a small portion of the training data set.

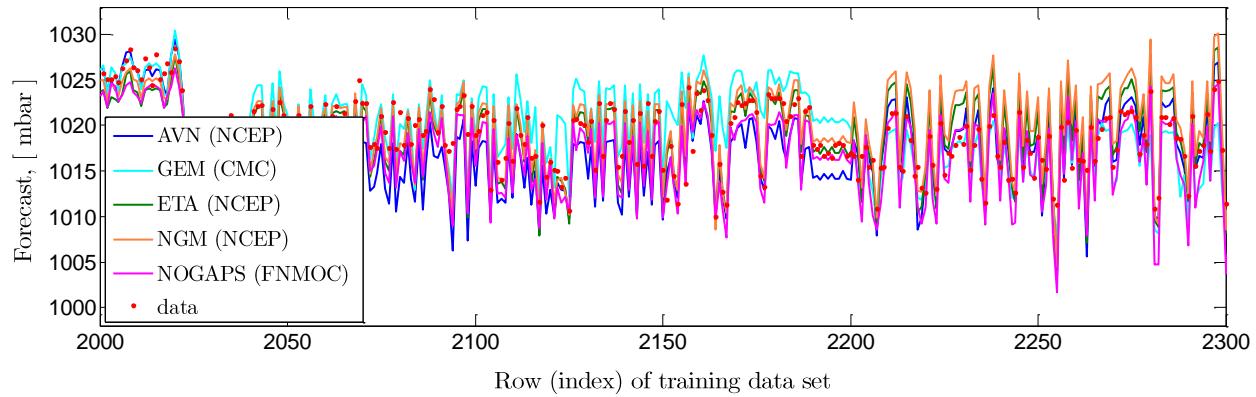


Figure 15: 48-h sea level pressure forecasts for the Pacific Northwest of the five member UW ensemble. I only display a small portion of the training data set. The red dots represent the verifying observations.

The models differ in their 48-h pressure forecasts, and the ensemble covers a large majority of the observed sea level pressure data (red dots). Note, that *Raftery et al.* (2005) used the exact same 25-day data set to introduce and benchmark the BMA method.

The file "example_3.m" summarizes the setup of case study 3 in the MODELAVG toolbox.

MATLAB input script "example_3.m", with data and setup of case study 3. The BMA method is used (method = 'bma') with a normal conditional pdf, options.PDF = 'normal' for the members' forecast distribution. As the field VAR of options is set to '2', each model is assumed to have a different standard deviation of the forecast distribution, but this standard deviation is constant. The sea pressure forecasts are verifying observations are stored in the ascii-file "pressure.txt". The built-in function `find` is used to extract the temperature forecasts of April 16 to June 9. This data is unpacked in the $n \times K$ matrix D and $n \times 1$ vector y. Bias correction is used of the forecasts of each member using the linear regression model of Equation (2).

```
% -----
%
% MM      MM    OOOOOOOO    DDDDDDDD    EEEEEEEE    LL          AAA        VV        VV    GGGGGGGG %
% MMM     MM    OOOOOOOOOO    DDDDDDDDDD    EEEEEEEE    LL          AA AA      VV        VV    GG      GG %
% MMMMM   MMMM   OO      OO    DD      DD    EE          LL          AA AA      VV        VV    GG      GG %
% MM MM  MM  MM   OO      OO    DD      DD    EEEEEEE    LL          AA AA      VV        VV    GGGGGGGG %
% MM   MMM   MM   OO      OO    DD      DD    EEEEEEE    LL          AAAAAAAA    VV      VV    GGGGGGGG %
% MM      MM   OO      OO    DD      DD    EE          LL          AA AA      VV      VV    GG      %
% MM      MM    OOOOOOOOOO    DDDDDDDDDD    EEEEEEEE    LLLLLLLL    AA      AA      VV      VV    GGGGGGGG %
% MM      MM    OOOOOOOO    DDDDDDDD    EEEEEEEE    LLLLLLLL    AA      AA      VVV     GGGGGGGG %
%
%
%% CASE STUDY III: 48-FORECASTS OF SEA SURFACE PRESSURE
%% CHECK: A.E. RAFTERY ET AL., MWR, 133, pp. 1155-1174, 2005.

%% DEFINE MODEL AVERAGING METHOD
method = 'bma'; % 'ewa'/'bga'/'aica'/'bica'/'gra'/'bma'/'mma'/'mma-s'

%% BMA -> CONDITIONAL DISTRIBUTION NEEDS TO BE DEFINED
options.PDF = 'gamma'; % gamma distribution
options.VAR = '3'; % individual non-constant variance
options.alpha = 0.95; % prediction intervals of BMA model (0.90/0.95/0.99)
options.print = 'yes'; % print output (figures, tables) to screen

%% NOW LOAD DATA
P = load('pressure.txt'); % 48-hour forecasts of air-pressure (mbar) and verifying data

%% DEFINE ENSEMBLE AND VECTOR OF VERIFYING OBSERVATIONS ( APRIL 16 TO JUNE 9, 2000 )
idx = find(P(:,1) == 2000 & P(:,2) == 4 & P(:,3) == 16); start_idx = idx(1);
idx = find(P(:,1) == 2000 & P(:,2) == 6 & P(:,3) == 9); end_idx = idx(end);
D = P(start_idx:end_idx,5:9); y = P(start_idx:end_idx,4);

%% APPLY LINEAR BIAS CORRECTION TO ENSEMBLE ( UP TO USER )
[ D , a , b ] = Bias_correction ( D , y );

%
% Run MODELAVG toolbox with two outputs
[ phi , output ] = MODELAVG ( method , D , y , options );
```

The BMA method is used for postprocessing of the 25-day sea level pressure ensemble. A normal distribution is used for the ensemble members' forecast distribution. The choice of this forecast distribution

is supported by the frequency distribution of the pressure observations of the training data set (see also Figure 4B). As air pressure observations have a constant measurement error, we assume that the forecast distribution has a fixed variance, but allow this variance to vary among the ensemble members. This thus involves the inference of $d = 2K = 10$ parameters with DREAM_(zs), namely the weight and standard deviation of each models' forecast distribution. Prior to this BMA model training, the individual members of the ensemble were bias corrected using simple linear regression of their forecasts on the verifying observations of the training data set.

Figure 16 presents histograms of the DREAM_(zs) derived marginal posterior distributions of the BMA weights and standard deviations of the different ensemble members. The optimal values derived with the EM algorithm are separately indicated in each panel with the black circle. These EM values are computed using the following command

$$[\text{beta}, \text{sigma}, \text{loglik}] = \text{EM_normal}(\text{D}, \text{Y}, \text{options}) \quad (41)$$

where `beta` ($1 \times K$ vector) and `sigma` ($1 \times K$ vector) return the maximum likelihood values of the weights and standard deviations of the members' forecast distribution, respectively, and variable `loglik` (scalar) contains the maximized value of the log-likelihood function of Equation (21). The function `EM_NORMAL` is discussed in Appendix B and part of the MODELAVG toolbox.

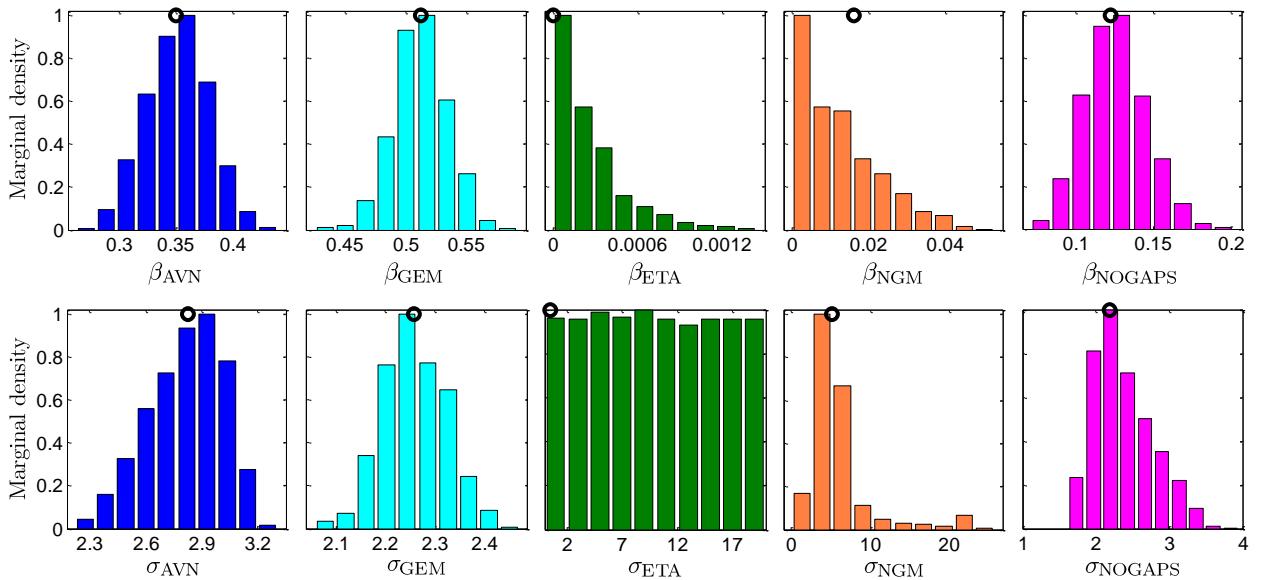


Figure 16: Histograms of the marginal posterior distributions of the BMA weights and standard deviations of the ensemble members' forecast distribution. The EM solution is separately indicated in each panel with the "o" symbol.

The results show an excellent agreement between the modes of the histograms derived from MCMC simulation with DREAM_(zs) (`loglik = -9864.1`) and the maximum likelihood values of the EM algorithm (`loglik = -9863.7`). Hence, previous applications of the EM algorithm are likely to have yielded robust estimates for the BMA weights and standard deviations (variances) (see also Vrugt *et al.* (2008b)). However, DREAM_(zs) has the desirable feature that it not only returns to the user the maximum likelihood values of

the BMA weights and standard deviations of the members' forecast distribution, but also their underlying posterior distribution and (posterior) correlation among ensemble members. This information is of great practical value as it helps determine each ensemble members' contribution to the predictive skill, and the dependencies among the forecasts of the K different members. This is crucial information to help determine which ensemble members to keep and which ones to discard as it takes time to setup and run each model of the ensemble. For instance, the ETA and NGM models receive very low weights in the present application. In fact, the forecast distribution of the ETA model is particularly ill-defined as its standard deviation has an almost uniform marginal distribution. The weights and standard deviations of the other models forecast distributions (AVN, GEM, and NOGAPS) appear much better defined, hence their forecasts play a key role in the construction of the BMA model.

I now turn attention to the DREAM_(zs) algorithm. Figure 17 displays trace plots of the sampled weights of the AVN (top), GEM (middle) and NOGAPS (bottom) model. I use color coding for each different Markov chain sampled by DREAM_(zs). The maximum likelihood values are indicated at the right hand side with the "x" symbol. The EM solutions are also presented separately with the black "o" symbol.

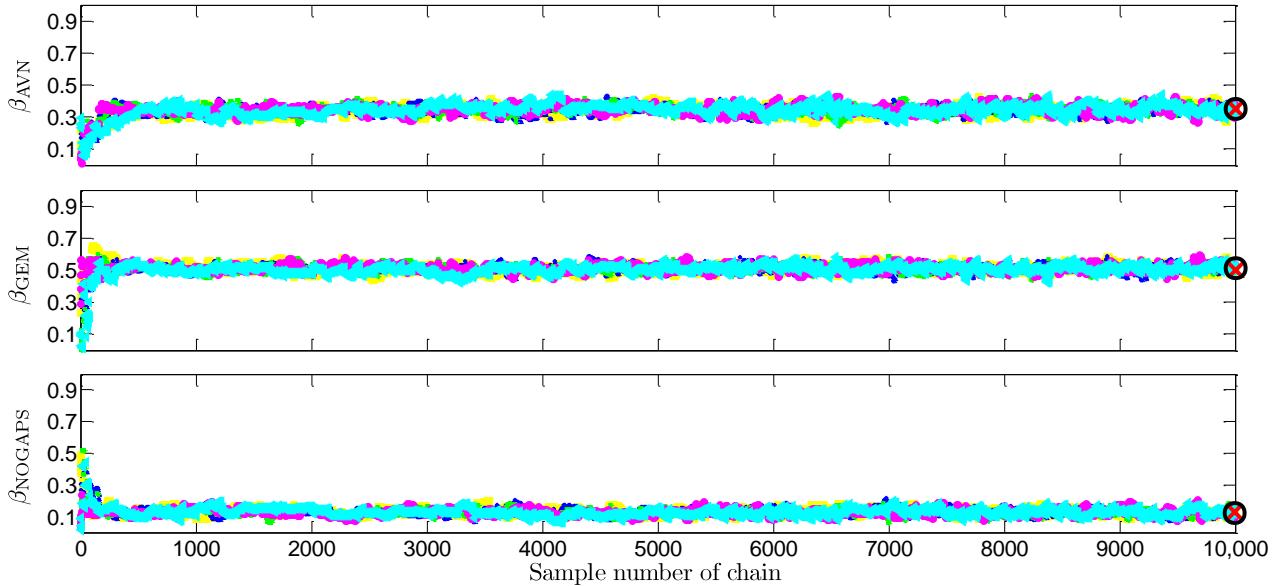


Figure 17: Trace plots of the sampled BMA weights of the AVN, GEM, and NOGAPS model in the Markov chains generated with DREAM_(zs). The maximum likelihood estimates computed of DREAM_(zs) and the EM algorithm are separately indicated at the right hand side in each panel using the \times and \circ symbols.

During the initial stages of the search (first few hundred BMA_CALC evaluations), the chains occupy different parts of the weight space, resulting in a relatively high value for the \hat{R} -convergence diagnostic (not shown). After this, the five chains settle down in the approximate same region of the parameter space and successively visit solutions stemming from a stable distribution. This demonstrates convergence to a limiting distribution.

Figure 18 presents the 95% prediction uncertainty of the maximum likelihood BMA mixture distribution. The weighted average (deterministic) point forecast of the BMA model is separately indicated with the solid black line. This point predictor is derived from Equation (16) using the maximum likelihood values of the

MODELAVG MANUAL

weights, $\hat{\beta}_{\text{BMA}}$, and forecasts of the members of the ensemble stored in the $n \times K$ matrix \mathbf{D} .

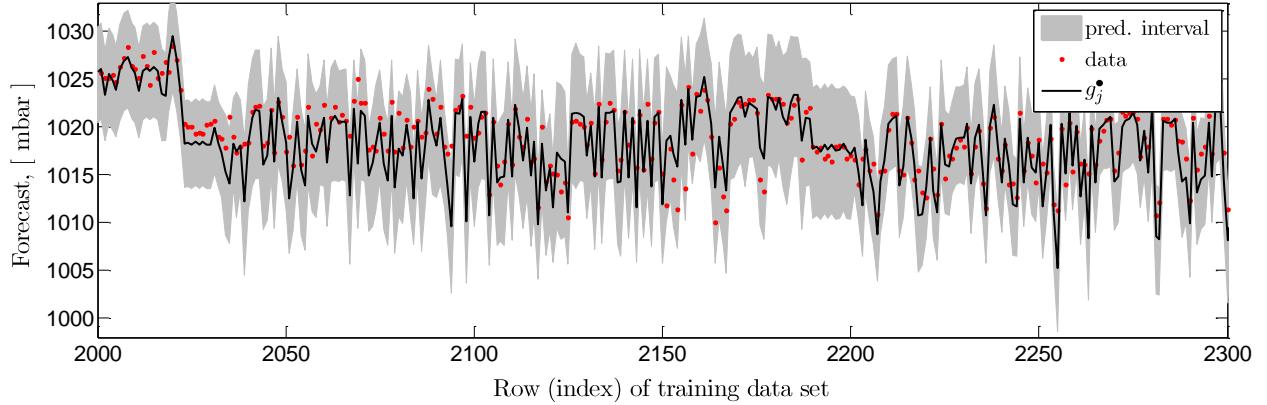


Figure 18: 95% prediction intervals (gray region) of the BMA predictive density for a representative portion of the 25-day training data set. The black line displays the mean point forecast of the BMA model derived from Equation (16), whereas the red dots signify the verifying observations.

The 95% prediction uncertainty ranges of the BMA model appear rather large with average spread of about 10.86 mbar and coverage of 94.36% of the observations. The RMSE of the BMA point forecast (mean of the mixture density) is approximately 2.80 mbar which is equivalent to the RMSE of the best model (NOGAPS) of the ensemble (see Table 8).

Table 8 lists the (maximum likelihood) values of the weights for each of the model averaging methods of the MODELAVG toolbox.

Table 8: Results of different model averaging methods for the five-member multimodel ensemble of 48-h forecasts of sea level pressure in the Pacific Northwest of the USA. The first two columns list the names of each model of the ensemble and their corresponding RMSE values (in mbar), respectively. Subsequent columns list the (maximum likelihood) values of the weights of each model averaging method of the toolbox. The bottom part of the Table lists the RMSE of the deterministic forecast of each method. This point forecast is calculated with Equation (4) using the ensemble forecasts of \mathbf{D} and maximum likelihood weights.

Model	RMSE	EWA	BGA	AICA [†]	BICA [†]	GRA	BMA	MMA	MMA ^{S †,‡}
AVN	2.8507	0.2000	0.2304	0.0000	0.0000	0.5402	0.2538	0.5048	0.2774
GEM	2.9769	0.2000	0.2112	0.0000	0.0000	0.3461	0.1834	0.3457	0.2038
ETA	3.4640	0.2000	0.1560	0.0000	0.0000	0.0005	0.0003	0.0606	0.0000
NGM	3.5664	0.2000	0.1472	0.0000	0.0000	-0.4903	0.0009	-0.5305	0.0000
NOGAPS	2.7084	0.2000	0.2552	1.0000	1.0000	0.6035	0.5616	0.6194	0.5188
Point	2.8734	2.7875	2.7084	2.7084	2.7084	2.4327	2.5731	2.4337	2.5714

†: I assume $p = 20 * \text{ones}(1, K)$ (twenty parameters assigned to each model)

‡: Presence of numerous local optima on likelihood surface

Information criterion averaging (AICA and BICA) assigns the NOGAPS model a weight of unity, whereas all other models are given a zero weight. The BMA method distributes the weights more equally among the different ensemble members, yet assigns almost zero weights to ETA and NGM. Discarding these two models hardly affects the predictive skill of the averaged model and (in case of BMA) the forecast density.

The GRA and MMA methods receive the lowest RMSE for their deterministic (point) forecasts. These are the only two methods that do not restrict the weights to lie on the unit simplex.

Finally, I follow *Vrugt et al.* (2008b) and benchmark the performance of the DREAM_(zs) algorithm by comparing the maximum likelihood estimates of this method against those derived separately using the EM algorithm. The function EM_NORMAL of Appendix A applies only to Gaussian forecast distributions, and hence we compare both methods using a normal forecast distribution for the members of the sea surface pressure ensemble. Figure 19 compares the log-likelihood estimates derived from the EM method (solid red line) and DREAM_(zs) (blue squares) for different lengths of the training data set (x-axis). I also compare the corresponding estimates of the spread (B) and coverage (C) of the 95% BMA model prediction ranges. Bias correction was applied to each training data set using simple linear regression of D_k on \tilde{y} of the training data set.

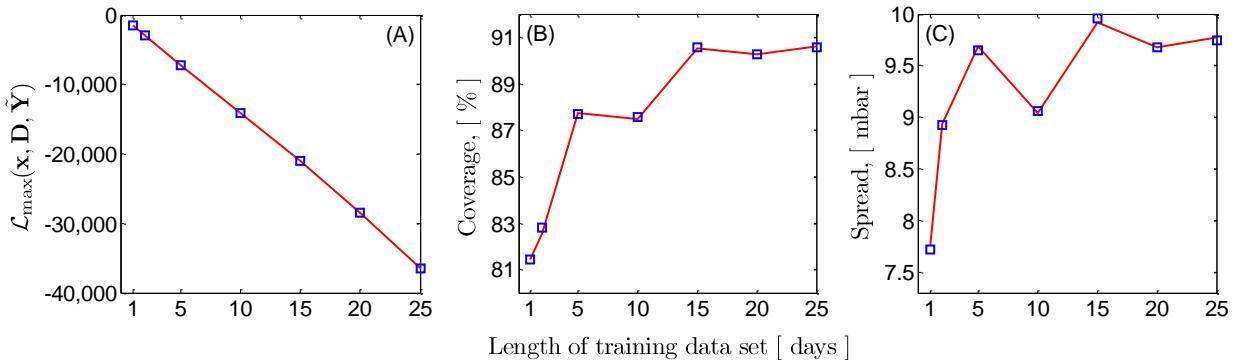


Figure 19: Maximum value of the log-likelihood of the EM (red) and DREAM_(zs) (blue) method (A) and corresponding coverage (B) and spread (C) of the BMA prediction intervals.

The panels plot only the results for the training data set. The results presented here highlight several important observations. First, the DREAM_(zs) algorithm and EM method derive exactly similar estimates of the log-likelihood value of the BMA model, irrespective of the length of the training data set. Secondly, the log-likelihood increases linearly with length of the training data set. This is simply the effect of the number of observations, n . Thirdly, the coverage of the 95% prediction intervals increases with length of the training data set. The same observation is made for the spread of the 95% prediction ranges. Both findings are not surprising as longer training data sets typically involve a larger diversity of weather events, requiring a larger standard deviation of the forecast distribution.

5. Recent developments

The original BMA approach presented by *Raftery et al.* (2005) assumes that the conditional pdf of each individual model is adequately described with a univariate normal or Gamma distribution, possibly with a heteroscedastic variance. Recently, I have added to this mix the generalized normal distribution. It is up to users to evaluate the merits of this "new" density. Preliminary numerical experiments with the built-in examples of the MODELAVG toolbox demonstrated a further increase in the log-likelihood of the BMA model compared to the use of a normal distribution with a constant or non-constant variance. I like to point

to the work of *Rings et al.* (2012) who introduced a variant of BMA that uses a flexible representation of the conditional forecast distribution. This approach uses a particle filter to derive, as closely and consistently as possible, the evolving forecast density (conditional pdf) of each constituent ensemble member. A mixture distribution is then fitted to the forecast distribution at each time, the weights of which are then determined from a training data period using maximization of the likelihood. The mixture distributions of each ensemble member are then combined to derive one overall predictive distribution. Benchmark studies demonstrate that this revised BMA method receives substantially lower forecast errors than the default BMA method (due to filtering) with associated prediction intervals that are considerably smaller (due to time-variant conditional pdf) but statistically coherent.

In the latest release (V2.0), I also added the lognormal, truncated normal and Weibull distributions as possible parametric forms for the conditional PDFs. The manual has been updated to reflect this addition. Furthermore, the user can also specify multiple different significance levels in field `alpha` of structure `options` (see Fig. 10). The corresponding prediction quantiles of the BMA model are returned in the field `pred` of structure `output`. Lower and upper prediction quantiles are computed almost exactly by evaluating repeatedly the CDF of the BMA model using an iterative root finding method. This implementation is presented in the `BMA_QUANTILE` function. Then, a new function `BMA_sample` draws samples from the BMA forecast density and returns analytic estimates of its mean and variance at each time. These new options are exercised in `example_12`. This example also computes so-called scoring rules of the BMA forecast density. Please refer to *Vrugt* (2023) for a detailed description of the computatoin of the scoring rules and their implementation for model evaluation and diagnostics.

6. Summary

In this manual I have introduced a MATLAB package, entitled MODELAVG, which provides interested users with a simple toolbox for postprocessing of forecast ensembles. This toolbox implements equal weight averaging, Bates-Granger averaging, information criterion averaging, Granger-Ramanathan averaging, Bayesian model averaging and Mallows model averaging. For those averaging methods for which an iterative solution is required to derive the weights and/or variance(s) of the conditional forecast distribution, MCMC simulation with the DREAM_(ZS) algorithm is used, and a sample of the posterior distribution is generated. Three different case studies were used to illustrate the main capabilities and functionalities of the MATLAB toolbox. These example studies are easy to run and adapt and serve as templates for other modeling problems and watershed data sets.

The toolbox allows for different formulations of the BMA conditional forecast distribution. Forecast densities that differ from a normal, lognormal, truncated normal, generalized normal, gamma or Weibull distribution are readily implemented in the source code of the MODELAVG toolbox by adding new "cases" to the MATLAB functions BMA_CALC and BMA_QUANTILE. In fact, one can further generalize the BMA method by allowing the ensemble members to differ in their parametric forms of the forecast PDF. The easiest way to do so is through an extension of the generalized normal distribution; fundamentally, each ensemble member can then have its own unique forecast PDF. This may sharpen further the BMA mixture distribution and, thus, reduce predictive uncertainty (the spread at different confidence levels, α). This would probably achieve better results if coupled with model calibration so that each member of the ensemble is tasked with specific aspects of the forecast distribution.

7. Acknowledgements

The MATLAB toolbox of MODELAVG is available upon request from the author, jasper@uci.edu. I would like to thank Guilherme Gomes for careful proofreading of the final document.

Appendix A. The Expectation-Maximization algorithm

The Expectation - Maximization (EM) algorithm is a broadly applicable approach to the iterative computation of maximum likelihood estimates, useful in a variety of incomplete-data problems (*Dempster et al.*, 1997; *McLachlan and Krihnan*, 2008). I use herein the index t to mean "for all $t \in (1, \dots, n)$ " and the index k to mean "for all $k \in (1, \dots, K)$ " and use the latent variable z_{tk} to help find the optimal values of the BMA weights and standard deviations. This unobserved quantity has a value of unity if ensemble member k is the best forecast of \tilde{y}_t and zero otherwise. For each observation of the training data set, only one of the (z_{t1}, \dots, z_{tK}) values is thus equal to one, and all the other values are zero.

The EM algorithm alternates between an expectation step and a maximization step until convergence is achieved. In the E-step, the values of z_{tk} are calculated given the current values of the BMA weights and variances. For the BMA model of Equation (13) and (15) the E-step is given by

$$\hat{z}_{tk}^{(j)} = \frac{\beta_k f_{\mathcal{N}}(\tilde{y}_t | d_{tk}, \sigma_k^{2(j-1)})}{\sum_{i=1}^K \beta_i f_{\mathcal{N}}(\tilde{y}_t | D_{ti}, \sigma_i^{2(j-1)})} \quad (\text{Expectation Step}), \quad (\text{A.1})$$

where the function $f_{\mathcal{N}}(a|b,c)$ returns the density at a of a normal distribution $\mathcal{N}(b,c)$ with mean b and variance c , and the superscript j signifies iteration counter. In the subsequent maximization or M-step, the values of β_k and σ_k^2 are updated using the current estimates of z_{tk} , i.e. $\hat{z}_{tk}^{(j)}$ as follows

$$\begin{aligned} \beta_k^{(j)} &= \frac{1}{n} \sum_{t=1}^n \hat{z}_{tk}^{(j)} \\ &\quad (\text{Maximization Step}) \quad (\text{A.2}) \\ \sigma_k^{2(j)} &= \frac{\sum_{t=1}^n \hat{z}_{tk}^{(j)} (\tilde{y}_t - d_{tk})^2}{n \sum_{t=1}^n \hat{z}_{tk}^{(j)}} \end{aligned}$$

where n denotes the number of observations of the training data set. By alternating between Equations (A.1) and (A.2) the EM algorithm improves iteratively the values of β_k and σ_k^2 . Convergence is achieved when the values of the likelihood (= denominator of Equation (A.1)), weights, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$, variances, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)^\top$ and $\hat{z}_{tk}^{(j)}$'s remain constant from one iteration to the next.

The function `EM_NORMAL` listed below implements the EM algorithm in MATLAB. This subroutine has three input arguments, including `D` (matrix with ensemble forecasts), `y` (calibration data vector), and structure `options` (for field `VAR` with variance option) and returns the maximum likelihood values of the BMA weights, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^\top$ and variance, $\boldsymbol{\sigma}^2$ (`options.VAR = '1'`) or variances, $\boldsymbol{\sigma}^2 = (\sigma_1^2, \dots, \sigma_K^2)^\top$ (`options.VAR = '2'`) and corresponding log-likelihood, $\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\sigma}^2 | \mathbf{D}, \tilde{\mathbf{y}})$ in Equation (21).

MATLAB code of the Expectation-Maximization algorithm: This function calculates the maximum likelihood values of the BMA weights, beta and standard deviations, sigma of each members normal forecast distribution for a given ($n \times K$) matrix of ensemble forecasts, and ($n \times 1$) vector of training observations. The third input argument, options is a structure with field VAR that determines whether to use a single common variance for all forecast distributions of the members of the BMA ensemble (options.VAR = '1') or a member specific variance (options.VAR = '2'). Notation and variable names are consistent with main text and vectorization is used to minimize somewhat the number of lines of the code. Built-in functions are highlighted with a low dash. The binary singleton expansion function bsxfun(FUNC, A, B) applies an element-by-element binary operation to the ($n \times K$) matrix A and ($n \times 1$) vector B using either a right array divide (FUNC = @rdivide) or a minus operation (FUNC = @minus). The function abs(X) calculates the absolute value of the elements of phi, and sum(X, dim) sums along the dimension dim (1: vertical, 2: horizontal).

```

function [ beta , sigma , loglik , j ] = EM_normal ( D , y , options )
% Expectation-Maximization method for training of BMA model
%
% -----
% SYNOPSIS      [ibeta,sigma,lik] = EM_normal(D,Y);          %
%               [ibeta,sigma,lik] = EM_normal(D,Y,options);        %
% INPUT         D          (n x K)-matrix of ensemble forecasts    %
%               Y          (n x 1)-vector with training data set      %
%               options     structure with settings                  %
% OUTPUT        beta       (1 x K)-vector of BMA weights           %
%               sigma      (1 x K)-vector of BMA standard deviations   %
%               loglik     log-likelihood value                      %
%
% -----
% MAIN ASSUMPTION: Gaussian forecast (= conditional) distribution
if nargin < 2
    error('EM_normal ERROR: Function EM_normal requires at least two input arguments');
end
if nargin < 3
    warning('EM_normal WARNING: Unknown variance option => Default options.VAR = ''1''');
    options.VAR = '1';
end
[n,K] = size(D);                                % Matrix of ensemble forecasts
Y = repmat(y,1,K);                            % Replicate training data observations
beta = ones(1,K)/K;                           % Initial weights (at random)
sigma = std(y)*ones(1,K); z_j = zeros(n,K);    % Initial standard devs + latent variables
loglik_j = -Inf; err = 1; j = 0; max_j = 1e4; % Settings/constraints while loop
while ( max(err) > 1e-6 ) && ( j < max_j )      % Until ... do
    loglik = loglik_j; z = z_j;                 % Copy z and loglik
    for k = 1:K
        z_j(:,k) = beta(k)*normpdf(y,D(:,k),sigma(k)); % EXPECTATION STEP
    end
    loglik_j = sum(log(sum(z_j,2)));            % Log-likelihood BMA model
    z_j = bsxfun(@rdivide,z_j,sum(z_j,2));      % Normalize latent variables
    beta_j = sum(z_j)/n;                         % MAXIMIZATION STEP
    sigma2_j = sum(z_j.* (Y-D).^2) ./ sum(z_j);
    if strcmp(options.VAR,'1')
        sigma2_j = mean(sigma2_j)*ones(1,K);
    end
    err(1) = max(abs(beta_j - beta));
    err(2) = max(abs(log(sigma2_j./sigma.^2)));
    err(3) = max(max(abs(z - z_j)));           50
    err(4) = max(abs(loglik - loglik_j));
    beta = beta_j; sigma = sqrt(sigma2_j);
    j = j + 1;
end
% End while loop

```

Appendix B. The DREAM_(zs) algorithm

The DREAM_(zs) algorithm is an efficient multi-chain MCMC simulation method that uses differential evolution as genetic algorithm for population evolution with a Metropolis selection rule to decide whether candidate points should replace their parents or not. In DREAM_(zs), N different Markov chains are run simultaneously in parallel. If the state of a single chain is given by the d -vector \mathbf{x} , then at each generation $t - 1$ the N chains in DREAM_(zs) define a population \mathbf{X} , which corresponds to an $N \times d$ matrix, with each chain as a row. If A is a subset of d^* -dimensions of the original parameter space, $\mathbb{R}^{d^*} \subseteq \mathbb{R}^d$, then a jump, \mathbf{SX}^i in the i^{th} chain, $i = (1, \dots, N)$ at iteration $t = (2, \dots, T)$ is calculated from a steadily growing archive \mathbf{Z} of m points sampled by the joint chains using differential evolution (*Storn and Price, 1997; Price et al., 2005*)

$$\begin{aligned}\Delta \mathbf{X}_A^i &= \zeta_{d^*} + (\mathbf{1}_{d^*} + \boldsymbol{\lambda}_{d^*}) \gamma_{(\delta, d^*)} \sum_{j=1}^{\delta} (\mathbf{Z}_A^{\mathbf{a}_j} - \mathbf{Z}_A^{\mathbf{b}_j}) \\ \Delta \mathbf{X}_{\neq A}^i &= 0,\end{aligned}\tag{B.1}$$

where $\gamma = 2.38/\sqrt{2\delta d^*}$ is the jump rate, δ denotes the number of chain pairs used to generate the jump, and \mathbf{a} and \mathbf{b} are vectors consisting of δ integers drawn without replacement from $(1, \dots, m)$. The default value of $\delta = 3$, and results, in practice, in one-third of the proposals being created with $\delta = 1$, another third with $\delta = 2$, and the remaining third using $\delta = 3$. The values of $\boldsymbol{\lambda}$ and ζ are sampled independently from $\mathcal{U}_{d^*}(-c, c)$ and $\mathcal{N}_{d^*}(0, c_*)$, respectively, the multivariate uniform and normal distribution with, typically, $c = 0.1$ and c_* small compared to the width of the target distribution, $c_* = 10^{-6}$ say. In 20% of the proposals, I use a jump rate of unity, $p_{(\gamma=1)} = 0.2$, to enable the chains of DREAM_(zs) to jump directly between disconnected posterior modes. The candidate point of chain i at iteration t then becomes

$$\mathbf{X}_p^i = \mathbf{X}^i + \Delta \mathbf{X}^i,\tag{B.2}$$

and the Metropolis ratio

$$p_{\text{acc}}(\mathbf{X}^i \rightarrow \mathbf{X}_p^i) = \min[1, p(\mathbf{X}_p^i)/p(\mathbf{X}^i)],\tag{B.3}$$

is used to determine whether to accept this proposal or not. If $p_{\text{acc}}(\mathbf{X}^i \rightarrow \mathbf{X}_p^i) \geq \mathcal{U}(0, 1)$ the candidate point is accepted and the i^{th} chain moves to the new position, that is $\mathbf{x}_t^i = \mathbf{X}_p^i$, otherwise $\mathbf{x}_t^i = \mathbf{x}_{t-1}^i$. The default equation for γ should, for Gaussian and Student target distribution, result in optimal acceptance rates close to 0.44 for $d = 1$, 0.28 for $d = 5$, and 0.23 for large d (please refer to section 7.84 of *Roberts and Casella (2004)* for a cautionary note on these references acceptance rates). The historical archive \mathbf{Z} is drawn from the prior parameter distribution, and appended periodically, at a fixed rate, with the latest states of the N chains. This approach does not satisfy detailed balance, but diminishing adaptation of \mathbf{Z} ensures convergence to the target distribution. A snooker update is included as well to diversify the candidate points. Details of this procedure can be found in several publications, most recently ?.

The d^* -members of the subset A are sampled from the entries $(1, \dots, d)$ (without replacement) and define the dimensions of the parameter space to be sampled by the proposal. This subspace spanned by A is construed in the DREAM_(zs) algorithm with the help of a crossover operator. This genetic operator is applied before each proposal is created and works as follows. First, a crossover value, cr is sampled from a geometric sequence of n_{CR} different crossover probabilities, $\mathbf{CR} = (\frac{1}{n_{\text{CR}}}, \frac{2}{n_{\text{CR}}}, \dots, 1)$ using the discrete multinomial

distribution, $\mathcal{M}(\mathbf{CR}, \mathbf{p}_{\text{CR}})$ on \mathbf{CR} with selection probabilities \mathbf{p}_{CR} . Then, a d -vector $\mathbf{z} = (z_1, \dots, z_d)$ is drawn from a standard multivariate normal distribution, $\mathbf{z} \stackrel{\mathcal{D}}{\sim} \mathcal{U}_d(0, 1)$. All those values j which satisfy $z_j \leq \text{cr}$ are stored in the subset A and span the subspace of the proposal that will be sampled using Equation (B.1). If A is an empty set then one dimension of $(1, \dots, d)$ will be sampled at random to avoid the jump vector to have zeros everywhere.

The use of a vector of crossover probabilities enables single-site Metropolis (A has one element), Metropolis-within-Gibbs (A has one or more elements) and regular Metropolis sampling (A has d elements), and constantly introduces new directions in the parameter space that chains can take outside the subspace spanned by their current positions. What is more, the use of subspace sampling allows for $N < d$, thereby reducing as much as possible the total number of function evaluations required for burn-in. Subspace sampling as implemented in DREAM_(ZS) adds one extra algorithmic variable, n_{CR} to the algorithm. The default setting of $n_{\text{CR}} = 3$ has shown to work well in practice, but larger values of this algorithmic variable might seem appropriate for high-dimensional target distributions, say $d > 50$, to preserve the frequency of low-dimensional jumps. Note, more intelligent subspace selection methods can be devised for target distributions involving many highly correlated parameters.

To enhance search efficiency the selection probability of each crossover value, stored in the n_{CR} -vector \mathbf{p}_{CR} , is tuned adaptively during burn-in by maximizing the distance traveled by each of the N chains. This adaptation is described in detail in *Vrugt et al. (2008a, 2009); Vrugt (2016)*, and a numerical implementation of this approach appears in the MATLAB code of the DREAM_(ZS) algorithm below.

The core of the DREAM_(ZS) algorithm can be written in about 50 lines of code (see MATLAB code below) and includes the function handles `prior` and `pdf` and the values of `N` (number of chains), `T` (number of iterations) and `d` (number of parameters).

MATLAB code of the DiffeRential Evolution Adaptive Metropolis (DREAM_(ZS)) algorithm. Based on input arguments `prior` (handle), `pdf` (handle), `N` (number of chains), `T` (number of generations) and `d` (number of parameters), the DREAM_(ZS) algorithm evolves N different trajectories simultaneously to produce a sample of the posterior target distribution. `prior` is an anonymous function that draws N samples from a d -variate prior distribution, and similarly `pdf` is a function handle which computes the posterior density of a proposal (candidate point). The output arguments `phi` and `p_x` store the sampled Markov chain trajectories and corresponding density values, respectively. Built-in functions are highlighted with a low dash. The jump vector, $\Delta X(i, 1:d)$ of the i^{th} chain contains the desired information about the scale and orientation of the proposal distribution and is derived from the historical archive `Z` of states visited by the N chains. `deal()` assigns default values to the algorithmic variables of DREAM_(ZS) and `sum()` computes the sum of the columns `A` of the chain pairs `a` and `b`.

```

function [x,p_x,Z] = dream_zs(prior, pdf, N, T, d)
% Differential Evolution Adaptive Metropolis with sampling from past archive and snooker update

[delta,c,c_star,n_CR,p_g,k,p_s] = deal(1,0.01,1e-12,3,0.2,10,0.1); % Default of algorithmic parameters
m0 = max(N,20*d); n_d = 3*delta;                                     % DREAM_ZS algorithmic variables
CR = (1:n_CR)/n_CR; p_CR = ones(1,n_CR)/n_CR;                         % Crossover values and select. prob
x = nan(T,d,N); p_x = nan(T,N);                                         % Preallocate chains and density
Z = prior(m0,d);                                                       % Create initial collection of points

for i = m0-N+1:m0, Z(i,d+1) = pdf(Z(i,1:d)); end
X = Z(m0-N+1:m0,1:d); p_X = Z(m0-N+1:m0,d+1); m = m0;
x(1:d,1:N) = reshape(X',1,d,N); p_x(1,1:N) = p_X';

for t = 2:T      % Dynamic part: Evolution of N chains
    dx = zeros(N,d);
    lambda = unifrnd(-c,c,N,1);
    R = randsample(1:m,N*n_d,'false'); R = reshape(R,n_d,N);
    method = randsample({'parallel' 'snooker'},1,'true',[1-p_s p_s]);
    alfa_sn = ones(N,1);
    for i = 1:N
        D = randsample(1:delta,1,'true');
        a = R(1:D,i); b = R(D+1:2*D,i); c_sn = R(2*D+1:3*D,i);
        if strcmp(method,'parallel')
            id = randsample(1:n_CR,1,'true',p_CR);
            z = rand(1,d);
            A = find(z < CR(id));
            d_star = numel(A);
            if d_star == 0, [~,A] = min(z); d_star = 1; end
            gamma_d = 2.38/sqrt(2+d_star);
            g = randsample([gamma_d 1],1,'true',[1-p_g p_g]);
            dx(i,A) = c_star*randn(1,d_star) + ...
                (1+lambda(i))*g*g*sum(Z(a,A)-Z(b,A),1);
        elseif strcmp(method,'snooker')
            F = X(i,1:d)-Z(a,1:d); D_e = F*F' + realmint;
            zP = F*(sum((Z(b,1:d)-Z(c_sn,1:d)).*F)/D_e);
            g = 1.2 + rand;
            dx(i,1:d) = c_star*randn(1,d) + (1+lambda(i))*g*zP;
        end
        Xp(i,1:d) = X(i,1:d) + dx(i,1:d);
    end
    if strcmp(method,'snooker')
        alfa_sn = (sum((Xp-Z(R(1,1:N),1:d)).^2,2) ./ sum((X - ...
            Z(R(1,1:N),1:d)).^2,2)).^(d-1)/2;
    end
    for i = 1:N
        p_Xp = pdf(Xp(i,1:d));
        alfa = p_Xp/p_X(i,1);
        p_acc = min(1,alfa_sn(i)*alfa);
        if p_acc > rand
            X(i,1:d) = Xp(i,1:d); p_X(i,1) = p_Xp;
        end
    end
    x(t,1:d,1:N) = reshape(X',1,d,N); p_x(t,1:N) = p_X';
    if (mod(t,k) == 0), Z(m+1:m+N,1:d+1) = [X p_X]; m = m + N; end
end
% End dynamic part

```

The MATLAB code listed above implements the different steps of the DREAM_(ZS) algorithm as detailed in the main text of this Appendix. Variable names correspond with their symbols used in Equations (B.1)

and (B.2). Indents and comments are used to enhance readability and to convey the main intent of each line of code. The computational efficiency of this code can be improved considerably, for instance through vectorization of the inner for loop, but this will affect negatively readability. Note that this basic code of DREAM_(ZS) does not monitor convergence of the sampled chain trajectories nor does it adapt the selection probabilities of the n_{CR} crossover values.

For those proficient in statistics, computer coding and numerical computation, the DREAM_(ZS) code listed above will be sufficient to solve for the posterior distribution of the BMA, MMA, and MMA^S parameters. Yet, for other users this code might not suffice as it has very few built-in options and capabilities. I therefore refer to the MATLAB toolbox of the DREAM_(ZS) algorithm described in *Vrugt (2016)*.

Appendix C. Download and installation

The MODELAVG code can be downloaded from my website at the following link <http://faculty.sites.uci.edu/jasper/software>. Please scroll down to the appropriate link (under MODEL AVERAGING) and save the file called "MATLAB-pCode-MODELAVG-V1.0" to your hard disk, for instance, in the directory "D:\Downloads\Toolboxes\ MATLAB\MODELAVG". Now open Windows explorer in this directory (see Figure C1).

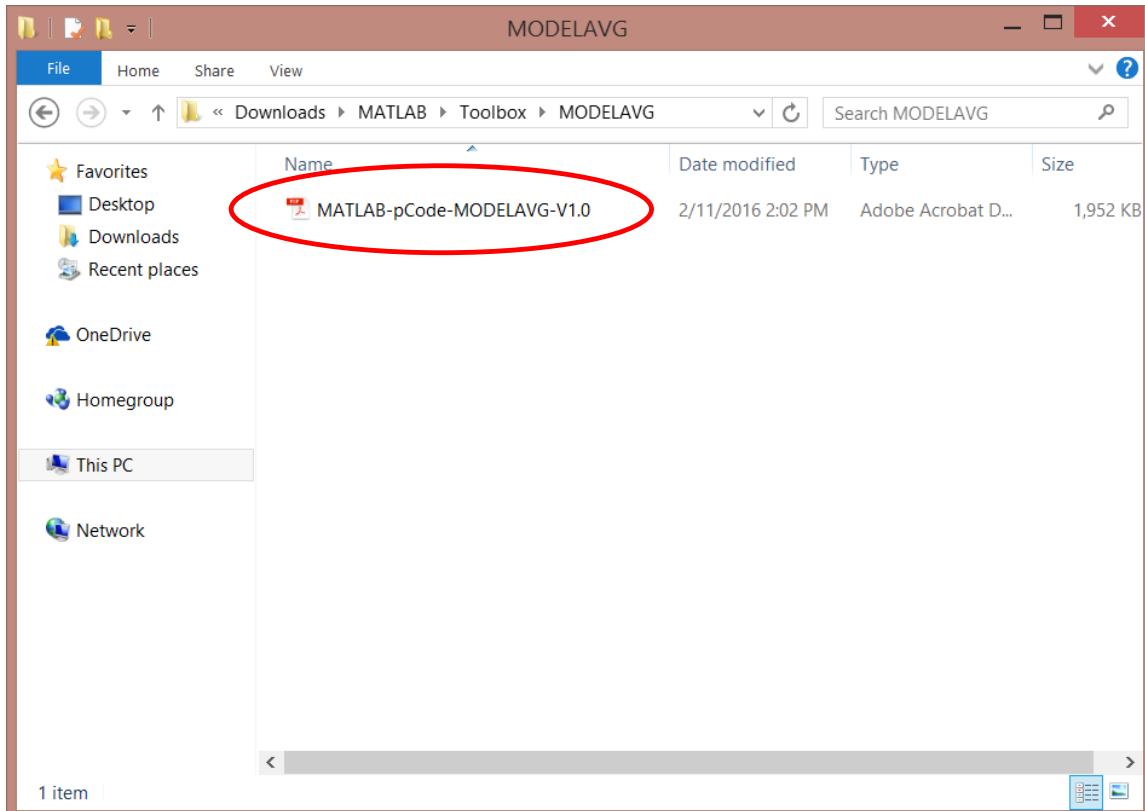


Figure C1

You will notice that the file does not have an extension - it is just called **MATLAB-pCode-MODELAVGV1.0**. That is because Windows typically hides extension names.

If you can already see file extensions on your computer, then please skip the next step. If you cannot see the file extension, please click the **View** tab. Then check the box titled "File name extensions" (see Figure C2).

MODELAVG MANUAL

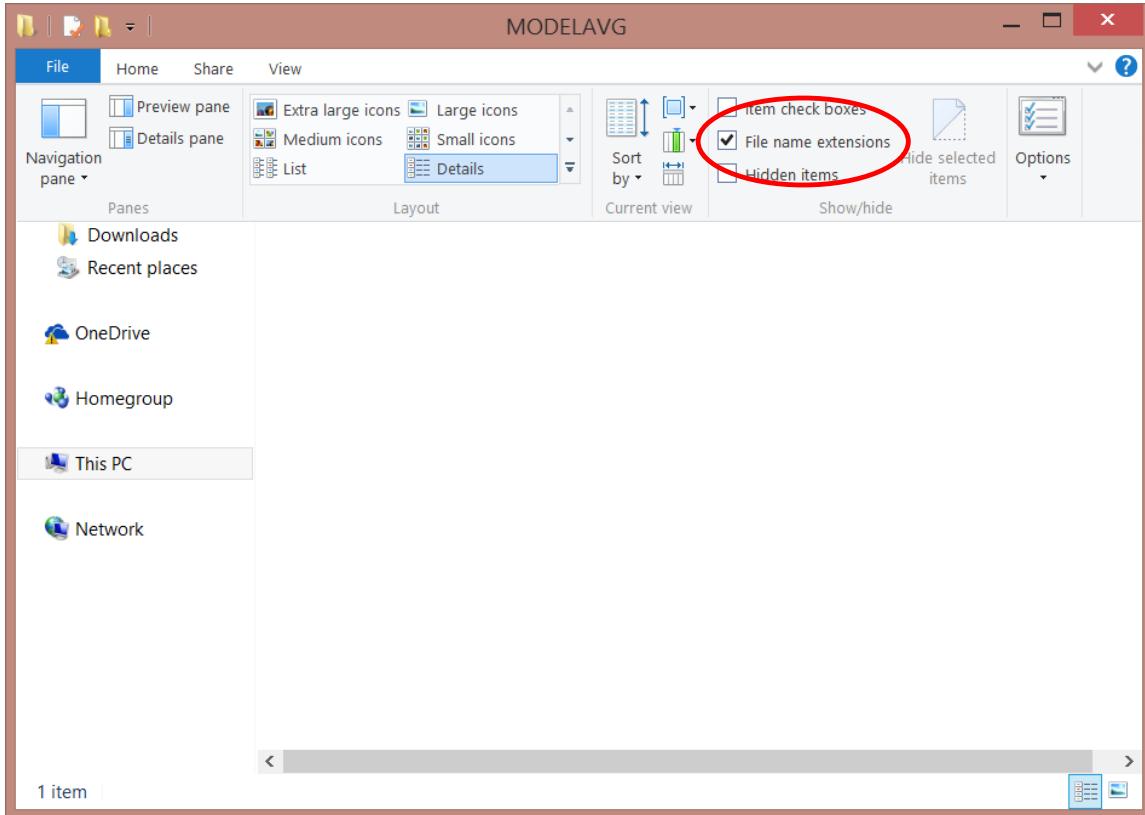


Figure C2

Now you should be able to see the file extension. Right-click the file name and select **Rename** (see Figure C3).

MODELAVG MANUAL

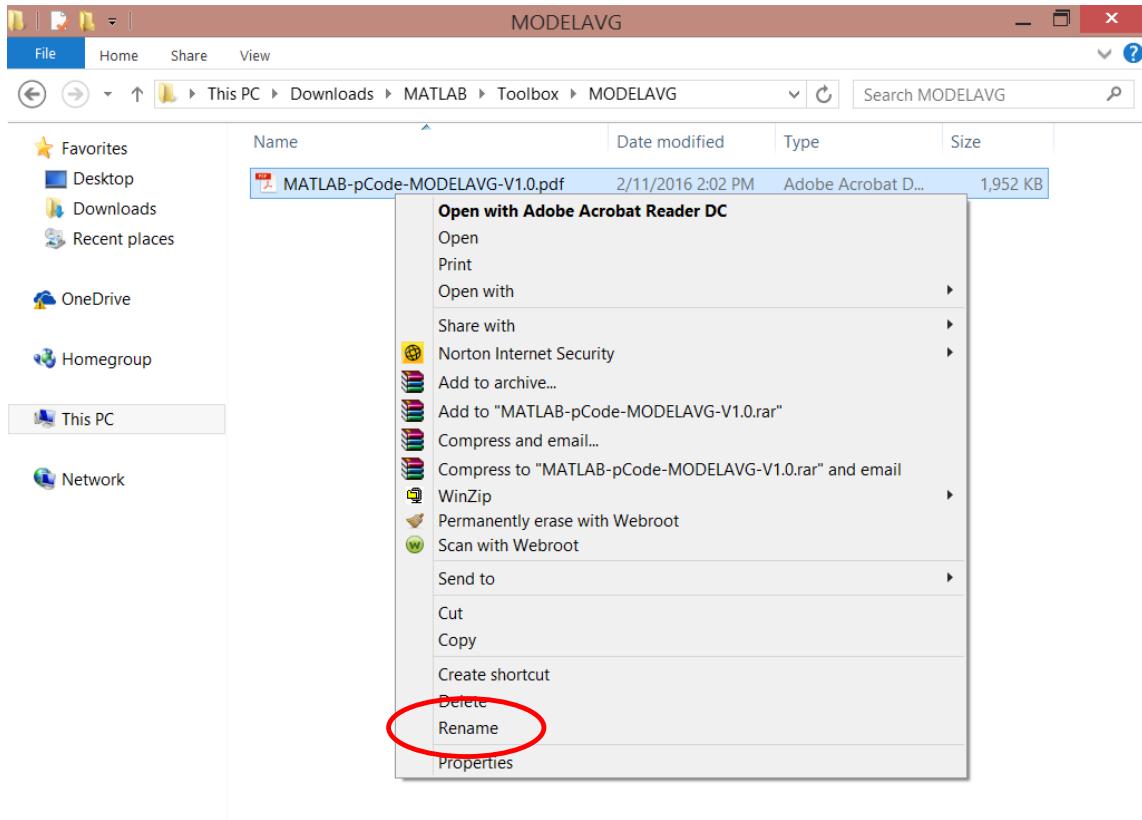


Figure C3

Now change the extension of "MATLAB-pCode-MODELAVG-V1.0" from ".pdf" to ".rar" (see Figure C4).

MODELAVG MANUAL

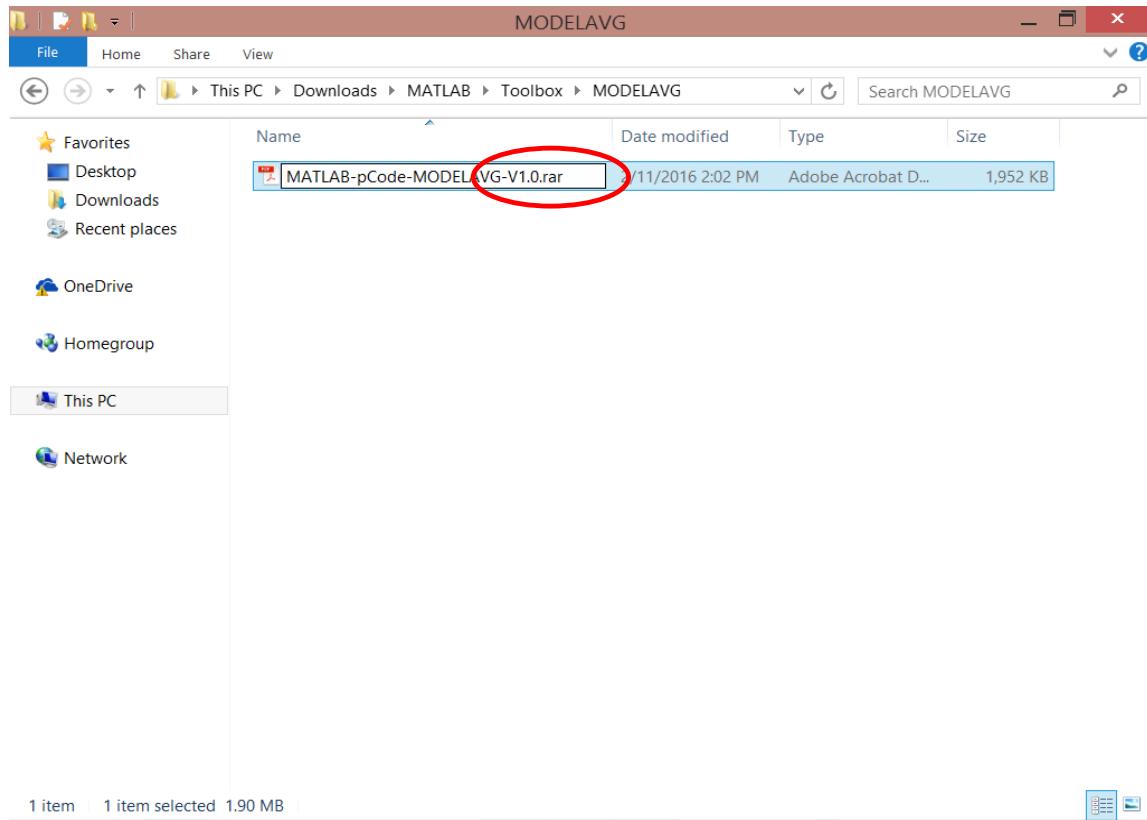


Figure C4

After entering the new extension, hit the **Enter** (return) key. Windows will give you a warning that the file may not work properly (see Figure C5). This is quite safe - remember that you can restore the original extension if anything goes wrong.

MODELAVG MANUAL

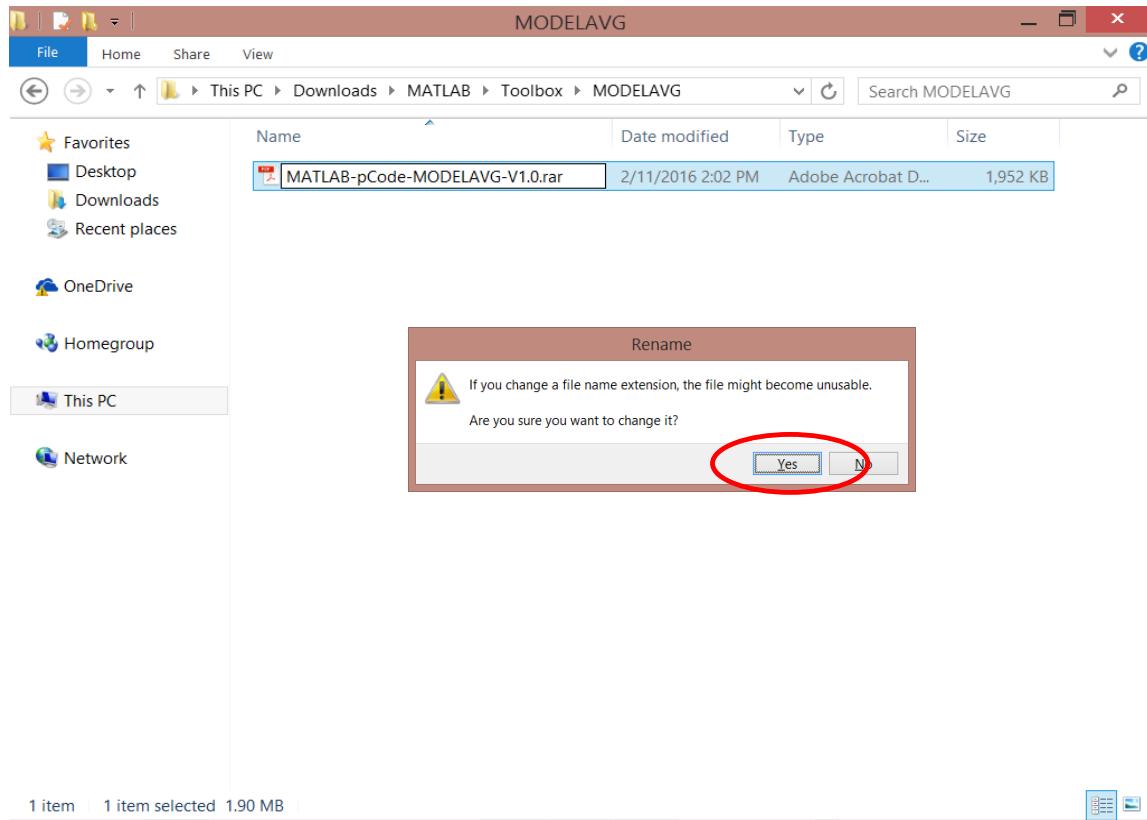


Figure C5

It is also possible that you might get another message telling you that the file is "read-only". In this case either say yes to turning off read-only, or right-click the file, select **Properties** and uncheck the **Read-only** box.

If you do not have permission to change the file extension, you may have to login as Administrator. Another option is to make a copy of the file, rename the copy and then delete the original.

Now you have changed the extension of the file to ".rar" you can use the program WinRAR to extract the files to whatever folder your desire, for instance "D:\Downloads\Toolboxes\MATLAB\MODELAVG". Right-click the file name and select **Extract Here** (see Figure C6).

MODELAVG MANUAL

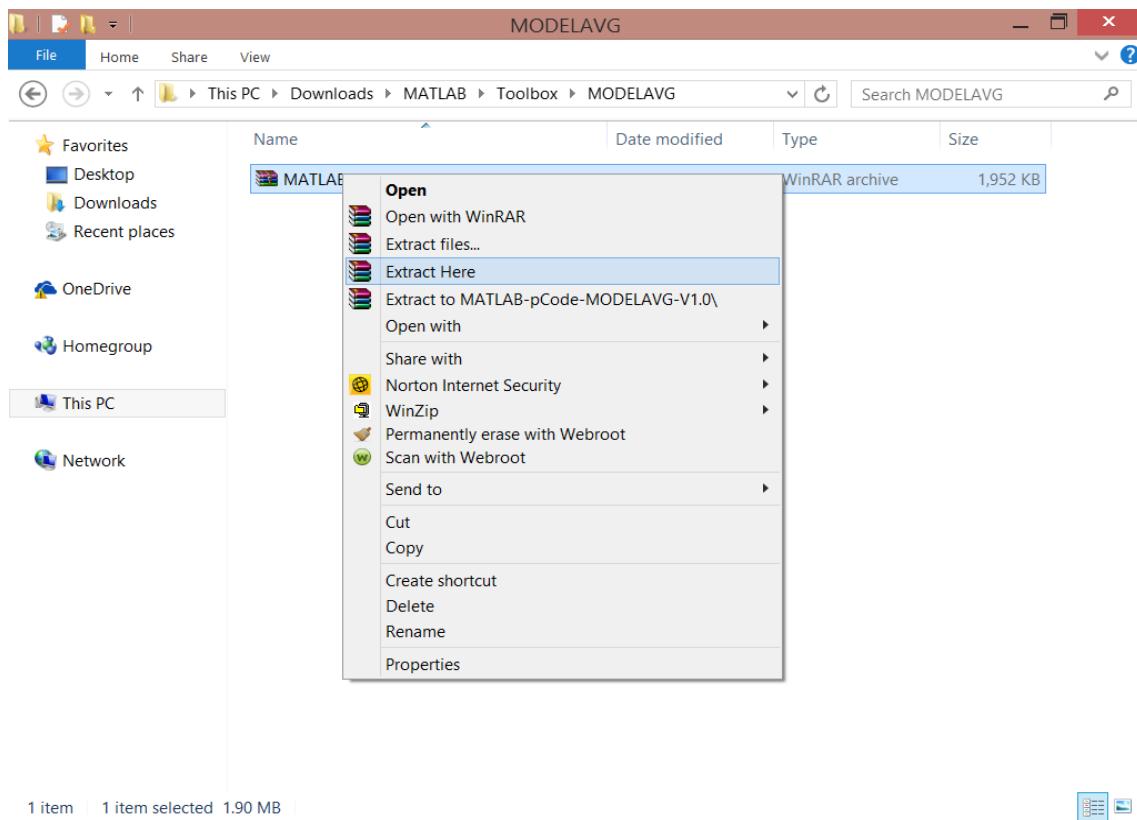


Figure C6

Now WinRAR should extract the files to your folder. The end result should look as in Figure C7.

MODELAVG MANUAL

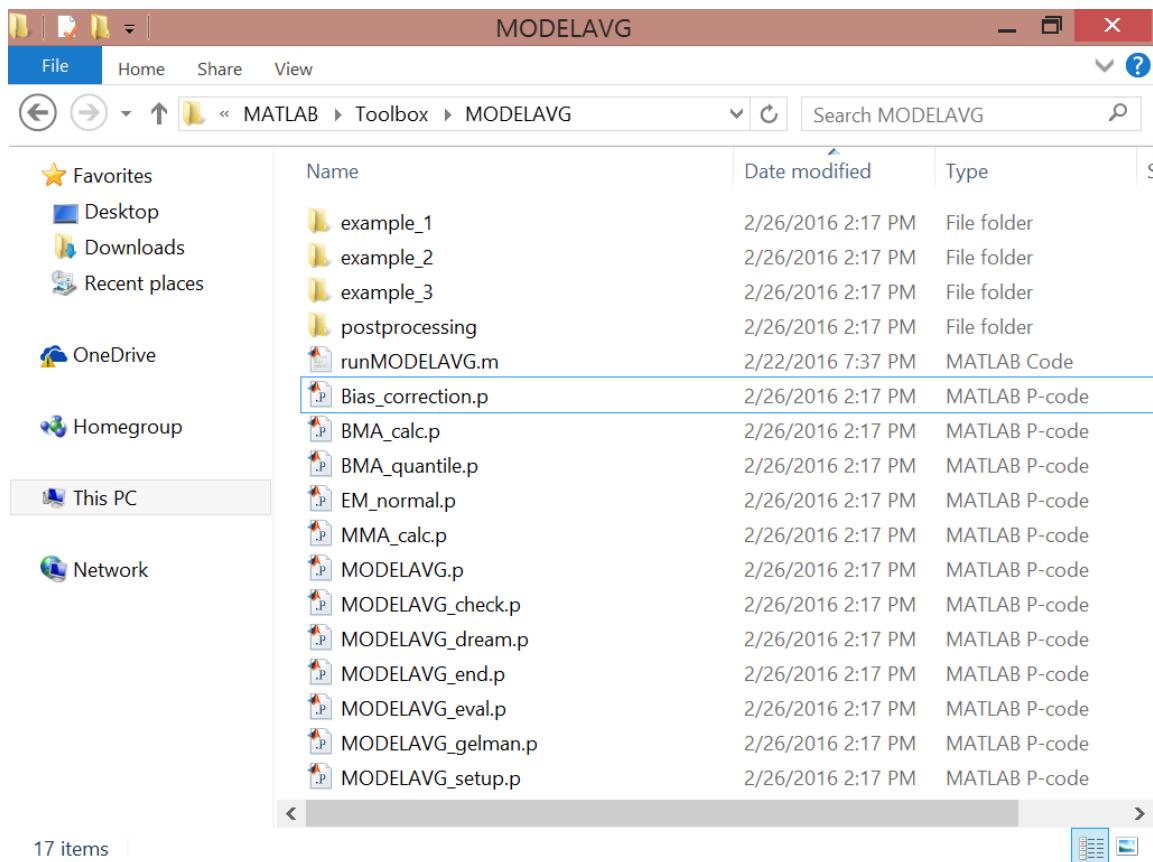


Figure C7

As last step, please open MATLAB and go to the appropriate directory with the MODELAVG files, for instance "D:\Downloads\Toolboxes\MATLAB\MODELAVG". Now execute the following statement in the MATLAB prompt: `addpath(pwd,[pwd '/postprocessing'])`. By adding the main FDCFIT and postprocessing directories to the search path, the user can execute the toolbox from any other directory. Now the MODELAVG toolbox is ready for use. If you want to execute the first built-in case study, then please change the current working directory of MATLAB to "...\\example_1" via `cd example_1` or `cd('D:\Downloads\Toolboxes\MATLAB\MODELAVG\example_1')`, and then type in the MATLAB prompt: `example_1`.

Appendix D. Main functions of MODELAVG toolbox

Table D1 summarizes, in alphabetic order, the different files of the MODELAVG package. The main program MODELAVG is called on the last line of each of the "example_X.m" input files on Pages 32lstlisting.-4, 37lstlisting.-5 and 42lstlisting.-6, where $X = \{1, \dots, 3\}$ and returns to the user the values of the weights (all methods except BMA) or weights and standard deviations (or proxies thereof) of the forecast density (if BMA is used). What is more, a structure `output` with results of each model averaging method is produced, and tables and figures are printed to the screen if the print option in field `print` of structure `options` is activated and thus set to 'yes'. The data of each case study are stored in their respective folders in the root directory of the MODELAVG toolbox (see Figure C7 for a screen shot of these directories). Graphical output is suppressed if the field `print` of structure `options` = 'no'.

The function MODELAVG_DREAM_ZS contains a basic implementation of the DREAM_(ZS) algorithm. This function is used to derive the posterior distribution of the BMA, MMA, and MMA^S weights and standard deviation(s) of the members' forecast distribution (BMA). Users are referred to the DREAM toolbox of *Vrugt* (2016).

The folder "postprocessing" in the root directory of the MODELAVG toolbox contains the script MODELAVG_POSTPROC which tabulates and visualizes the results of each model averaging method. The table is written to the file "MODELAVG_output.txt" and opened automatically in the MATLAB editor (see Appendix E) at termination of the code. the end of calculation. whereas figures are printed directly to the screen, including a time series plot of the ensemble members, the verifying observations and the averaged forecast, an autocorrelation function and a quantile-quantile graph of the error residuals of this point predictor. If BMA, MMA or MMA^S are used, many more figures are created from the DREAM_(ZS) output including trace plots of the sampled chain trajectories and \hat{R} -convergence diagnostic, and histograms of the marginal distributions of the parameters sampled by DREAM_(ZS) posterior samples (among others). Appendix E presents the screen output of the second case study produced by the function MODELAVG_postproc.

MODELAVG MANUAL

Table D1: Description of the MATLAB functions and scripts (.m files) used by MODELAVG, version 1.0.

Name of function	Description
BIAS_CORRECTION	Applies linear bias correction of each member of ensemble
BMA_CALC	Calculates the log-likelihood of BMA model parameters
BMA_QUANTILE	Computes the prediction intervals of BMA mixture distribution
EM_NORMAL	Expectation maximization algorithm for BMA model training
MMA_CALC	Calculates the log-likelihood of the MMA weights
MODELAVG	Main function of the toolbox - returns output arguments phi and structure output
MODELAVG_CHECK	Verifies input arguments of MODELAVG toolbox
MODELAVG_DREAM_ZS	Basic implementation of the DREAM(zs) algorithm for BMA and MMA model training
MODELAVG_END	Prepares graphical output and return arguments of MODELAVG toolbox
MODELAVG_EVAL	Calculates statistics of independent evaluation data set
MODELAVG_GELMAN	Calculates the \hat{R}^d and \hat{R} -convergence diagnostics
MODELAVG_SETUP	Setup of computational framework of MODELAVG toolbox
README.TXT	Text file (ascii format) with details how to setup and use the MODELAVG toolbox

Appendix E. Screen output

The MODELAVG toolbox presented herein returns to the user tables and figures which jointly summarize the results of the toolbox. This appendix displays all this output for the second case study involving application of the BMA method to the five-member ensemble of temperature forecasts in Pacific Northwest of the USA. A normal forecast distribution was assumed for each model of the ensemble. The standard deviation of this distribution was assumed to be constant, yet member-dependent.

Figure E1 displays the ascii file "MODELAVG_output.txt" which is created by the function MODELAVG_POSTPROC of the toolbox and printed automatically to the MATLAB editor after completion of the calculations.

```
Editor - D:\Algorithm\Own_Algorithm\MODELAVG\example_2\MODELAVG_output.txt

EDITOR      VIEW
File  Edit  Find  Files  Compare  Go To  Comment  Breakpoints
New  Open  Save  Close  Print  Find  Find  Insert  Indent  Outdent
ALMAGAM_output.txt  FDCCFIT_plot.m  MODELAVG_postproc.m  example_2.m  warning_file.txt  MODELAVG_output.txt  +
1 Model averaging method: BMA with normal conditional pdf (check manual)
2 =====
3 MODEL  PARAMETER    OPT   STD    COMPLEXITY  RMSE
4 =====
5 1  beta_1    +0.391  +0.019  +3.058
6 2  beta_2    +0.231  +0.015  +3.343
7 3  beta_3    +0.283  +0.018  +3.114
8 4  beta_4    +0.036  +0.013  +3.188
9 5  beta_5    +0.059  +0.011  +3.402
10 =====
11 - sigma_1   +2.454  +0.116
12 - sigma_2   +3.389  +0.167
13 - sigma_3   +1.916  +0.100
14 - sigma_4   +6.176  +1.026
15 - sigma_5   +1.587  +1.322
16 =====
17
18 =====
19 METHOD  RMSE  R
20 =====
21 BMA  +2.359  +0.861
22 =====
23
24 =====
25 BMA CONDITIONAL PDF  Coverage  Spread  Log-likelihood
26 =====
27 Normal Distribution  +90.736  +49.741  -34747.623
28 =====
29
30 ====== CORRELATION COEFFICIENTS ======
31 beta_1  beta_2  beta_3  beta_4  beta_5  sigma_1  sigma_2  sigma_3  sigma_4  sigma_5
32 beta_1  1.00  -0.43  -0.39  -0.23  -0.26  0.49  0.32  -0.27  -0.12  -0.04
33 beta_2  -0.43  1.00  -0.31  0.14  -0.28  -0.26  -0.08  -0.02  -0.03  -0.10
34 beta_3  -0.39  -0.31  1.00  -0.46  -0.46  -0.38  -0.03  0.57  0.07  -0.09
35 beta_4  -0.23  0.14  -0.46  1.00  -0.22  -0.09  -0.00  -0.49  0.11  -0.42
36 beta_5  -0.26  -0.28  0.04  -0.22  1.00  -0.17  -0.41  0.08  0.22  -0.32
37 sigma_1  0.49  -0.26  -0.35  0.20  -0.17  1.00  0.47  -0.63  -0.01  -0.56
38 sigma_2  0.32  -0.08  -0.01  -0.00  -0.41  0.47  1.00  -0.17  -0.03  -0.58
39 sigma_3  -0.27  -0.02  0.55  -0.39  0.08  -0.63  -0.17  1.00  0.03  0.35
40 sigma_4  -0.12  -0.03  -0.07  0.11  0.22  -0.01  -0.03  0.03  1.00  -0.40
41 sigma_5  -0.04  -0.10  0.39  -0.62  0.32  -0.56  -0.58  0.35  -0.40  1.00
```

Figure E1: Screen copy of ascii file "MODELAVG_output.txt" which appears in the MATLAB editor at the end of all calculations. Notation in this table matches exactly the names of the variables used in the Equations and main text.

The toolbox also presents to the user a large number of figures that visualize the results. I now present all these figures, two per page.

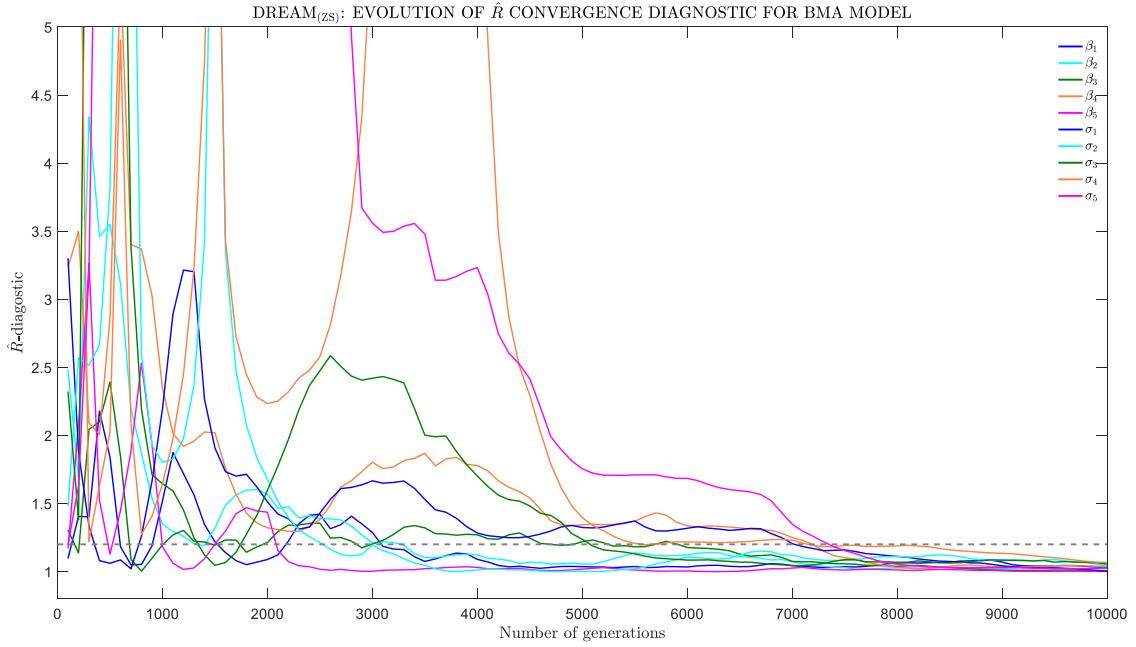


Figure E2: Traceplot of \hat{R} convergence diagnostic of *Gelman and Rubin* (1992) for the BMA weights and BMA standard deviations of each members' normal forecast distribution.

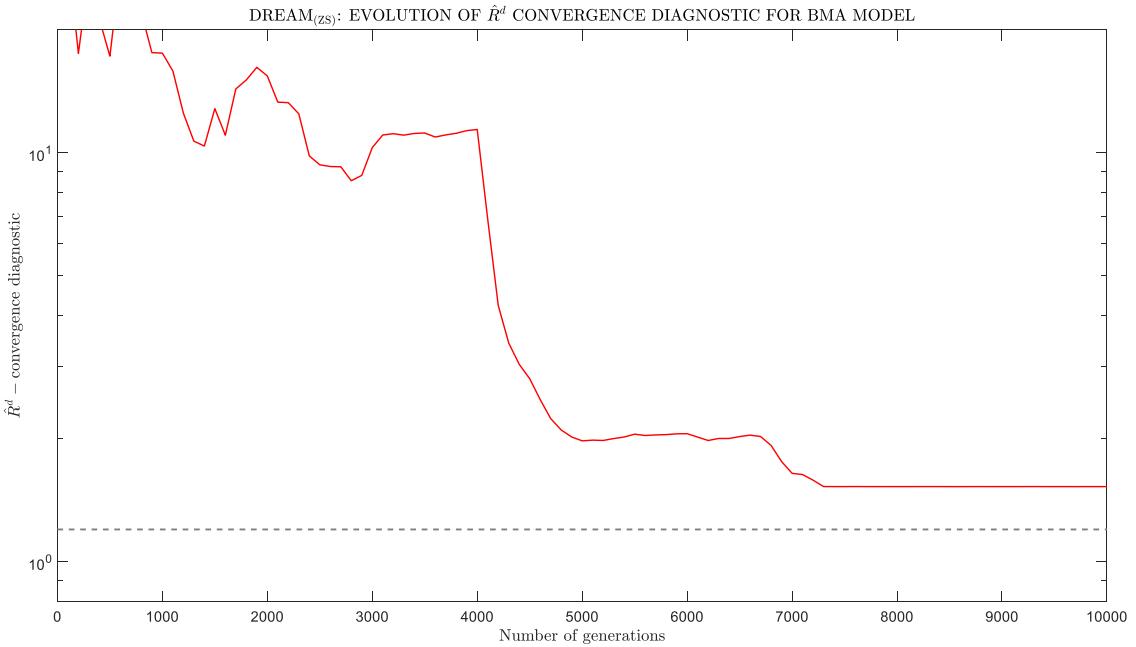


Figure E3: Traceplot of \hat{R}^d convergence diagnostic of *Brooks and Gelman* (1998) for the BMA model parameters.

MODEL AVG MANUAL

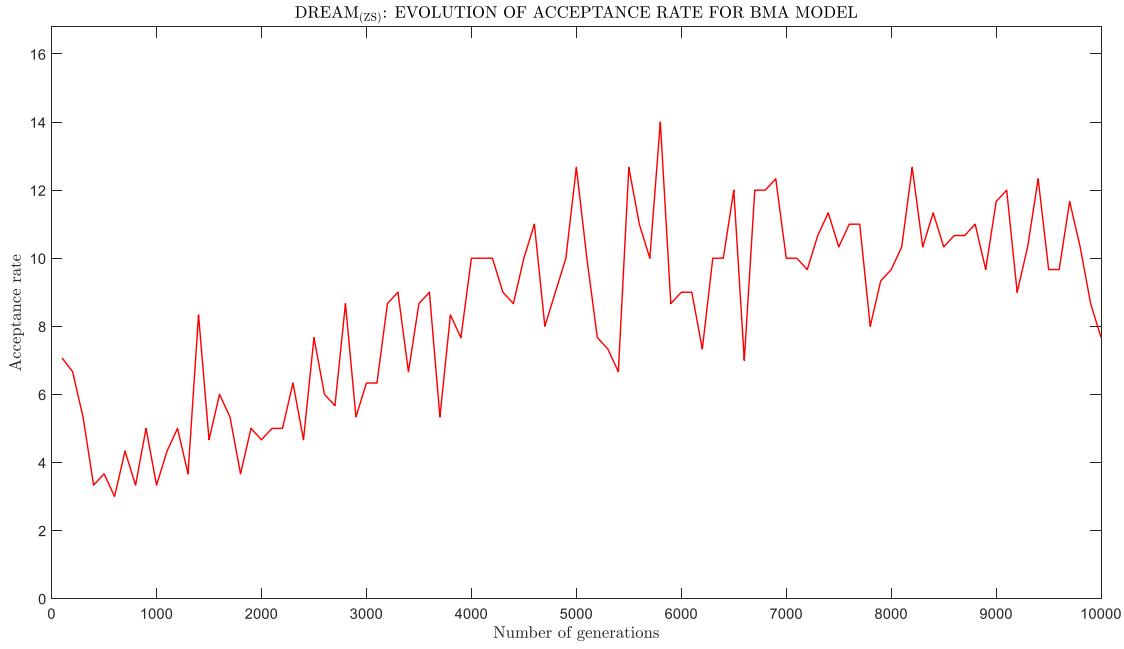


Figure E4: Evolution of acceptance rate of the DREAM_(ZS) sampled candidate points.

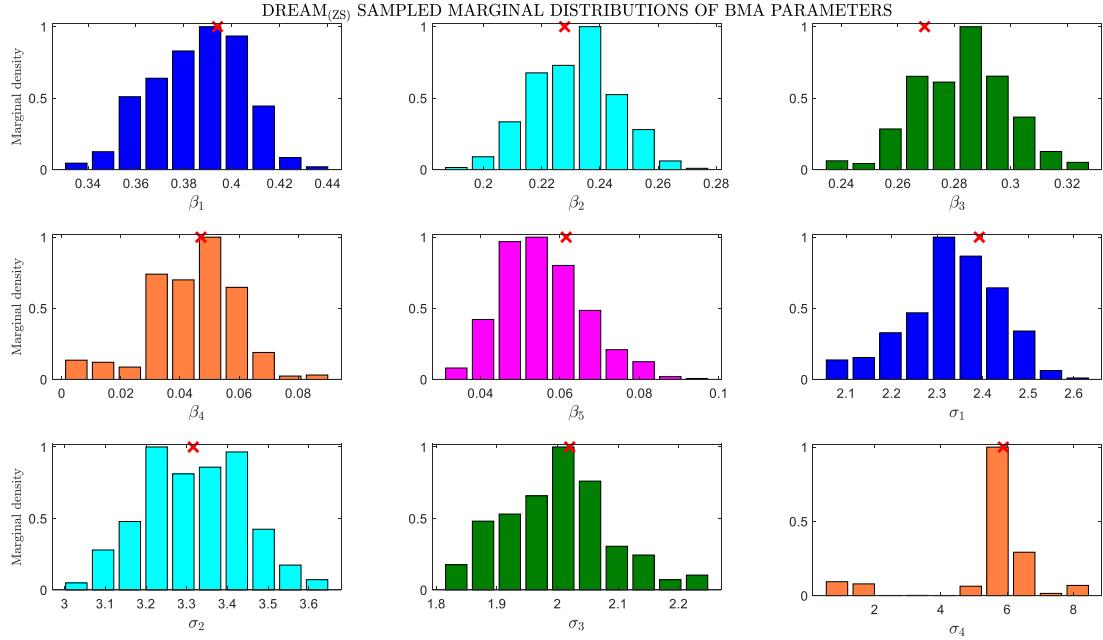


Figure E5: Histograms of the DREAM_(ZS) sampled marginal posterior distributions of the BMA model parameters.

MODEL AVG MANUAL

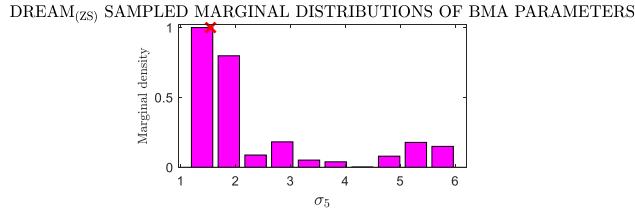


Figure E6: Histogram of the DREAM_(ZS) sampled marginal posterior distribution of BMA model parameter, σ_5 .

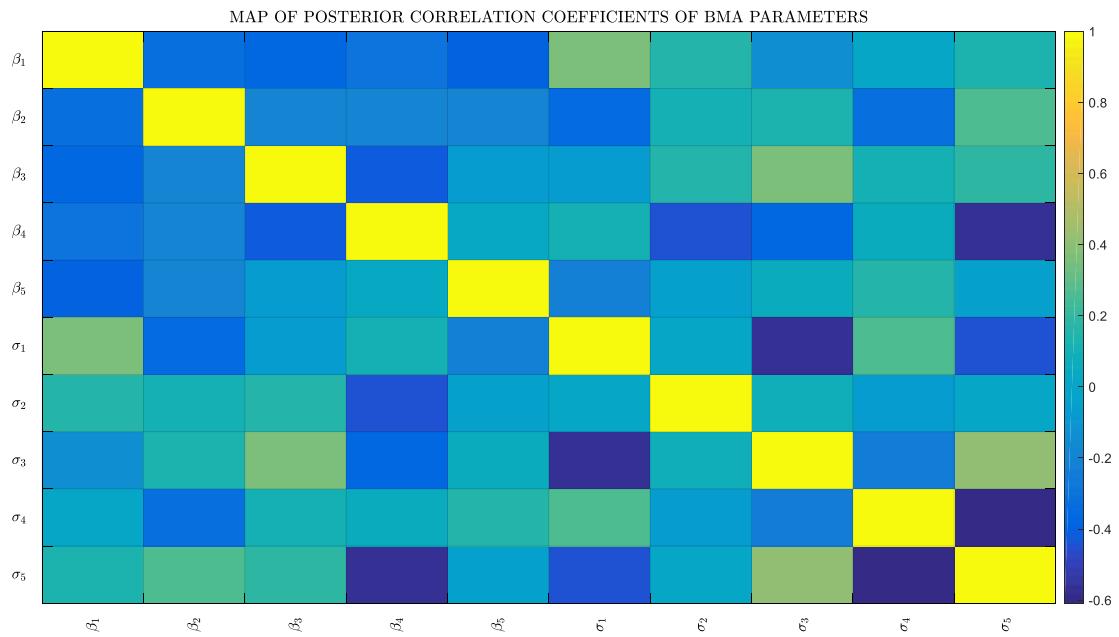


Figure E7: Map of the posterior correlation coefficients of the BMA model parameters.

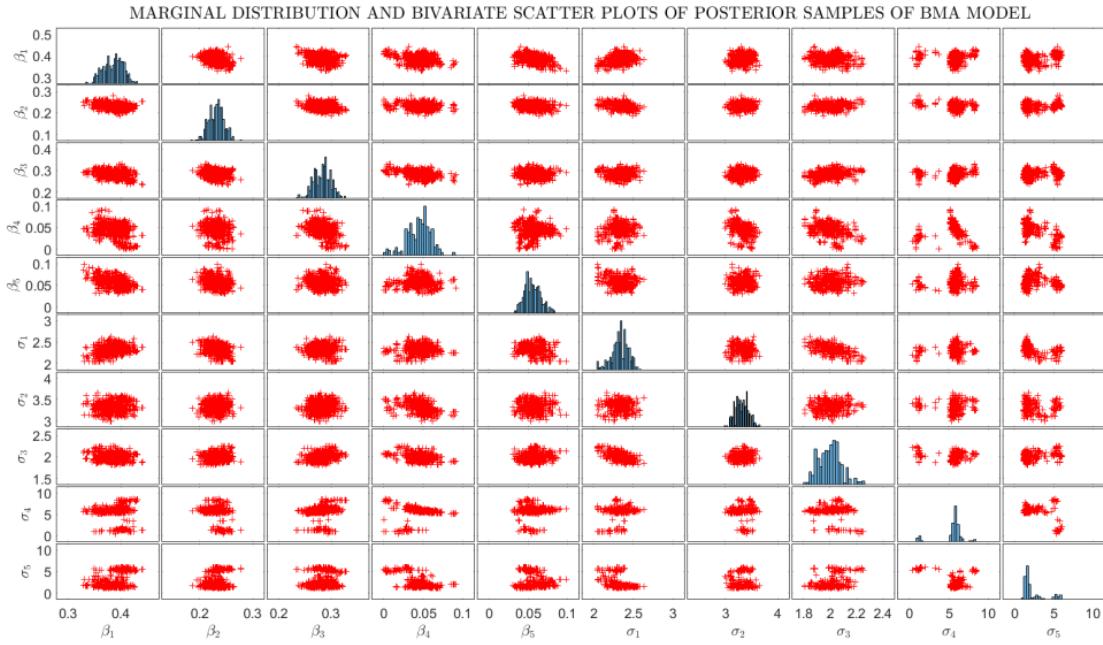


Figure E8: Scatter plot matrix of the DREAM_(ZS) sampled posterior solutions. The main diagonal displays histograms of the marginal distribution of each individual BMA parameter, whereas the off-diagonal graphs display bivariate scatter plots of the posterior samples.

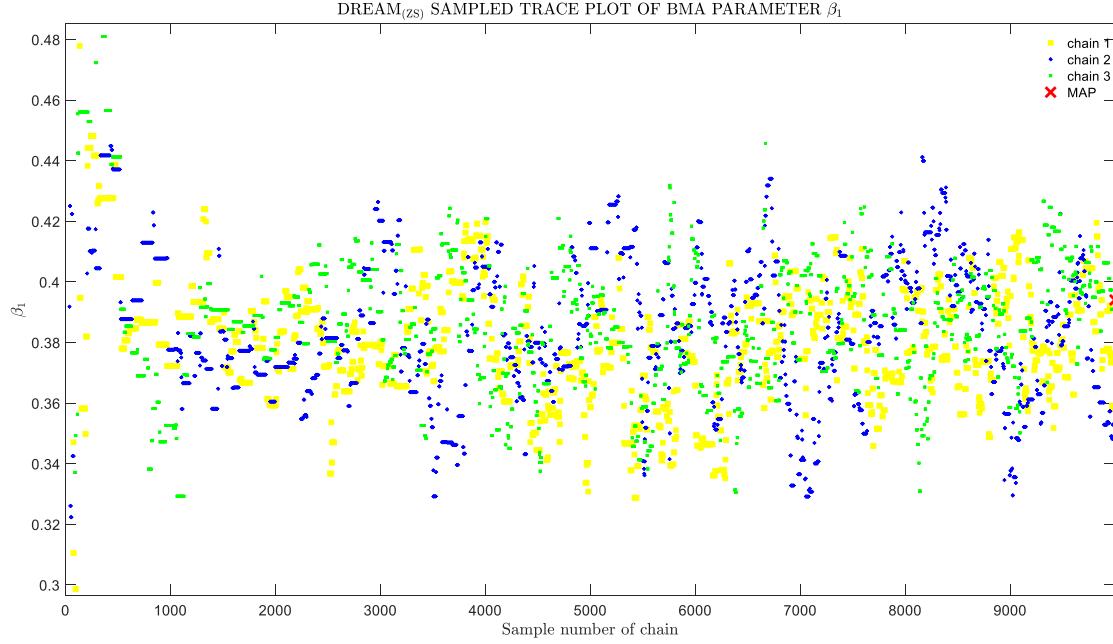


Figure E9: Traceplot of the sampled values of β_1 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

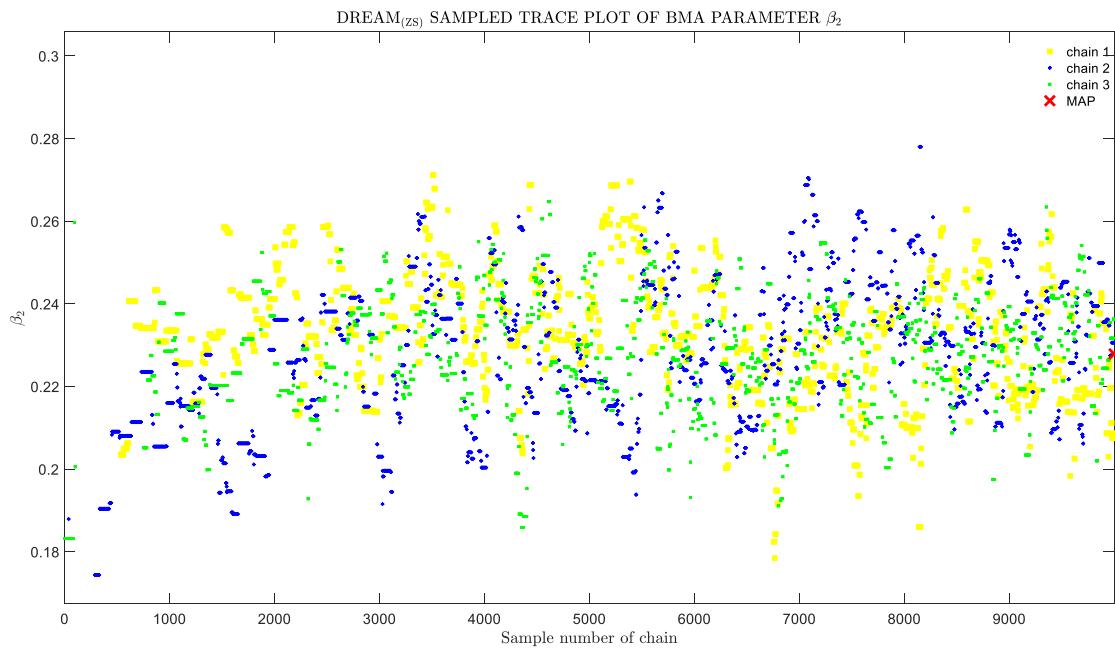


Figure E10: Traceplot of the sampled values of β_2 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

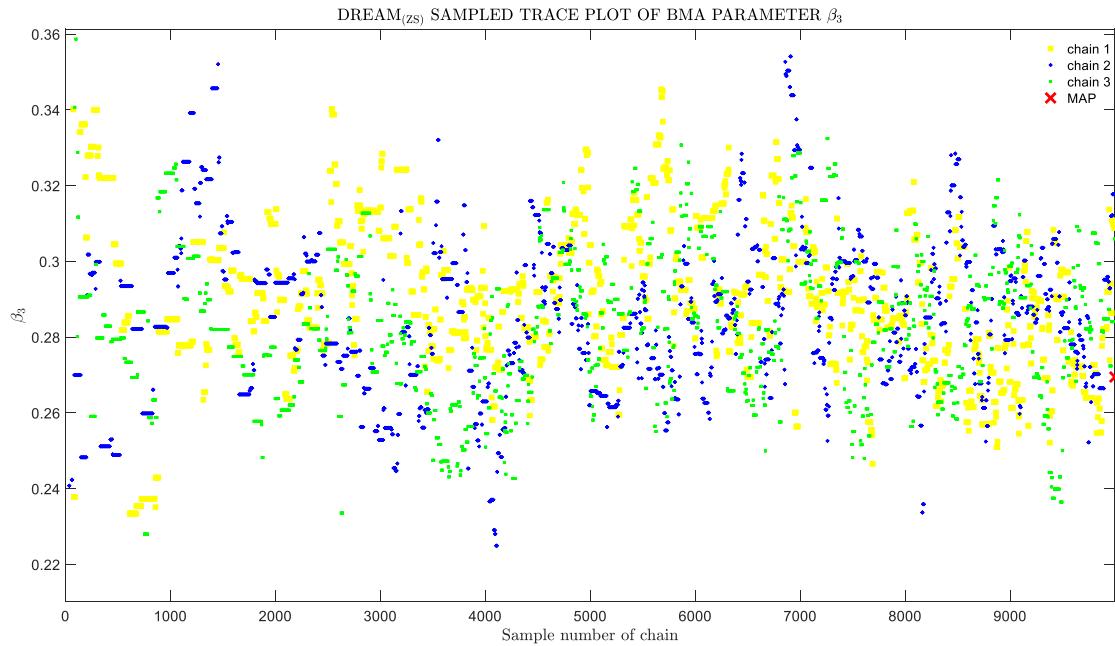


Figure E11: Traceplot of the sampled values of β_3 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

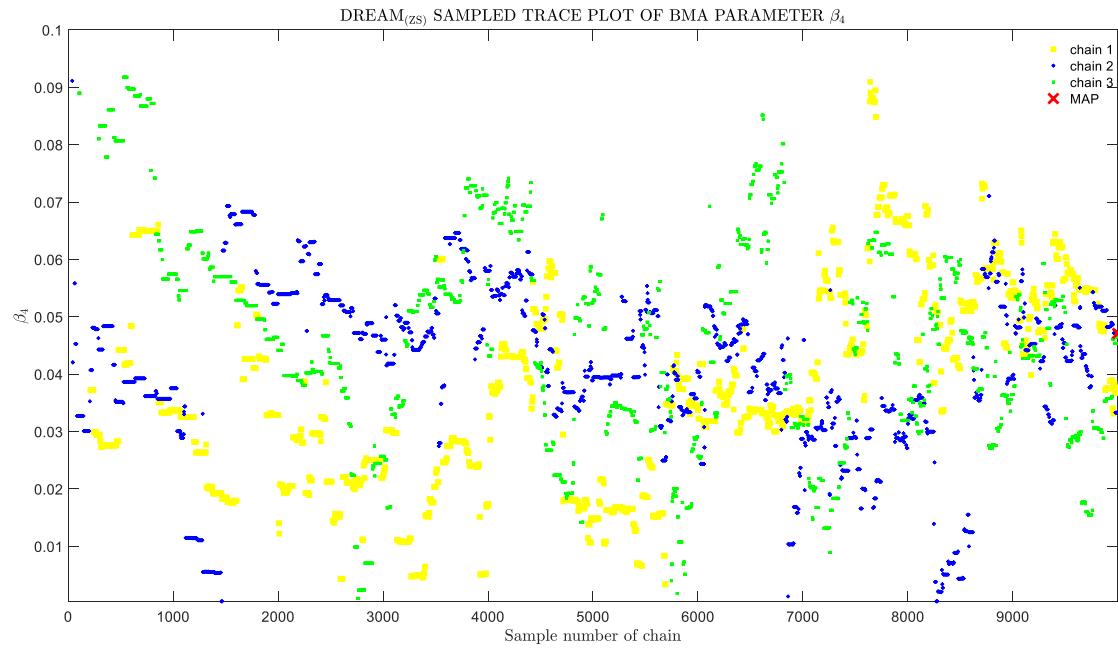


Figure E12: Traceplot of the sampled values of β_4 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

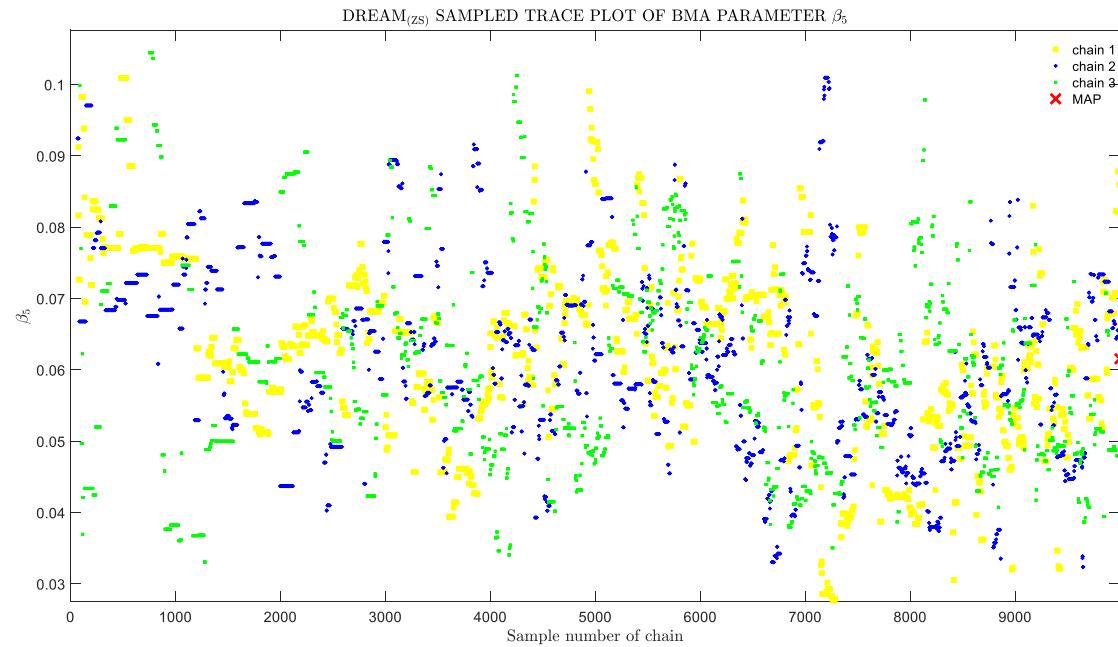


Figure E13: Traceplot of the sampled values of β_5 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

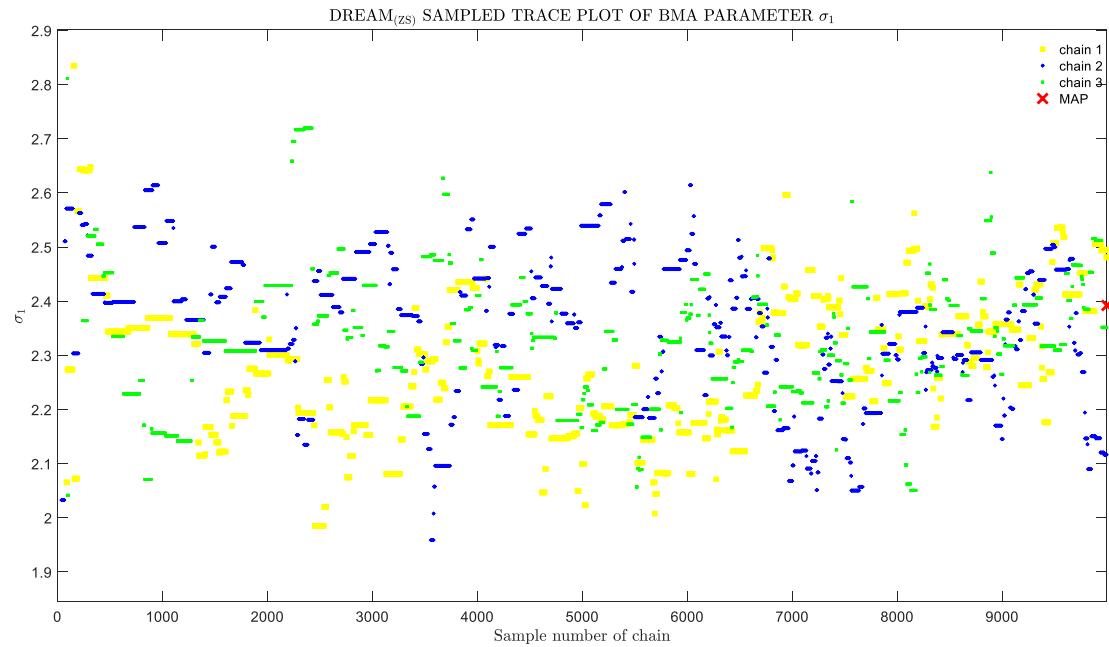


Figure E14: Traceplot of the sampled values of σ_1 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

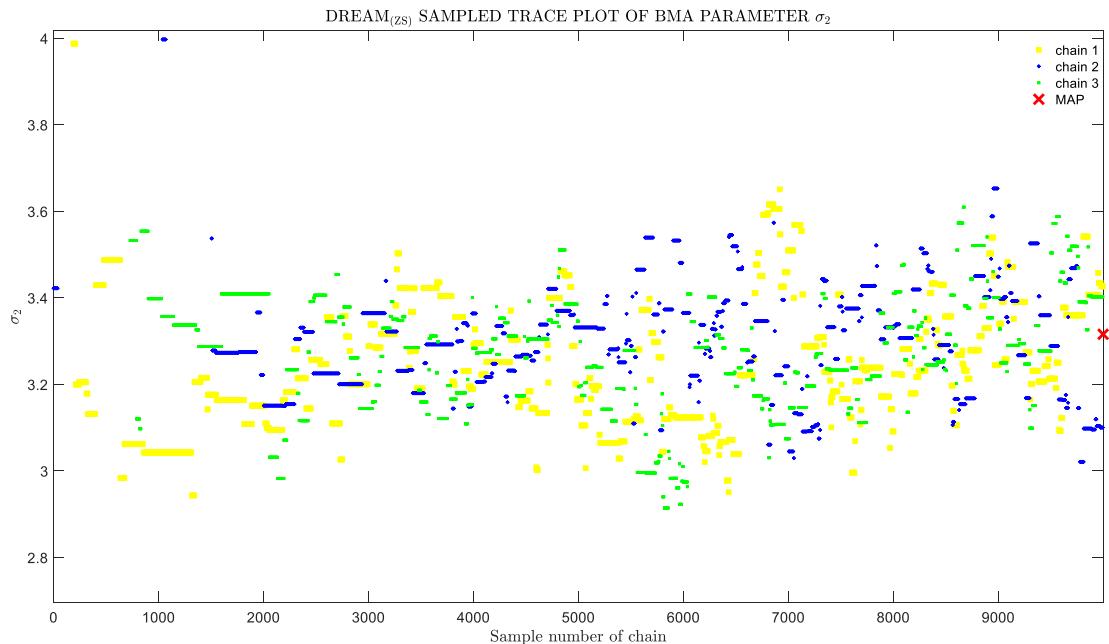


Figure E15: Traceplot of the sampled values of σ_2 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

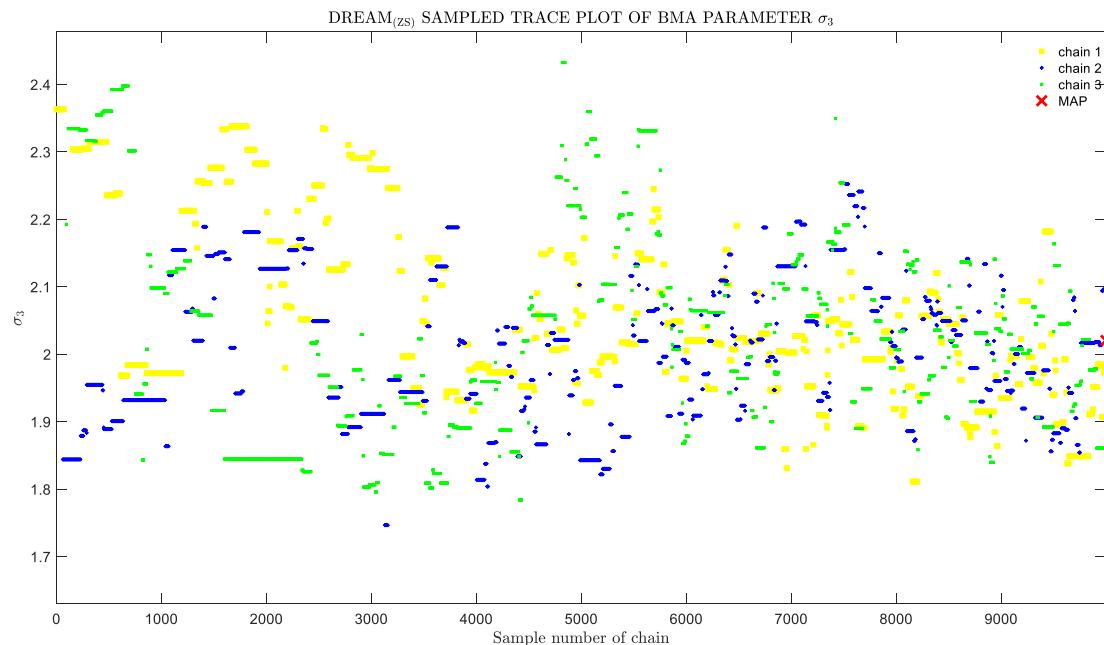


Figure E16: Traceplot of the sampled values of σ_3 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

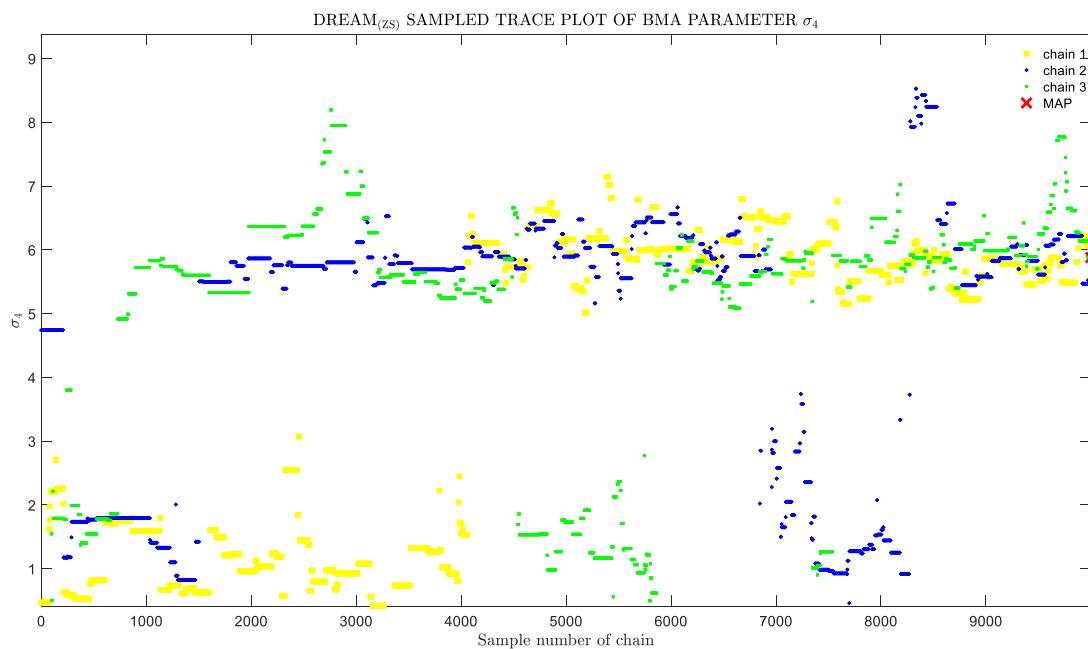


Figure E17: Traceplot of the sampled values of σ_4 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

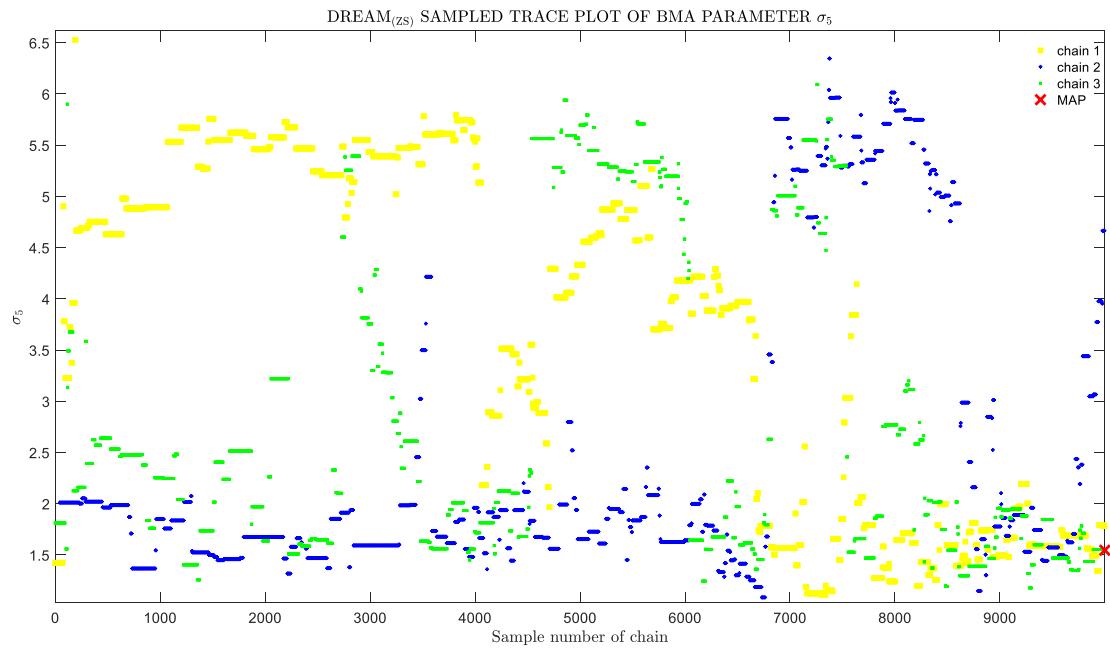


Figure E18: Traceplot of the sampled values of σ_5 in the $N = 3$ different chains simulated by the DREAM_(ZS) algorithm.

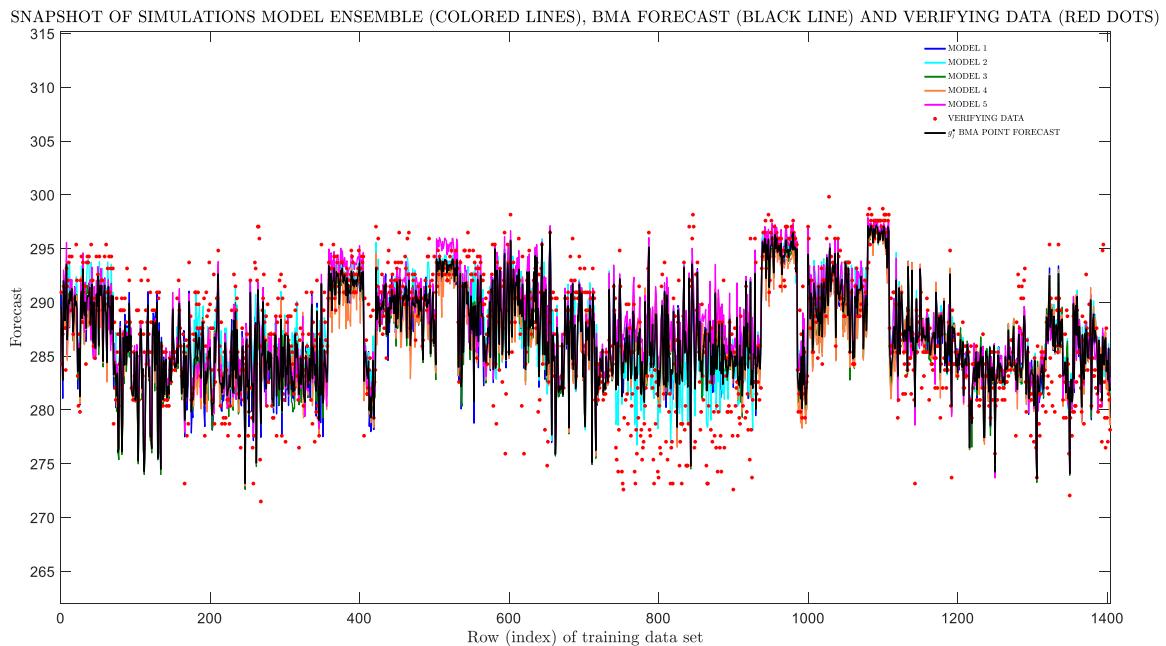


Figure E19: Snapshot of model ensemble, mean forecast of the BMA model, and the verifying data for a representative portion of the training data set.

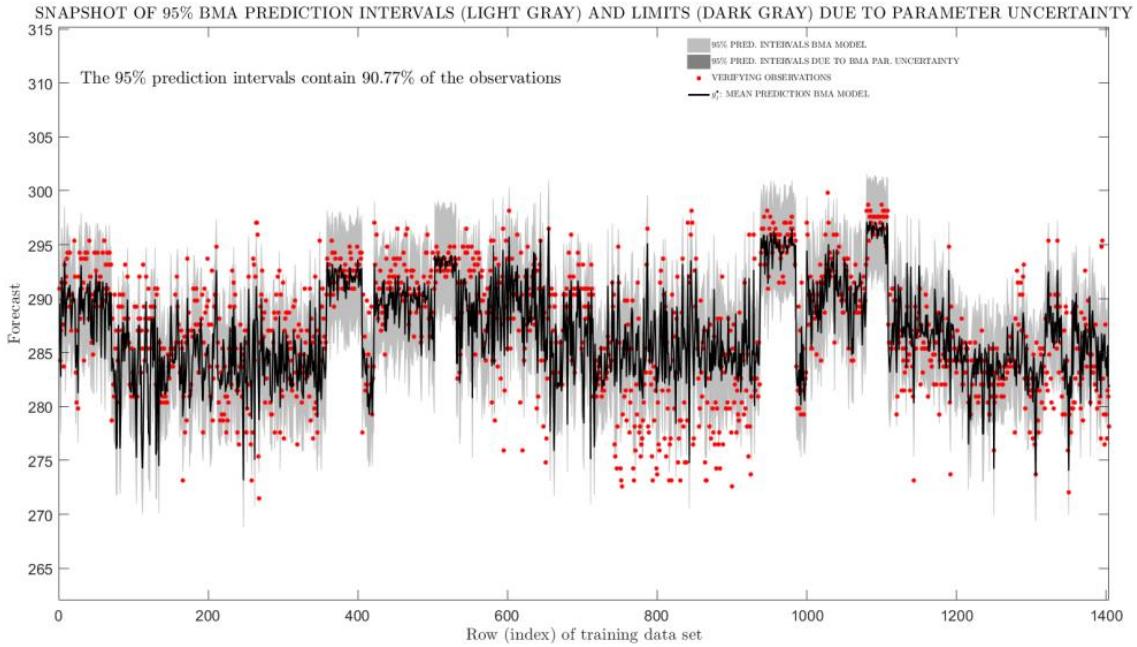


Figure E20: Snapshot of 95% prediction uncertainty ranges of the BMA due to parameter (dark grey) and total (light grey) uncertainty for a representative portion of the training data set. The mean forecast of the BMA model (solid black line), and the verifying data (red dots) are separately indicated.

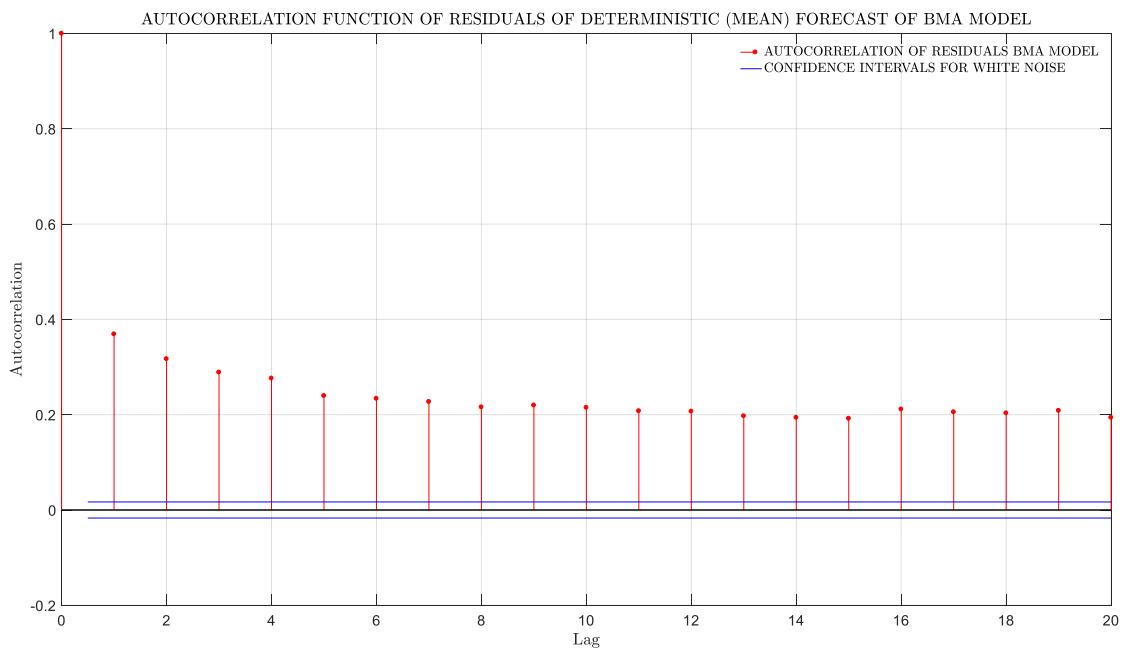


Figure E21: Autocorrelation function of the residuals of the mean forecast of the BMA model. The solid blue lines signify the 95% ranges of the autocorrelation coefficients for white noise.

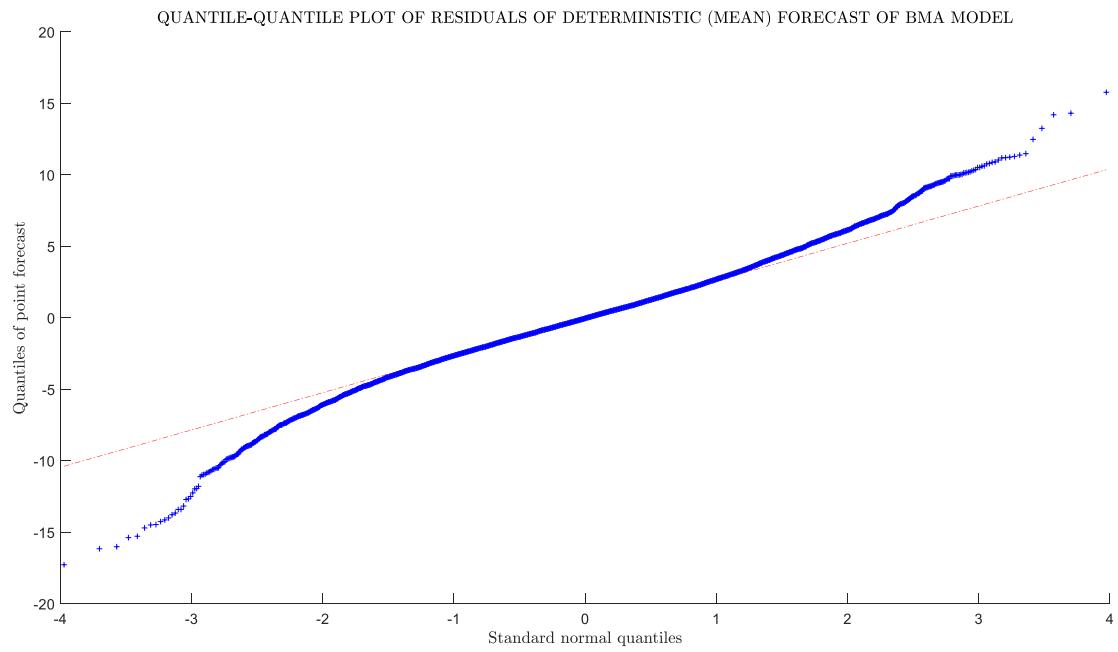


Figure E22: Quantile-quantile plot of the residuals of the mean forecast of the BMA model. The dotted red line signifies Gaussian residuals.

Appendix F. References

References

- J.M. Bates and C.M.W. Granger, "The combination of forecasts," *Operations Research Quarterly*, vol. 20, pp. 451-468, 1969.
- C.H. Bishop and K.T. Shanley, "Bayesian modeling averaging's problematic treatment of extreme weather and a paradigm shift that fixes it," *Monthly Weather Review*, vol. 136, pp. 4641-4652, 2008.
- G.E.P. Box, and D.R. Cox, "An analysis of transformations," *Journal of the Royal Statistical Society, Series B*, vol. 26 (2), pp. 211-252, 1964.
- C.J.F. ter Braak, "A Markov chain Monte Carlo version of the genetic algorithm differential evolution: easy Bayesian computing for real parameter spaces," *Statistics & Computing*, vol. 16, pp. 239-249, 2006.
- C.J.F. ter Braak, and J.A. Vrugt, "Differential evolution Markov chain with snooker updater and fewer chains," *Statistics & Computing*, vol. 18 (4), pp. 435-446, doi:10.1007/s11222-008-9104-9, 2008.
- S.P. Brooks, and A. Gelman, "General methods for monitoring convergence of iterative simulations," *Journal of Computational and Graphical Statistics*, vol. 7, pp. 434-455, 1998.
- S.T. Buckland, K.P. Burnham, and N.H. Augustin, "Model selection: An integral part of inference," *Biometrics*, vol. 53, pp. 603-618, 1997.
- K.P. Burnham, and D.R. Anderson, "Model selection and multimodel inference: A practical information-theoretic approach," 2nd edition, Springer, New York, 2002.
- A.P. Dempster, N.M. Laird, and D.B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, vol. 39(B), pp. 1-39, 1977.
- C.G.H. Diks, and J.A. Vrugt, "Comparison of point forecast accuracy of model averaging methods in hydrologic applications," *Stochastic Environmental Research and Risk Assessment*, 24(6), pp. 809-820, doi:10.1007/s00477-010-0378-z, 2010.
- Q.Y. Duan, S. Sorooshian, and V.K. Gupta, "Effective and efficient global optimization for conceptual rainfall-runoff models," *Water Resources Research*, 28 (4), pp. 1015-1031, doi:10.1029/91WR02985, 1992.
- A.G. Gelman, and D.B. Rubin, "Inference from iterative simulation using multiple sequences," *Statistical Sciences*, vol. 7, pp. 457-472, 1992.
- T. Gneiting, A.E. Raftery, A.H. Westveld, and T. Goldman, "Calibrated probabilistic forecasting using ensemble model output statistics and CRPS estimation," *Monthly Weather Review*, vol. 133, pp. 1098-1118, 2005.
- C.W.J. Granger and R. Ramanathan, "Improved methods of combining forecast accuracy," *Journal of Forecasting*, vol. 3, pp. 197-204, 1984.
- E.P. Grimit, and C.F. Mass, "Initial results of a mesoscale shortrange ensemble forecasting system over the Pacific Northwest", *Weather Forecasting*, vol. 17, pp. 192-205, 2002.
- B.E. Hansen, "Least-squares model averaging," *Econometrica*, vol. 75, pp. 1175-1189, 2007.
- B.E. Hansen, "Least-squares forecast averaging," *Journal of Econometrics*, vol. 146, pp. 342-350, 2008.
- J.A. Hoeting, D. Madigan, A.E. Raftery, and C.T. Volinsky, "Bayesian model averaging: A tutorial," *Statistical Science*, vol. 14, pp. 382-417, 1999.
- J. Geweke, "Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments," in Bayesian Statistics 4, edited by J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F.M. Smith, pp. 169-193, Oxford University Press, 1992.
- E. Laloy, and J.A. Vrugt, "High-dimensional posterior exploration of hydrologic models using multiple-try DREAM_(zs) and high-performance computing," *Water Resources Research*, vol. 48, W01526, doi:10.1029/2011WR010608, 2012.
- G. McLachlan, and T. Krishnan, The EM algorithm and extensions: 2nd Edition, 400 pages, Wiley, Apr. 2008.
- S.P. Neuman "Maximum likelihood Bayesian averaging of uncertain model predictions," *Stochastic Environmental Research and Risk Assessment*, vol. 17, pp. 291-305, 2003.
- K.V. Price, R.M. Storn, and J.A. Lampinen, Differential evolution, A practical approach to global optimization, Springer, Berlin, 2005.
- A.E. Raftery, and S.M. Lewis, "One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo," *Statistical Science*, vol. 7, pp. 493-497, 1992.
- A.E. Raftery, and S.M. Lewis, "The number of iterations, convergence diagnostics and generic Metropolis algorithms," in Practical Markov chain Monte Carlo, edited by W.R. Gilks, D.J. Spiegelhalter and S. Richardson, London, U.K., Chapman and Hall, 1995.

- A.E. Raftery, D. Madigan, and J.A. Hoeting, "Bayesian model averaging for linear regression models," *Journal of the American Statistical Association*, vol. 92, pp. 179-191, 1997.
- A.E. Raftery, T. Gneiting, F. Balabdaoui, and M. Polakowski, "Using Bayesian model averaging to calibrate forecast ensembles," *Monthly Weather Review*, vol. 133, pp. 1155-1174, 2005.
- J. Rings, J.A. Vrugt, G. Schoups, J.A. Huisman, and H. Vereecken, "Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments," *Water Resources Research*, 48, W05520, doi:10.1029/2011WR011607, 2012.
- C.P. Roberts, and G. Casella, "Monte Carlo statistical methods," 2nd edition, Springer, New York, 2004.
- M. Sadegh, and J.A. Vrugt, "Approximate Bayesian computation using Markov chain monte Carlo simulation: DREAM_(ABC)," *Water Resources Research*, vol. 50, doi:10.1002/2014WR015386, 2014.
- J.M. Sloughter, A.E. Raftery, T. Gneiting, and C. Fraley, "Probabilistic quantitative precipitation forecasting using Bayesian model averaging," *Monthly Weather Review*, vol. 135, pp. 3209-3220, 2007.
- J.M. Sloughter, T. Gneiting, and A.E. Raftery, "Probabilistic wind speed forecasting using ensembles and Bayesian model averaging," *Monthly Weather Review*, vol. 105, no. 489, pp. 25-35, doi:10.1198/jasa.2009.ap08615, 2010.
- R. Storn, and K. Price, "Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, pp. 341-359, 1997.
- J.A. Vrugt, H.V. Gupta, W. Bouten, and S. Sorooshian, "A Shuffled Complex Evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters," *Water Resources Research*, vol. 39 (8), 1201, doi:10.1029/2002WR001642, 2003.
- J.A. Vrugt, M.P. Clark, C.G.H. Diks, Q. Duan, and B.A. Robinson, "Multi-objective calibration of forecast ensembles using Bayesian model averaging," *Geophysical Research Letters*, vol. 33, L19817, doi:10.1029/2006GL027126.
- J.A. Vrugt, and B.A. Robinson, "Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging," *Water Resources Research*, vol. 43, W01411, doi:10.1029/2005WR004838, 2007.
- J.A. Vrugt, C.J.F. ter Braak, M.P. Clark, J.M. Hyman, and B.A. Robinson, "Treatment of input uncertainty in hydrologic modeling: Doing hydrology backward with Markov chain Monte Carlo simulation," *Water Resources Research*, vol. 44, W00B09, doi:10.1029/2007WR006720, 2008a.
- J.A. Vrugt, C.G.H. Diks, and M.P. Clark, "Ensemble Bayesian model averaging using Markov chain Monte Carlo sampling," *Environmental Fluid Dynamics*, vol 8, pp. 579-595, 2008b.
- J.A. Vrugt, C.J.F. ter Braak, C.G.H. Diks, D. Higdon, B.A. Robinson, and J.M. Hyman, "Accelerating Markov chain Monte Carlo simulation by differential evolution with self-adaptive randomized subspace sampling," *International Journal of Non-linear Sciences and Numerical Simulation*, vol. 10, no. 3, pp. 273-290, 2009.
- J.A. Vrugt, and C.J.F. ter Braak, "DREAM_(D): an adaptive Markov chain Monte Carlo simulation algorithm to solve discrete, noncontinuous, and combinatorial posterior parameter estimation problems," *Hydrology and Earth System Sciences*, vol. 15, pp. 3701-3713, doi:10.5194/hess-15-3701-2011, 2011.
- J.A. Vrugt, and M. Sadegh, "Toward diagnostic model calibration and evaluation: Approximate Bayesian computation," *Water Resources Research*, vol. 49, doi:10.1002/wrcr.20354, 2013.
- J.A. Vrugt, "Multi-criteria Optimization using the AMALGAM software package: Theory, concepts, and MATLAB Implementation," *Manual, Version 1.0*, pp. 1-53, 2015a.
- J.A. Vrugt, "FDCFIT: A MATLAB Toolbox of parametric expressions of the flow duration curve," *Manual, Version 1.0*, pp. 1-35, 2015b.
- J.A. Vrugt, "Markov chain Monte Carlo simulation using the DREAM software package: Theory, concepts, and MATLAB Implementation," *Environmental Modeling & Software*, vol. 75, pp. 273-316, 10.1016/j.envsoft.2015.08.013, 2016.
- J.A. Vrugt, "Distribution-based Model Evaluation and Diagnostics: Elicitability, Propriety and Scoring Rules for Hydrograph Functionals," *Water Resources Research*, In Review, 2023. Earlier draft of this paper: <https://www.authorea.com/users/597576/articles/632989-the-promise-of-diversity-distribution-based-hydrologic-model-evaluation-and-diagnostics>
- T. Wöhling, and J.A. Vrugt, "Combining multiobjective optimization and Bayesian model averaging to calibrate forecast ensembles of soil hydraulic models," *Water Resources Research*, vol. 44, W12432, pp. 1-18, 2008.
- M. Ye, P.D. Meyer and S.P. Neumann, "On model selection criteria in multimodel analysis," *Water Resources Research*, vol. 44, W03428, pp. 1-12, 2008.
- X. Zhang, A.T.K. Wan, and G. Zou, "Least squares model combining by Mallows criterion," *SSRN working paper*, 2008.