Multi-objective calibration of forecast ensembles using Bayesian model averaging

Jasper A. Vrugt,[1] Martyn P. Clark,[2] Cees G. H. Diks,[3] Qinyun Duan,[4] and Bruce A. Robinson[1]

Received 6 June 2006; revised 9 August 2006; accepted 24 August 2006; published 12 October 2006.

[1] Bayesian Model Averaging (BMA) has recently been proposed as a method for statistical postprocessing of forecast ensembles from numerical weather prediction models. The BMA predictive probability density function (PDF) of any weather quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts from a set of different models. However, current applications of BMA calibrate the forecast specific PDFs by optimizing a single measure of predictive skill. Here we propose a multi-criteria formulation for postprocessing of forecast ensembles. Our multi-criteria framework implements different diagnostic measures to reflect different but complementary metrics of forecast skill, and uses a numerical algorithm to solve for the Pareto set of parameters that have consistently good performance across multiple performance metrics. Two illustrative case studies using 48-hour ensemble data of surface temperature and sea level pressure, and multi-model seasonal forecasts of temperature, show that a multi-criteria formulation provides a more appealing basis for selecting the appropriate BMA model. **Citation:** Vrugt, J. A., M. P. Clark, C. G. H. Diks, Q. Duan, and B. A. Robinson (2006), Multi-objective calibration of forecast ensembles using Bayesian model averaging, *Geophys. Res. Lett.*, *33*, L19817, doi:10.1029/2006GL027126.

1. Introduction and Scope

[2] Ensemble Bayesian Model Averaging (BMA) has recently been proposed by *Raftery et al.* [2005] as a new method for statistical postprocessing of forecast ensembles from numerical weather prediction models. The BMA predictive probability density function (PDF) of any quantity of interest is a weighted average of PDFs centered on the bias-corrected forecasts from a set of different models. The weights assigned to each model reflect that model's contribution to the forecasting skill over a training period. Various recent studies have demonstrated that the BMA method yields prediction intervals that more closely describe the forecast error than the raw ensemble, and BMA provides deterministic point forecasts that have better predictive performance than the best of the ensemble members, or the ensemble mean [*Raftery et al.*, 2005; *Sloughter et al.*, 2006; *Min and Hense*, 2006; *Vrugt and Robinson*, 2005].

[3] Successful implementation of the BMA method, however, requires calibration of the forecast specific predictive PDFs. In their seminal paper, *Raftery et al.* [2005] proposes to calibrate these individual predictive PDFs by maximum likelihood (ML) from a training data set. The likelihood function is defined as the probability of the training data given the model parameters to be estimated. While the ML estimator exhibits many optimality properties [*Casella and Berger*, 2001], a single metric cannot reflect all the aspects of forecast skill deemed important for accurate probabilistic weather forecasting. For instance, the goal of postprocessing ensembles is not only to obtain deterministic point forecasts that exhibit the smallest possible quadratic forecast errors, but also to obtain predictive PDFs that exhibit appropriate coverage and are as sharp as possible [*Gneiting et al.*, 2003]. By sharp we mean that the prediction intervals are as small as possible, while still being statistically meaningful and significant.

[4] In this paper we propose a hybrid multi-criteria optimization approach for calibration of a forecast ensemble using Bayesian Model Averaging (BMA). Our framework implements different diagnostic measures to reflect different but complementary aspects of forecast skill, and the resulting multiobjective optimization problem is solved by means of a numerical evolutionary algorithm [J. A. Vrugt and B. A. Robinson, Improved evolutionary optimization from genetically adaptive multi-method search, submitted to *Proceedings of the National Academy of Sciences of the United States of America*, 2006, hereinafter referred to as Vrugt and Robinson, submitted manuscript, 2006]. Performance metrics that will be used in the multi-criteria analysis include the Root Mean Square Error (*RMSE*) measuring the average quadratic forecast error of the BMA deterministic point prediction, and the ignorance score [*Good*, 1952] and continuous ranked probability score (*CRPS*) [*Brown*, 1974; *Matheson and Winkler*, 1976; *Hersbach*, 2000], to measure the sharpness of the BMA predictive PDF. Recently, *Carney and Cunningham* [2005] have presented a similar multi-objective idea to the calibration of probability density forecasts. However, our proposed multi-objective framework is more general, as it not only applies to BMA, but other statistical methods for ensemble postprocessing as well.

[5] Sections 2–4 describe the BMA method, discuss the multi-criteria optimization methodology and illustrate the

[1]Earth and Environmental Sciences Division, Los Alamos National Laboratory, Los Alamos, New Mexico, USA.
[2]National Institute for Water and Atmospheric Research, Christchurch, New Zealand.
[3]Center for Nonlinear Dynamics in Economics and Finance, University of Amsterdam, Amsterdam, Netherlands.
[4]Lawrence Livermore National Laboratory, Livermore, California, USA.

0094-8276/06/2006GL027126$05.00

power of the approach through a case study using MM5 ensemble data of 48-h forecasts of surface temperature and sea level pressure in the North American Pacific Northwest [*Grimit and Mass*, 2002], and 3-month seasonal forecasts of surface temperature over the contiguous United States using the DEMETER ensemble prediction system [*Palmer et al.*, 2004].

2. Ensemble Bayesian Model Averaging (BMA)

[6] Bayesian model averaging was originally developed as a way to combine inferences and predictions of several competing statistical models and to assess their joint predictive uncertainty. *Raftery et al.* [2005] recently extended BMA to ensembles of dynamical models and demonstrated how it can be used to postprocess forecast ensembles from dynamic weather models.

[7] To explicate the BMA method developed by *Raftery et al.* [2005], let f $\{f_1,..f_K\}$ denote an ensemble of predictions obtained from K different models, and Δ be the quantity of interest. In BMA, each ensemble member forecast, f_k, is associated with a conditional PDF, $g_k(\Delta \mid f_k)$, which can be interpreted as the PDF of Δ given f_k. The BMA predictive model for dynamic ensemble forecasting can then be expressed as:

$$p(\Delta|f_1,\ldots,f_k) = \sum_{k=1}^{K} w_k g_k(\Delta|f_k) \tag{1}$$

where w_k denotes the posterior probability of forecast k being the best one. The w_k's are nonnegative and add up to one, and they can be viewed as weights reflecting an individual model's relative contribution to predictive skill over the training period.

[8] The original ensemble BMA method described by *Raftery et al.* [2005] assumes that the conditional PDFs $g_k(\Delta|f_k)$ of the different ensemble members can be approximated by a normal distribution centered at a linear function of the original forecast, $a_k + b_k f_k$ and standard deviation σ:

$$\Delta|f_k \sim N\left(a_k + b_k f_k, \sigma^2\right) \tag{2}$$

The values for a_k and b_k are bias-correction terms that are derived by simple linear regression of Δ on f for each of the individual ensemble members. The approximation used in equation (2) seems to be reasonable for variables such as temperature and sea level pressure considered in this paper. For other variables such as precipitation and streamflow, the normal distribution is not appropriate and other conditional PDFs need to be implemented. Recent work presented by *Sloughter et al.* [2006] and *Vrugt and Robinson* [2005] discusses how to use the BMA method for distributions other than the normal PDF, such as the gamma distribution.

[9] The BMA predictive mean can be computed as:

$$E[\Delta|f_1,\ldots,f_K] = \sum_{k=1}^{K} w_k(a_k + b_k f_k) \tag{3}$$

which is a deterministic forecast, whose predictive performance can be compared with the individual forecasts in the ensemble, or with the ensemble mean. If we denote space and time with subscripts s and t respectively, so that f_{kst} is the kth forecast in the ensemble for place s and time t, the associated variance can be computed as [*Raftery et al.*, 2005]:

$$Var[\Delta_{st}|f_{1st},\ldots,f_{Kst}] = \sum_{k=1}^{K} w_k \left((a_k + b_k f_{kst}) - \sum_{l=1}^{K} w_l(a_l + b_l f_{lst})\right)^2 + \sum_{k=1}^{K} w_k \sigma_k^2 \tag{4}$$

The BMA variance defined in equation (4) consists of two terms, the first representing the ensemble spread, and the second representing the within-forecast variance.

[10] The crux of the BMA method is estimating appropriate values for the weights and variance parameter. *Raftery et al.* [2005] implement a step-wise approach to BMA parameter estimation in which they first remove any long-term biases from each individual model, then use the Expectation-Maximization (EM) algorithm to estimate the weight and variance parameters (by maximizing the log-likelihood value), and finally they tune the BMA variance to optimize the continuous ranked probability score (*CRPS*). Such a step-wise parameter estimation approach appears inefficient, and does not address interactions between different forecasting objectives. In the next section we suggest a more appealing alternative for estimating the BMA weights and variance parameters.

3. Multi-Criteria Calibration of BMA Model

[11] The maximum likelihood estimator used by *Raftery et al.* [2005] is the value of the parameter vector that maximizes the likelihood function, that is, the value of the parameter vector under which the observed data were most likely to have been observed. This estimator exhibits many optimality properties [*Casella and Berger*, 2001] but emphasizes predictive skill in terms of average distance of the verifying observations to the BMA predictive mean. There are however, various other attributes of forecast quality, such as sharpness that are also considered important for accurate probabilistic weather forecasting. *Murphy and Winkler* [1992] provide an overview of the various attributes of forecast quality that are deemed important for probabilistic forecasting.

[12] Instead of using a single measure of predictive skill to estimate the BMA weights and variance, we here propose using multiobjective optimization for calibration of the forecast ensemble. Our multi-criteria framework implements three different diagnostic measures to reflect different but complementary measures of forecast skill. These metrics include the Root Mean Square Error (*RMSE*) to measure the average quadratic forecast error of the BMA deterministic point prediction, and the ignorance score (or negative log predictive density - *NLL*) and *CRPS* to measure the sharpness or spread of the BMA predictive PDF [*Good*, 1952; *Brown*, 1974; *Matheson and Winkler*, 1976; *Hersbach*, 2000; *Raftery et al.*, 2005]. These three measures provide

complementary information about the BMA calibrated forecast ensembles and can be computed as follows:

$$RMSE = \sqrt{\frac{\sum_{s,t}\left(\sum_{k=1}^{K} w_k f_k - \Delta_{st}\right)^2}{n}},$$

$$NLL = -\frac{1}{n}\sum_{s,t} \log[h_{st}(y|\Delta_{st})],$$

$$CRPS = \frac{1}{n}\sum_{s,t} \int_{-\infty}^{\infty} (H_{st}(y) - 1\{y \geq \Delta_{st}\})^2 dy \quad (5)$$

where $H(y)$ and $h(y)$ denote the cumulative distribution and probability density function of the BMA predictive PDF respectively, and $1\{y \geq \Delta_{st}\}$ is the Heaviside function, that attains the value 1 if $y \geq \Delta_{st}$ and the value 0 otherwise, and n is the total number of observations. Note that the *NLL* score equals the average of the natural logarithms of the BMA PDFs evaluated at the observations, and the *CRPS* is equivalent to the mean absolute error for a deterministic forecast [*Hersbach*, 2000]. Smaller values of the diagnostic measures are preferred and indicate better performance of the BMA predictive model. Please note, that in this paper the ignorance score and *CRPS* are optimized for the training data set. This is unlike typical applications of these scores.

[13] In the case of multiple performance metrics, there will generally not be a single combination of values for w_k, $k = 1,\ldots K$ and σ^2 that optimizes all performance diagnostics simultaneously, but rather a Pareto set of solutions corresponding to trade-offs among the various performance metrics. This Pareto set defines the minimum uncertainty in the BMA weights and variance that can be specified without stating a subjective preference for minimizing one specific performance measure at the expense of another.

[14] If we proceed with our multi-criteria calibration framework, we need an algorithm that can efficiently solve for the Pareto set of solutions. An efficient evolutionary algorithm for solving the multiobjective problem has recently been developed (Vrugt and Robinson, submitted manuscript, 2006). This approach is called A Multi-ALgorithm Genetically Adaptive Multiobjective or AMALGAM method, and combines two new concepts, simultaneous multi-method search, and self-adaptive offspring creation, to ensure a fast and computationally efficient solution to multiobjective optimization problems. Experiments conducted using standard, synthetic multi-objective test problems have shown that the AMLAGAM method is on the order of 3–10 times more efficient than current state-of-the art multiobjective optimization algorithms, and provides a final population that closely approximates the Pareto solution space. A detailed description and explanation of the method is given by Vrugt and Robinson (submitted manuscript, 2006), and so will not be repeated here.

4. Case Studies

[15] We illustrate the usefulness and applicability of the multi-criteria optimization approach to probabilistic weather forecasting for two case studies with increasing complexity. The first case study considers the 48-h forecasts of surface temperature and sea level pressure in the North American Pacific Northwest in January–June 2000 using the University of Washington (UW) mesoscale short-range ensemble system [*Grimit and Mass*, 2002]. This is a five member multianalysis ensemble (hereafter referred to as the UW ensemble) consisting of different runs of the fifth-generation Pennsylvania State University – National Center for Atmospheric Research Mesoscale Model (MM5), in which initial conditions are taken from different operational centers. In the second case study, we use 3-month seasonal forecasts of surface temperature over the contiguous United States for each of the four calendar seasons for the period 1980–2001. This data was taken from the DEMETER project (Development of a European Multi-model Ensemble system for seasonal to inTERannual prediction) and consists of a seven member multi-model ensemble generated by seven state-of-the-art global coupled ocean-atmosphere models [*Palmer et al.*, 2004].

[16] In both case studies, the Pareto optimal solutions space for the three-criterion {*RMSE, NLL, CRPS*} calibration was estimated with the AMALGAM method using a population size of 100 points in combination with 100 generations. We used a uniform prior distribution of [0,1], and $[0,3\cdot Var(\Delta)]$ for the individual weights and variance respectively. To satisfy the constraint that the sum of the weights equals one, each AMALGAM generated combination of weights is first divided by its respective sum before evaluating the BMA model and computing the performance measures in equation (5).

4.1. Case Study I: UW Ensemble Forecasts of Temperature and Sea Level Pressure

[17] We consider 48-h forecast of surface temperature and sea level pressure using data from the UW ensemble consisting of forecasts and observations in the 0000 UTC cycle from 12 January to 9 June 2000. Following *Raftery et al.* [2005] a 25-day training period between April 16 and 9 June 2000 is used for BMA model calibration, whereas the remainder of the data set (12 January–16 April) is used to evaluate the model performance. 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 bias correction terms a_k and b_k in equations (2)–(4) for the individual ensemble members were derived using simple linear regression of Δ_{st} on f_{kst} for the training data set.

[18] Figure 1 presents normalized parameter plots of the results for the three criterion BMA model calibration with the AMALGAM method for the temperature (Figure 1a) and sea level pressure data set (Figure 1b). Each line going from left to right across the plot corresponds to a different parameter combination. The gray lines represent members of the Pareto set (appears as a band), whereas the black lines refer to the optimal solution derived by maximum likelihood estimation using the Expectation – Maximization (EM) algorithm. The six BMA model parameters are listed along the x – axis, and the y – axis corresponds to the parameter values, normalized by their prior uncertainty ranges. The two plots at the right-hand side in Figure 1 depict three-dimensional projections of the trade-off surface of the three diagnostic performance measures for both data sets. The Pareto rank 1 solutions in these plots are indicated

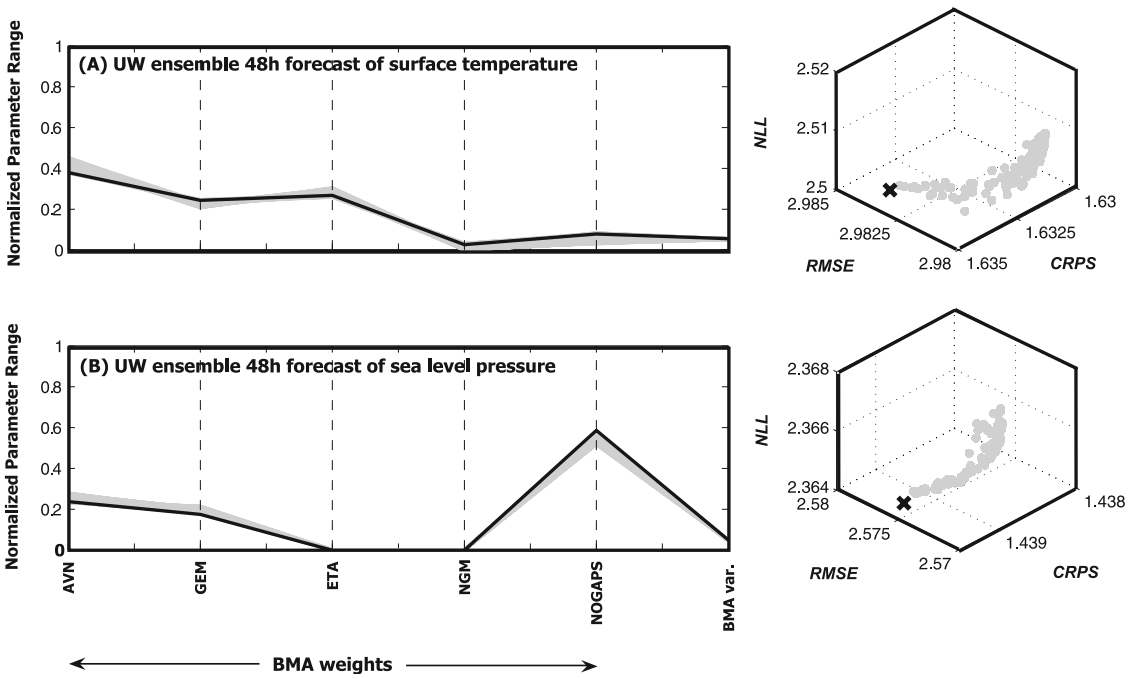

Figure 1. Normalized parameter plots for the BMA weights and variance using a three-criterion (*RMSE*, *NLL*, *CRPS*) calibration with the AMALGAM algorithm for the UW 48-hour ensemble data sets of (a) surface temperature and (b) sea level pressure. Each line across the graph denotes a single parameter set: shaded is Pareto solution set; solid lines denote maximum likelihood (ML) estimates separately derived with the EM algorithm. The plots at the right-hand side are three-dimensional projections of the objective space of the Pareto set of solutions. The ML estimator is indicated with the 'x' symbol.

with the gray dots, whereas the solution corresponding to the maximum likelihood estimator, separately derived with the EM algorithm, is denoted with the 'x' symbol.

[19] The results presented in Figure 1 demonstrate that the Pareto solution space spans only a very small region interior to the prior defined plausible parameter space for both data sets. This illustrates that the BMA model is well defined by calibration against the three individual performance criteria. This is further confirmed by a check of the three-dimensional projections of the {*RMSE, NLL, CRPS*} space of the Pareto set of solutions at the right hand side. Both these plots show very small trade-off in the fitting of the various performance measures, and suggest that it is possible to identify a single BMA model that has good and consistent performance for each of the individual performance metrics. The user who requires a best parameter set from the Pareto region can select the maximum likelihood estimator, which albeit not being part of the Pareto front, seems to provide an acceptable trade-off between the performance measures.

[20] In general we can conclude that the multi-criteria optimization helps to guide the search for an appropriate BMA model, and provides useful information about the trade-offs between the various performance metrics. The current practice of optimizing the BMA model using maximum likelihood theory seems to result in a calibrated forecast ensemble that receives good performance in terms of quadratic forecast error, and sharpness of the prediction intervals. However, such consistent performance of the ML method cannot be guaranteed for all forecasting problems.

4.2. Case Study II: DEMETER Seasonal Forecasts of Surface Temperature

[21] The DEMETER project used seven state-of-the-art global coupled ocean-atmosphere models to produce long-lead (6-month) seasonal climate forecasts [*Palmer et al.*, 2004]. The models used were CERFACS [*Déqué*, 2001], ECMWF [*Gregory et al.*, 2000], INVG [*Roeckner*, 1996], LODYC [*Gregory et al.*, 2000], Meteo-France [*Déqué*, 2001], the U.K. Meteorological Office [*Pope et al.*, 2000], and MPI [*Marsland et al.*, 2003]. Please refer to *Palmer et al.* [2004] for more details on each of the seven models.

[22] Each model in the DEMETER project was used to produce forecasts (or, more exactly, hindcasts) over the past several decades. While the ECMWF, Meteo-France, and the U.K. Meteorological Office produced hindcasts for the period 1959–2001, all seven models produced hindcasts for the period 1980–2001. In each year, 6-month hindcasts were initialized starting on 1 February, 1 May, 1 August, and 1 December. Each of the seven models was run with nine ensemble members.

[23] In this study we restrict attention to 3-month forecasts of precipitation and temperature over the contiguous United States for each of the four calendar seasons for the period 1980–2001. That is, we use forecasts for March, April, and May (MAM) initialized on 1 February, forecasts for June, July and August (JJA) initialized on 1 May, forecasts for September, October and November (SON) initialized on 1 August, and forecasts for December, January and February (DJF) initialized on 1 November. For convenience, we use the ensemble mean for each of the seven

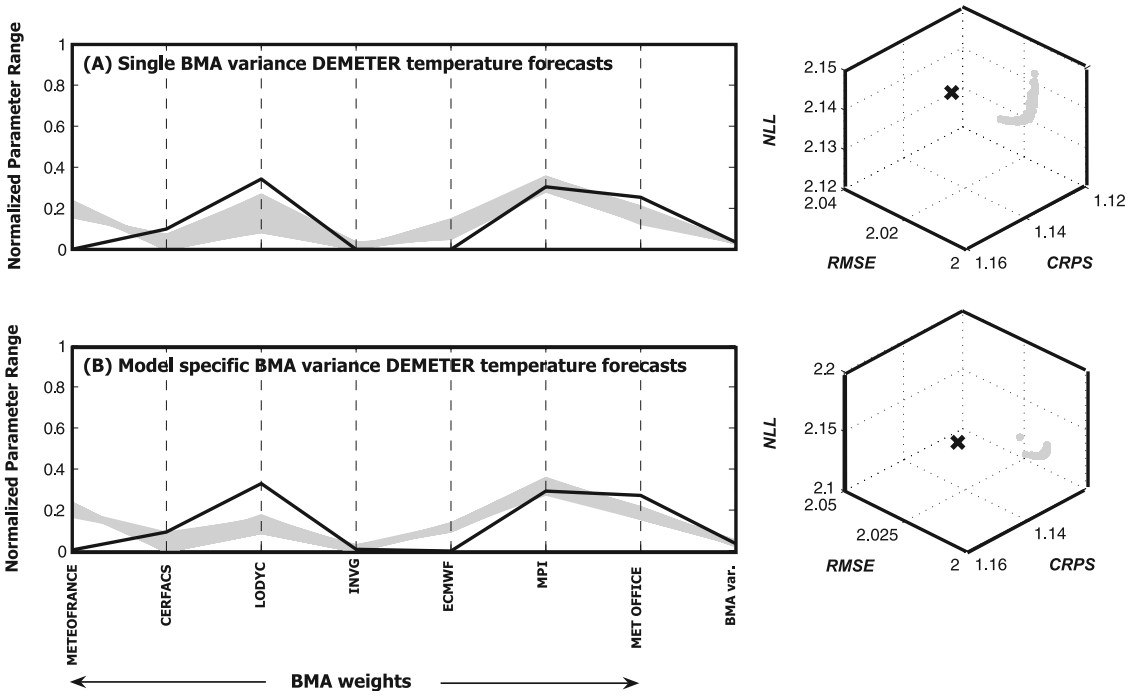

Figure 2. Normalized parameter plots for the DEMETER seasonal forecasts of temperature throughout the United States using either (a) a single variance for the individual models of the ensemble or (b) different model specific BMA variances. Each line across the graph denotes a single parameter set: shaded denote Pareto solutions; solid lines represent the maximum likelihood estimates derived with the EM algorithm. The plots at the right-hand side denote three-dimensional projections of the objective space of the Pareto set of solutions. In these plots, rank 1 solutions are indicated by solid dots (appears as a cluster), whereas the ML solution is indicated with a cross symbol.

models in our BMA analysis. In future work we will attempt to further exploit the information contained in the ensemble of predictions from each individual model.

[24] Figure 2 presents the results of the three criteria {*RMSE, NLL, CRPS*} BMA model calibration with the AMALGAM algorithm in the normalized parameter and objective space for two cases. In the first case (Figure 2a) a single BMA variance was used for the different model forecasts, whereas in the second case (Figure 2b) different variance parameters for the individual models of the ensemble were used. The gray lines represent members of the final Pareto solution set (appears as a band), and the solid black line denotes the maximum likelihood estimator for the same training data set, separately derived with the EM algorithm. The various model parameters are listed along the x – axis, where the y – axis corresponds to the parameter values, normalized by their initial uncertainty ranges. For the second case, we averaged the individual optimized BMA variances of the Pareto region to result in a single average BMA variance, and facilitate comparison with the results of the first case. The squared plots at the right hand-side denote three dimensional projections of the Pareto performance diagnostic trade-off surface. In this plot, gray dots denote Pareto solutions, whereas the black 'x' symbol represents the performance of the maximum likelihood estimator.

[25] The results presented in Figure 2 highlight several important observations. First, notice that the BMA weight and variance trade-off region (Pareto solution space) is quite small compared to the initial uncertainty ranges. Also, for this data set the BMA model is well defined by Pareto calibration against the {*RMSE, NLL, CRPS*} performance diagnostics, and not much trade-offs appears in the fitting of these various measures of forecast skill. Even though the maximum likelihood estimator is not part of the Pareto trade-off region, it seems that this estimator still provides an acceptable trade-off between the three performance metrics.

[26] The second interesting observation is that the size of the Pareto space of the BMA weights is smaller for the second case study, where we calibrated different BMA variances for the individual models of the ensemble. This is an interesting result, because added complexity usually allows for increasing trade-offs between performance metrics.

[27] A third significant and interesting observation is that the AMALGAM algorithm seems to provide a very efficient search of the BMA weight and variance space. With only 10,000 function evaluations the algorithm is able to create 100 Pareto solutions that span the entire trade-off surface and include the theoretical single objective solutions at the extreme end of the Pareto front. This is especially impressive for the second case study reported in Figure 2b, where the AMALGAM method is used to simultaneously optimize 14 BMA parameters (7 weights + 7 variances). Our conjecture is that a multi-criteria calibration can increase the identifiability of the global optimum, and therefore reduces the number of function evaluations needed to find Pareto solutions.

5. Conclusions

[28] In this paper we have presented a multi-criteria optimization approach to calibration of a forecast ensemble

using Bayesian Model Averaging (BMA). Our multi-criteria framework implements different diagnostic measures to reflect different but complementary metrics of forecast skill, and uses a genetically adaptive evolutionary algorithm to solve for the non-dominated or Pareto set of solutions. Through two case studies we have shown that our multi-objective calibration approach is practical and relatively simple to implement, and helps guide the selection of the appropriate values of the BMA weights and variances. Interestingly, for the case studies presented in this paper, little trade-off appeared in the fitting of the *RMSE*, *NLL*, and *CRPS* measures. Of course, we could have considered other objectives that would have shown more trade-off. As illustration, however, we preferred to use performance metrics that are commonly used to evaluate probabilistic forecasts.

[29] The small trade-off between forecasting objectives occurs because the three objective functions examined in this study are highly correlated. As such, the maximum likelihood estimator, albeit not being part of the Pareto set in both case studies, still results in an acceptable trade-off between the various performance metrics. However, the statistical similarity between forecasting objectives cannot be guaranteed for all forecasting problems. For example, our recent other studies that apply the multi-criteria concept to calibration of multi-model streamflow forecasts show significant trade-off, both in the parameter and objective space. In those cases, the user has to decide which Pareto solution provides the most acceptable trade-off for the forecasting application at hand. We therefore suggest that multi-criteria optimization methods are a more robust and defensible choice for post-processing of forecast ensembles.

[30] **Acknowledgments.** The first author is supported by the LANL Director's Funded Postdoctoral program. The second author acknowledges support from the New Zealand Foundation for Research Science and Technology (contract C01X0401). We thank Adrian Raftery from the Department of Statistics at the University of Washington for providing the MM5 ensemble weather data. In addition, detailed reviews of Richard Ibbitt and two anonymous reviewers greatly improved the current manuscript.

References

Brown, T. A. (1974), Admissible scoring systems for continuous distributions, *Rand Corp. Rep.*, *P-5235*, 22.

Carney, M., and P. Cunningham (2005), Calibrating probability density functions with multi-objective search, *Tech. Rep. TCD-CS-2006-07*, Comput. Sci. Dep., Trinity Coll. Dublin, Ireland. (Available at https://www.cs.tcd.ie/publications/tech-reports/reports.06/TCD-CS-2006-07.pdf).

Casella, G., and R. L. Berger (2001), *Statistical Inference*, 2nd ed., 660 pp., Duxbury, Pacific Grove, Calif.

Déqué, M. (2001), Seasonal predictability of tropical rainfall: Probabilistic formulation and validation, *Tellus, Ser. A*, *53*, 500–512.

Gneiting, T., A. E. Raftery, F. Balabdaoui, and A. Westveld (2003), Verifying probabilistic forecasts: Calibration and sharpness, paper presented at Workshop on Ensemble Forecasting, Val-Morin, Que., Canada, 18–20 Sept.

Good, I. J. (1952), Rational decisions, *J. R. Stat. Soc., Ser. B*, *14*, 107–114.

Gregory, D., J. J. Morcrette, C. Jakob, A. C. M. Beljaars, and T. Stockdale (2000), Revision of convection, radiation, and cloud schemes in the ECMWF integrated forecasting system, *Q.J.R. Meteorol. Soc.*, *126*, 1685–1710.

Grimit, E. P., and C. F. Mass (2002), Initial results of a mesoscale short-range ensemble forecasting system over the Pacific Northwest, *Weather Forecasting*, *17*, 192–205.

Hersbach, H. (2000), Decomposition of the continuous ranked probability score for ensemble prediction systems, *Weather Forecasting*, *15*, 559–570.

Marsland, S. J., H. Haak, J. H. Jungclaus, M. Latif, and F. Röske (2003), The Max-Planck-Institute global ocean/sea ice model with orthogonal curvilinear coordinates, *Ocean Modell.*, *5*(2), 91–127.

Matheson, J. E., and R. L. Winkler (1976), Scoring rules for continuous probability distributions, *Manage. Sci.*, *22*, 1087–1095.

Min, S., and A. Hense (2006), A Bayesian approach to climate model evaluation and multi-model averaging with an application to global mean surface temperatures from IPCC AR4 coupled climate models, *Geophys. Res. Lett.*, *33*, L08708, doi:10.1029/2006GL025779.

Murphy, A. H., and R. L. Winkler (1992), Diagnostic verification of probability forecasts, *Int. J. Forecasting*, *7*, 435–455.

Palmer, T. N., et al. (2004), Development of a European multi-model ensemble system for seasonal-to-interannual prediction (DEMETER), *Bull. Am. Meteorol. Soc.*, *85*, 853–872, doi:10.1175/BAMS-85-6-853.

Pope, V. D., M. L. Gallani, P. R. Rowntree, and R. A. Stratton (2000), The impact of new physical parameterizations in the Hadley Centre climate model: HadAM3, *Clim. Dyn.*, *16*, 123–146.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski (2005), Using Bayesian model averaging to calibrate forecast ensembles, *Mon. Weather Rev.*, *133*, 1155–1174.

Roeckner, E. (1996), The Atmospheric General Circulation Model ECHAM-4: Model description and simulation of present-day climate, *Tech. Rep. 218*, 90 pp., Max-Planck-Inst. für Meteorol., Hamburg, Germany.

Sloughter, J. M., A. E. Raftery, and T. Gneiting (2006), Probabilistic quantitative precipitation forecasting using Bayesian model averaging, *Tech. Rep. 496*, 20 pp., Dep. of Stat., Univ. of Wash., Seattle.

Vrugt, J. A., and B. A. Robinson (2006), Treatment of uncertainty using ensemble methods: Comparison of sequential data assimilation and Bayesian model averaging, *Water Resour. Res.*, doi:10.1029/2005WR004838, in press.

M. P. Clark, NIWA, Private Bag 8602, Riccarton, Christchurch, New Zealand.

C. G. H. Diks, Center for Nonlinear Dynamics in Economics and Finance, University of Amsterdam, Roetersstraat 11, 1018 WB Amsterdam, Netherlands.

Q. Duan, LLNL, P.O. Box 808, L103, Livermore, CA 94550, USA.

B. A. Robinson and J. A. Vrugt, Earth and Environmental Sciences Division, Los Alamos National Laboratory, Mail Stop T003, Los Alamos, NM 87545, USA. (vrugt@lanl.gov)