The Overweighting of Climatology in Multi-Model Post-Processing Using Kernel Density Methods

Daniel Hodyss, Elizabeth Satterfield, Justin McLay
Naval Research Laboratory,
Monterey, California, USA

Thomas M. Hamill
NOAA Earth System Research Laboratory
Boulder, Colorado, USA

Michael Scheuerer
CIRES, University of Colorado
Boulder, Colorado, USA

To be submitted to *Monthly Weather Review* as a regular article
May 14, 2015

Corresponding author:

Dr. Daniel Hodyss, Naval Research Laboratory, Marine Meteorology Division, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943, email: daniel.hodyss@nrlmry.navy.mil

Abstract

Ensemble post-processing is frequently applied to correct biases and deficiencies in the spread of ensemble forecasts. Kernel density methods (KDMs) addresses the typical biases and under-dispersion of ensembles through the bias correction of ensemble members followed by the generation of a probability density function (PDF) from the weighted sum of kernels fit around each corrected member. The weighting step accounts for the situation where the ensemble is constructed from different model forecasts or generated in some way that creates ensemble members that do not represent equally likely states. In the present work, it is shown that an over-weighting of climatology in KDMs can occur when one first performs a regression-based, state-dependent bias correction before weighting each member. This over-weighting of climatology results in an increase in the mean-squared error of the mean of the predicted PDF. The over-weighting of climatology is illustrated in a simulation study and a real-data study, where the reference is generated through a direct application of Bayes' rule. The real-data example is a comparison of a particular KDM referred to as Bayesian Model Averaging (BMA) and a direct application of Bayes'-rule for ocean wave heights using Navy and National Weather Service global deterministic forecasts.

Introduction

Ensemble prediction methods are now ubiquitous in weather and climate prediction. Multiple forecast simulations are generated, typically initialized with distinct model states that approximate draws from the distribution of analysis uncertainty. Model imperfections may be simulated through the use of multiple forecast models, multiple parameterization suites, a diversity of constants in the parameterizations, and/or stochastic prediction techniques. While ensemble weather prediction systems have improved greatly, the predictions are still frequently affected by systematic errors, including biased ensemble mean forecasts and often an insufficiency of ensemble spread. Consequently, much attention has been paid in recent years to statistical post-processing techniques, whereby the current guidance is adjusted based on relationships noted between past forecasts and observations/analyses. In many circumstances, the goal is to produce a probability density function (PDF) that is as sharp as possible while remaining reliable (Gneiting et al. 2007).

Many approaches have been proposed for statistical post-processing of ensemble forecasts. Non-homogeneous Gaussian regression (NGR, Gneiting et al. 2005) assumes an underlying Gaussian distribution for the posterior and estimates a state-dependent mean and spread of the distribution using the ensemble mean and spread as predictors. Forecast analogs (Hamill and Whitaker 2006, Delle Monache et al. 2013, Hamill et al. 2015) have been demonstrated to work well with large training sample sizes.

Other methods however approach the post-processing problem, either explicitly or implicitly, as a problem in kernel density estimation. In these methods, kernels are placed at pre-determined values of the variable to be post-processed in order to enhance the ensemble size

and/or smooth the resulting shape of the PDF for probabilistic forecasting. Examples of such methods include: Ensemble dressing (Roulston and Smith 2002; Wang and Bishop 2005; Fortin et al. 2006), which increases the ensemble size and spread by adding synthetic new ensemble members centered on existing or adjusted member forecasts. Bayesian Model Averaging (BMA; Raftery et al. 2005, R05 hereafter, Wilson et al. 2007) and related techniques like Ensemble Kernel Density Model Output Statistics (EKDMOS; Glahn et al. 2009) and ensemble regression (Unger et al. 2009) statistically adjust raw ensemble forecast guidance and produce PDFs through a weighted sum of kernels centered on the adjusted forecasts. In BMA, the post-processing takes place in multiple steps, following R05. The first step is a regression correction. If members have exchangeable errors, then following Fraley et al. (2010), the regression correction is applied to the ensemble-mean forecast, and the spread of the raw forecasts is preserved by re-centering the ensemble about the regression-corrected mean. If members do not have exchangeable error statistics, then regression corrections are commonly applied individually to each member. The second step of BMA is the application of kernels to each weighted member. R05 made the implicit assumption that regression-corrected forecasts were similar in quality, so that the kernel standard deviation could realistically be set to the same value for each forecast. With regression-corrected forecasts of very different quality, different kernel standard deviations are possible. In R05, the weights for each forecast and the kernel width were estimated with expectation-maximization methods (EM; Dempster et al. 1977 and references therein), though Vrugt et al. (2008) also showed that Markov-Chain Monte-Carlo methods can also be used.

Since R05, there have been a number of critiques of BMA and related methods. Hamill (2007) discussed issues related to overfitting; Bishop and Shanley (2008) discussed issues with

extreme forecasts. Wilks (2006) critiqued techniques like BMA that regress each member, demonstrating that for longer-lead forecasts, there is a tendency for the regressed values to coalesce toward a single number related to the climatology. In such a situation, BMA typically fits wide kernels to make up for the loss of spread in the original ensemble.

In this manuscript, we show that KDM applied to non-exchangeable ensembles have a problem that has not yet been thoroughly discussed. Namely, these techniques will not produce the minimum-error variance estimate of the mean that could be obtained from the direct application of Bayes' rule; they over-weight the climatological information, resulting in sub-optimal forecast skill. Previous work has illustrated such a direct application of Bayes' rule in statistical post-processing. For example, Krzysztofowicz and Evans (2008) introduced the Bayesian Processor of Forecasts (BPF), whereby climatological data transformed to a normal distribution provides the prior. This prior was updated based on a PDF estimated through a correction of the current forecast using regression relationships estimated from transformed forecast and observational data. Other Bayesian techniques in post-processing include Rajagopalan et al. (2002) and Luo et al. (2007).

To test the hypothesis of KDM sub-optimality, we first choose one of the most popular KDMs as a reference. Henceforth, we focus upon the behavior of BMA. We will first construct a hypothetical scenario of two forecasts which are purposely constructed to lack the correct climatological behavior that random draws from the true distribution would have. In this situation these forecasts should benefit from statistical post-processing that involves convolving them with climatological information. We then examine whether or not the mean of the posterior distribution estimated from BMA is equivalent to the mean estimated from a direct application of Bayes' rule, which provides the minimum error-variance estimate (section 2).

Section 3 extends the analysis to examine whether this overweighing of climatology by KDMs goes away with large numbers of forecasts, i.e. in the limit of large numbers of kernels. Section 4 introduces a new method of direct Bayesian estimation to be used as a control post-processing method in the further evaluation of BMA. This method extends the BPF concepts to work with ensemble data, allowing for members to be equally or unequally likely and accounting for possible correlation between members. Section 5 discusses issues related to the size of training data set; there are challenges with both BMA and Bayesian processing related to the "curse of dimensionality" (Bellman 2003). In section 6, we compare BMA to a Bayesian method through a real-data study, a prediction of wave height based on deterministic global weather forecasts from two operational forecast centers. The Bayesian process in this case will include an additional step, the log transformation of the data; this renders the data in this example to be more Gaussian before the analytic evaluation through Bayes' Rule. Section 7 provides some discussion and conclusions.

2. Differences in the weighting of climatology in Bayes Rule and BMA.

In this section we consider the similarity of BMA and related algorithms to a direct application of Bayes' rule. Differences are examined by considering the smallest-possible ensemble, two forecasts, under the simplest of possible assumptions, forecast with data drawn from a distribution with zero-mean errors. We use this to illustrate that BMA's process of generating state-dependent bias corrections (typically a linear regression step) followed by combining kernels centered around the bias-corrected forecasts will result in an overweighting of climatological information.

Specifically, imagine that the true state for the physical system under consideration to be drawn from a "climatological" pdf we label, $p(x)$, with the property that x is a random draw from $N(\overline{x}, P)$. We have available N_f = 2 unbiased forecasts of today's true state, $x = x_t$, viz.

$$\mathbf{x}_f = \begin{bmatrix} x_f^1 \\ x_f^2 \end{bmatrix} = \begin{bmatrix} x_t + \varepsilon_1 \\ x_t + \varepsilon_2 \end{bmatrix}. \tag{2.1}$$

The perturbations ε_1 and $\varepsilon_2 \sim N(\mathbf{0}, \mathbf{R})$, that is, the errors have zero mean and are Gaussian-distributed with a 2 x 2 covariance matrix **R**. The nth diagonal element of **R** denoted as r_n, and the covariance between the errors in the two forecasts will be denoted as ρ. Equation (2.1) implies that $p(\mathbf{x}_f|x)$ is also Gaussian with the statistics of the perturbations $N(x_t\mathbf{1}, \boldsymbol{R})$, where $\mathbf{1}$ is the one-vector with length N_f. This experimental design is admittedly simpler than what often occurs in actual weather prediction; here forecasts are assumed to have zero bias and no state-dependent error. The simple design here makes the underlying issue with KDMs easy to explain because we have the true posterior available as reference.

a. Bayesian solution.

A straightforward application of Bayes' rule would take the form

$$p(x|\mathbf{x}_f) = \frac{p(\mathbf{x}_f|x)p(x)}{p(\mathbf{x}_f)}, \tag{2.2}$$

where $p(\mathbf{x}_f)$ is $\sim N(\overline{x}\mathbf{1}, P\mathbf{I} + \boldsymbol{R})$. Deriving the posterior for this simple problem can be done analytically in several ways. Parameter estimation through a maximum-likelihood approach is straightforward, i.e., determining the parameters that minimize the log likelihood. Another method would be to attempt to minimize the error variance around a weighted linear estimator (Eq. 2.3a), which would deliver the mean and variance of the posterior. By employing either of these methods one obtains that $p(x|\mathbf{x}_f)$ is $N(\overline{x}_{Bayes}, Q)$ with

$$\overline{x}_{Bayes} = w_1 x_f^1 + w_2 x_f^2 + w_3 \overline{x}, \tag{2.3a}$$

$$Q = w_3 P. \tag{2.3b}$$

The weights are

$$w_1 = \frac{r_2 - \rho}{r_1 + r_2 - 2\rho + \frac{r_1 r_2 - \rho^2}{P}} \tag{2.4a}$$

$$w_2 = \frac{r_1 - \rho}{r_1 + r_2 - 2\rho + \frac{r_1 r_2 - \rho^2}{P}} \tag{2.4b}$$

$$w_3 = 1 - w_1 - w_2. \tag{2.4c}$$

These weights deliver the minimum-error variance estimate and show how the relative weight between the forecasts and climatology is determined in eq. (2.3a). It is important to realize that while these weights do sum to one as seen in (2.4c) it is not correct to interpret these weights as probabilities. In fact, we will show below that these weights can individually be larger than one and in some cases can even be negative. Equations (2.3b) and (2.4c) show that the error variance of the posterior mean in (2.3a) is simply a fraction of the climatological variance P; the weight applied to P is comparatively small when the sum of weights for the original forecasts is large, which happens when forecast-error variances r_1 and/or r_2 are << P.

b. Bayesian Model Averaging.

Here we derive the BMA weights for this same problem following R05. Equation (2) in that article states that for this experimental design

$$p\left(x|x_f^1, x_f^2\right) = m_1 g(x|x_c^1) + m_2 g(x|x_c^2). \tag{2.5}$$

The x_c^n are statistically post-processed forecasts that are created by separately regressing truth against each forecast to obtain coefficients a and b, such that

$$x_c^n = a_n x_f^n + b_n. \tag{2.6}$$

The kernel associated with the nth regression-corrected forecast is thus

$$g(x|x_c^n) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\frac{(x-x_c^n)^2}{\sigma^2}\right]. \tag{2.7}$$

The weights m_n and the kernel width parameter σ are commonly estimated using the expectation-maximization (EM) algorithm. This method can be extended to non-normally distributed variables (as in the Gamma-distribution kernels for quantitative nonzero precipitation of Sloughter et al. 2007; Also see section 6 of this manuscript) or to bivariate distributions (e.g., wind components in Sloughter et al. 2009). As discussed in Fraley et al. (2010), for the case of ensemble forecasts designed to be exchangeable, the forecast regression coefficients and BMA weights should be constrained to be equal. In the present situation, we omit the refinement of σ through CRPS minimization as discussed by Raftery et al. (2005) as the value of σ is not the determining factor in the results to be presented below. We do note that if $r_1 << r_2$ or vice versa, the assumption of a single value of σ is a further approximation, and ideally the EM procedure should fit different values for each member. We have re-run the experiments of this section with a modified EM procedure that obtained different σ values for un-equally likely forecasts and found that this only affects the weights w_1 and w_2 slightly, and definitely not by enough to explain the difference in BMA and the Bayesian technique.

The regression coefficients in eq. (2.6) are

$$a_n = \frac{P}{P+r_n}, b_n = \frac{r_n}{P+r_n}\overline{x}, \tag{2.8a,b}$$

assuming a training set of infinite length. These can also be experimentally verified through regression analysis. By inspection, the regressed forecast is simply a weighted linear combination of the original forecast and the climatological mean. We will see next that setting the weight with respect to climatology here in eqs. (2.8a,b) is a significant issue as the relative weighting between climatology and each forecast has now been fixed.

Recall that the BMA mean is simply

$$\overline{x}_{BMA} = m_1 x_c^1 + m_2 x_c^2. \tag{2.9}$$

This implies that by using eqs. (2.6) and (2.8a,b), we may write the BMA mean estimate as in eq. (2.3a):

$$\overline{x}_{BMA} = w_1^b x_f^1 + w_2^b x_f^2 + w_3^b \overline{x}, \tag{2.10}$$

where

$$w_1^b = m_1 \frac{P}{P+r_1}, \tag{2.11a}$$

$$w_2^b = m_2 \frac{P}{P+r_2}, \tag{2.11b}$$

$$w_3^b = m_1 \frac{r_1}{P+r_1} + m_2 \frac{r_2}{P+r_2}, \tag{2.11c}$$

$$m_1 + m_2 = 1. \tag{2.11d}$$

The question we wish to answer in this section is whether or not the weights (2.4a,b,c) are equal to (2.11a,b,c). In order for (2.4a,b,c) to equal (2.11a,b,c) the EM algorithm must choose the weights m_1 and m_2 appropriately. Note however that this amounts to setting the weights w_1^b, w_2^b, w_3^b equal to their counterparts in eqs. (2.4 a,b,c) by solving for m_1 and m_2. This turns eqs. (2.11 a,b,c,d) into four equations in two unknowns i.e., the system is over-determined. This suggests that eqs. (2.11a,b,c) cannot in all circumstances produce the Bayesian weights and therefore the BMA mean, eq. (2.10), cannot properly produce the minimum error variance estimate.

c. Comparison of Bayesian and BMA methods for two synthetic forecasts.

We now demonstrate the extent to which BMA over-weights climatology when only two forecasts are available. Further, we quantify how much the error in the state estimate is

increased by this over-weighting of climatology. The analytically derived Bayesian and the BMA EM-estimated weights are shown in Figs. 1 and 2, along with the differences in weights. In the construction of Figs. 1 and 2, a training sample of size 10^5 was used for the regression and EM used in BMA, and simulated forecasts and true values were generated consistent with eq. 2.1 and associated text. Samples of forecasts with correlated errors were generated following the methodology of Houtekamer (1993, eq. 13).

Figure 1 demonstrates several interesting differences between Bayesian and BMA weights. At high correlations between forecasts (Figs. 1c, 1f, 1i), the optimal weighting for the Bayesian forecasts is sometimes negative, while BMA constrains the weights to be positive. Consequently, at high correlations, there are large differences in the weight applied to the first forecast between Bayes and BMA; the differences are smaller with lower correlations between forecasts (Fig. 1g). The corresponding weights applied to climatology and their BMA-Bayesian differences are shown in Fig. 2. For lower correlations between forecast errors, Bayesian and BMA methods provide similar weights to climatology when one or both of the forecasts is relatively accurate (Fig. 2g). For high correlation between the forecasts, there are also similar weights when the forecasts have nearly equal magnitudes of error variances (Fig. 2i). On the other hand, at zero correlation between forecasts, when both forecasts have moderate or large error variances, BMA over-weights climatology by a substantial amount, in excess of 10% in many circumstances (Fig. 2g). At very high correlations between forecast errors (Fig. 2i), if one forecast is substantially lower in error than the other, climatology is over-weighted.

Does this overweighting increase the errors of the post-processed forecasts? To explore this, we reduce the training sample size to 100, similar in magnitude to training samples in R05, Wilson et al. (2007), and others. We then verify them with another 100 synthetic forecast and

truth samples. This training and validation procedure was then repeated 10^3 times to generate a large number of simulated Bayesian and BMA forecasts. Fig. 3 provides the root-mean-square error (RMSE) and their differences. Note two particular situations where BMA is notably higher in error than direct Bayesian methods. The first is when there are independent forecasts (Fig. 3g) and moderate to large error variances. The second is when the two forecasts have highly correlated forecast errors, but these errors are significantly different in their variances.

3. The potential to over-weight climatology with many ensemble members.

Consider a situation where the ensemble forecasts were known a priori to have no systematic errors. Generally, one would expect a more accurate resulting pdf from a KDM with a larger ensemble (with associated narrower kernels) than with a smaller ensemble (with wider kernels). Is it possible then that the issues with BMA over-weighting climatology demonstrated in section 2 were due to the use of a small ensemble (of size 2)? In this section we demonstrate that the over-weighting of climatology can still occur with large ensembles.

This overweighting is illustrated here with a simple experimental design, one for which we can analytically derive the correct Bayesian result in the limit of very large numbers of forecasts. Here, the forecasts are assumed to be equally likely and have uncorrelated errors. We fully realize that this is not a realistic model of the errors for an ensemble of weather predictions, especially short-lead predictions; in such situations the forecast errors are likely to have some correlation. This experimental design is rather meant to illustrate that KDMs do not necessarily converge to the correct predicted mean as the number of forecasts approaches infinity.

As in Section 2, imagine the climatology for the physical system under consideration to be drawn from $p(x)$, with the property that x is a random draw from $N(\overline{x}, P)$. We have available N_f unbiased forecasts of today's true state, $x = x_t$, viz.

$$\mathbf{x}_f = x_t \mathbf{1} + \boldsymbol{\varepsilon}, \tag{3.1}$$

where $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{R})$. $\mathbf{1}$ denotes an N_f x1 vector whose entries are 1. The perturbations $\boldsymbol{\varepsilon}$ have the property that $\mathbf{R}$ is N_f x N_f. The diagonal of $\mathbf{R}$ will be constant and denoted as r; off-diagonal elements of $\mathbf{R}$ are zero. Equation (3.1) thus implies that $p(\mathbf{x}_f|x)$ is Gaussian with the statistics of the perturbations, i.e. $N(\mathbf{0}, \boldsymbol{R})$.

a. Bayesian solution

A straightforward application of Bayes' rule would take the form of eq. (2.2). By using the information denoted above in equation (3.1) we obtain that $p(x|\mathbf{x}_f)$ is $N(\overline{x}_{Bayes}, Q)$ with

$$\overline{x}_{Bayes} = \overline{x} + P\mathbf{1}^T[\mathbf{1}\mathrm{P}\mathbf{1}^T + \boldsymbol{R}]^{-\mathbf{1}}(\mathbf{x}_f - \overline{x}\mathbf{1}), \tag{3.2a}$$

$$Q = (1 - P\mathbf{1}^T[\mathbf{1}P\mathbf{1}^T + \boldsymbol{R}]^{-\mathbf{1}}\mathbf{1})P. \tag{3.2b}$$

The Sherman-Morrison-Woodbury formula (Golub and Van Loan 1989, section 2.1.3) states that

$$[\mathbf{1}P\mathbf{1}^T + \boldsymbol{R}]^{-1} = \boldsymbol{R}^{-1} - \frac{P}{r^2 + rPN_f}\mathbf{1}\mathbf{1}^T, \tag{3.3}$$

and when this is used in eqs. (3.2a,b), we obtain

$$\overline{x}_{Bayes} = \frac{r}{r+PN_f}\overline{x} + \frac{P}{r+PN_f}\mathbf{1}^T\mathbf{x}_f, \tag{3.4a}$$

$$Q = \left(1 - \frac{PN_f}{r+PN_f}\right)P. \tag{3.4b}$$

Note that the vector product in eq. (3.4a) is the sum of the forecasts:

$$\mathbf{1}^T \mathbf{x}_f = \sum_{n=1}^{N_f} x_f^n, \tag{3.5}$$

and that (3.1) implies, in the sense of probability, that

$$\lim_{N_f \to \infty} \frac{1}{N_f} \sum_{n=1}^{N_f} x_f^n = x_t. \tag{3.6}$$

Using eqs. (3.5) and (3.6) in eq. (3.4a,b) reveals that the Bayesian solution has the property that the posterior mean approaches the truth, and the posterior variance vanishes, as $N_f \to \infty$.

b. Bayesian Model Averaging

Recall with BMA that the PDF is constructed with weighted kernels. Following R05, their eq. (2):

$$p\left(x|\mathbf{x}_f\right) = \sum_{n=1}^{N_f} w_n g_n(x|x_c^n). \tag{3.7}$$

Here the x_c^n are the statistically post-processed forecasts. Note that eqs. (2.8a,b), when applied to the model of equation (3.1), implies that

$$a_n = \frac{P}{P+r}, \; b_n = \frac{r}{P+r}\overline{x}, \tag{3.8a,b}$$

when provided with a training set of infinite length; the coefficients of eqs. (3.8a,b) are identical for all forecasts because the forecasts are equally likely. This is precisely the reason we have chosen equally likely forecasts, as this leads to a case in which we can predict precisely the regression coefficients and the result of the EM algorithm (e.g. Fraley et al. 2010). In this case we know that the weights are all equal to $1/N_f$.

Putting these results together allows one to show that the mean from BMA is

$$\overline{x}_{BMA} = \sum_{n=1}^{N_f} w_n x_c^n = \frac{r}{P+r}\overline{x} + \frac{P}{P+r}\frac{1}{N_f}\sum_{n=1}^{N_f} x_f^n. \tag{3.9}$$

Making use of eq. (3.6) shows that the BMA mean in the limit of an infinite number of members in eq. (3.7) equals

$$\overline{x}_{BMA} = \frac{r}{P+r}\overline{x} + \frac{P}{P+r}x_t. \tag{3.10}$$

Therefore, the BMA mean in the limit of an infinite number of members is a weighted average between the climatology and the truth, while the Bayesian solution (eqs. 3.4a,b) is simply the truth in that limit. Another way to see this over-weighting behavior of BMA is to note that eq. (3.10) states that, in the limit of an infinite number of kernels,

$$|\overline{x}_{BMA} - x_t| \leq |\overline{x} - x_t|, \tag{3.11}$$

with equality only in the trivial case of $P = 0$. Even though BMA cannot provide the correct weights in this situation note that it is closer to the truth than the climatological mean and hence this explains why one does obtain benefit from BMA in practice.

To the extent that this experimental design resembles possible ensembles, it shows that BMA can substantially over-weight climatology. Note however that in this experimental design, had the regression analysis been applied to $\overline{x}$ instead of to the individual members, it can be shown that the posterior mean would be identical to that from Bayes' rule. Therefore, the example of this section is simply meant to illustrate the behavior in the asymptotic limit of a large number of kernels. Recall that the typical result in KDMs is that the estimated PDF from KDMs become more accurate as the number of kernels increases. By contrast however we have shown that increasing the number of kernels (forecasts) does not eliminate this issue of overweighing climatology because this issue occurs in BMA because of the initial "bias-correction" step fixing the relative weight between the climatological mean and the forecasts and, as shown in section 2, this occurs more generally than just the specific experimental design of this section.

4. A particle-filter inspired procedure simulating Bayesian post-processing.

Sections 2 and 3 attempted to show that by regressing the forecasts to the truth *and then* weighting each with its own kernel was likely to lead to the wrong relative weight on each forecast as compared to the Bayesian result that delivers the minimum error variance estimate. This section will illustrate a very simple method that does not begin by regressing the forecasts to the truth and therefore does not suffer from this issue.

We begin by writing Bayes' rule again as

$$p(x|\mathbf{x}_f) = \frac{p(\mathbf{x}_f|x)p(x)}{p(\mathbf{x}_f)}. \tag{4.1}$$

Note that we may use the chain rule of probability to write the forecast likelihood as

$$p(\mathbf{x}_f|x) = p(x_f^1|x)p(x_f^2|x, x_f^1) \dots p(x_f^i|x, x_f^1, x_f^2, \dots, x_f^{i-1}), \tag{4.2}$$

where i = 1, … , N_f. Equation (4.2) suggests that we might begin by creating a succession of likelihoods from each forecast conditioned on specific other forecasts and truth. We will illustrate the procedure using Gaussian functions as representing the likelihoods, but the choice of what function to use should be made based on consideration of the characteristics of the data set. An example of how to choose these functions will be provided in section 6, and we present a method to fit generalized Gaussians in Appendix B.

We will assume below that the training set consists of N_f pairs of forecasts and their verification, which may be the analysis from a data assimilation system or observations. Because eq. (4.1) expresses the climatological pdf as separate from the forecast likelihoods, we may deduce the climatological pdf from an archive distinct from our set of truth-forecast pairs. Being separate, this archive may possibly be larger, and hence we refer to the number of members in our climatological archive as N_c.

The method to create this succession of likelihoods we use here proceeds as follows. We begin by scanning the training set for the ordering of the quality of the forecasts. We then sort the forecasts from best to worst, ranked by root-mean-squared-errors (RMSE). We start by fitting the best forecast with a Gaussian distribution:

$$p\left(x_f^1|x\right) = N_1 exp\left[-\frac{1}{2}\frac{\left(x_f^1 - H_1(x)\right)^2}{R_1}\right], \tag{4.3}$$

where N_1 is simply the normalization for the pdf and $H_1(x)$ is the function that results from regression for which the truths are the predictors and the x_f^1 are the predictands. Note that $H_1(x)$ is *not* what is commonly referred to as "state-dependent bias correction"; $H_1(x)$ maps the truth to the forecast, while standard state-dependent bias correction maps the forecast to the truth. This is important, as the likelihood $p\left(x_f^1|x\right)$ must be the distribution of x_f^1, and if one applied a "state-dependent bias correction" one would obtain the wrong distribution. This distinction is critical as this aspect is what allows this procedure to avoid overweighting climatology. The variance R_1 is obtained by calculating the variance across the training set of the difference between the x_f^1 and the regressed truth [i.e. $\left(x_f^1 - H_1(x)\right)$]. Additionally, it is important to recognize that $R_1 \neq r_1$ and that this is true for all forecasts.

This fitting procedure is then extended to the ith forecast, viz.

$$p\left(x_f^i|x, x_f^1, x_f^2, \dots, x_f^{i-1}\right) = N_i exp\left[-\frac{1}{2}\frac{\left(x_f^i - H_i\left(x, x_f^1, x_f^2, \dots, x_f^{i-1}\right)\right)^2}{R_i}\right], \tag{4.4}$$

where the function $H_i\left(x, x_f^1, x_f^2, \dots, x_f^{i-1}\right)$ is simply the function that results from multi-variate regression for which the predictors are the $x, x_f^1, x_f^2, \dots, x_f^{i-1}$ and the predictand is x_f^i. As before, the error variance R_i is the variance of $\left(x_f^i - H_i\left(x, x_f^1, x_f^2, \dots, x_f^{i-1}\right)\right)$ across the training set.

For completeness, we mention that one could simplify the succession of products in (4.2) by noting that because they are exponential the products can be re-written as a sum within the exponential function, i.e.

$$p(\mathbf{x}_f|x) = N_1 N_2 \ldots N_i exp\left[-\frac{1}{2}S\right], \tag{4.5}$$

$$S = \frac{\left(x_f^1 - H_1(x)\right)^2}{R_1} + \frac{\left(x_f^2 - H_2\left(x, x_f^1\right)\right)^2}{R_2} + \cdots + \frac{\left(x_f^i - H_i\left(x, x_f^1, x_f^2, \ldots, x_f^{i-1}\right)\right)^2}{R_i}. \tag{4.6}$$

After all the forecast likelihoods have been created we must now create the climatological PDF, $p(x)$. There are at least two ways to do this. The first way is to simply fit the climatological distribution to the characteristics of the training set. For example, we may calculate the mean, $\overline{x}$, and variance, P, of the N_c members of the distribution of the truth across our training set. This information allows us to simply fit the climatological pdf as

$$p(x) = N_c exp\left[-\frac{1}{2}\frac{(x-\overline{x})^2}{P}\right]. \tag{4.7}$$

Once (4.7) is determined all the information to evaluate equation (4.1) has been obtained and therefore the posterior pdf can simply be evaluated for whatever probabilistic prediction is required.

An alternative to fitting climatology to a function is to use a kernel density estimation procedure to represent the climatological pdf. The simplest way to do this is to use the framework of particle filtering (Doucet et al. 2000). In the particle filtering framework we assume that our samples from climatology are equally likely and have kernels that have zero width (i.e. they are the Dirac delta function rather than a Gaussian, as in BMA). We emphasize however that there is no requirement that the kernels have zero width; they could be Gaussians as in the equations for BMA [for example, eq. (3.7)]. The key defining difference between this technique and that of BMA is that BMA is applying kernel density estimation to estimate the

posterior, while under this technique kernel density estimation is used to represent the prior (climatological) distribution. This distinction is what will allow the following procedure to not overweight climatology.

As an example, we assume kernels of zero width (Dirac delta functions) to represent the climatological distribution. This allows one to determine the probability that the jth sample from the climatological distribution, x_j, is the true state given the forecasts, viz.

$$p_j = \frac{p(\mathbf{x}_f|x_j)}{\sum_{j=1}^{N_c} p(\mathbf{x}_f|x_j)}\,, \tag{4.8}$$

where N_c is the number of states from climatology that we have available in our training set. Hence, eq. (4.8) is essentially an "analog" approach in which we search the training set for states from climatology that, as measured by our forecast likelihoods, are likely to be today's truth. Note that the probability that the jth sample from the climatological distribution is the true state given the forecasts is also the weight for each of our delta function kernels. Hence, we may calculate the mean of the posterior as

$$\bar{x}_b = \sum_{j=1}^{N_c} p_j x_j. \tag{4.9}$$

Other moments of the posterior may be calculated similarly.

In some applications, an ensemble of equally likely members proves useful. To obtain this from eq. (4.8), we sort the x_j from smallest to largest and label this new set as x_j^s. We then re-organize the p_j into the same ordering as the x_j^s to obtain p_j^s. The p_j^s can now be cumulatively summed to determine the cumulative distribution function (CDF). The procedure for sampling from a CDF is well known and goes as follows. Draw a uniformly distributed random number on 0 to 1. Next, find the element of the CDF closest to this number. This element of the CDF corresponds to a particular x_j^s and therefore this value is the correct random

draw from the posterior, eq. (4.1). Repeat this procedure any number of times to obtain the ensemble of equally likely members.

The appropriateness of the choice between fitting climatology to a function or representing it with a kernel density method, such as the above described particle filtering algorithm, usually comes down to the size of the training set. If the training set is very small (less than 100 samples), then one is likely to do better by simply fitting the climatological distribution to some known function with reasonably accurate characteristics. Given a relatively larger climatological training sample (100s or more) one could effectively employ the kernel density algorithm using either Gaussian kernels or Dirac delta kernels (i.e. particle filter). The choice as to which method performs best is likely to be data-set dependent and is left for future work.

Lastly, we note that the direct application of Bayes' rule for use in post-processing is not a new concept. The idea of leveraging prior climatological information and updating with NWP information was presented in the "Bayesian Processor of Forecasts" of Krzysztofowicz and Evans (BPF; 2008). The above procedure explicitly illustrates how to extend to the multi-model ensemble case through the use of the chain rule of probability to compute the forecast likelihood. Further, this procedure directly accounts for correlations between forecasts and is flexible enough to be applied to equally likely ensemble members, subsets of equally likely members, or unequally likely members.

5. Effects of limited training sample size

In this section we will compare BMA to the Bayesian method of section 4 for different numbers of forecasts and different training lengths. The question to be answered in this section

is whether or not there is some advantage for small training size to performing state-dependent bias correction and then weighting each forecast (as in BMA) when compared to the method of section 4.

The problem we set up will be that of section 3, but we will now allow for unequally likely forecasts. As in section 3, we imagine climatology to be drawn from $p(x)$, with the property that x is a random draw from $N(1,1)$. We define the forecasts as in eq. (3.1), and as in section 3, the forecasts will not be correlated here. (The same experiments as to be described below were run for correlations as high as 0.98 and the result that the Bayesian technique had smaller MSE was also found.) Two cases will be run, one with un-equally likely forecasts, and the other with equally likely forecasts. In the equally likely forecast case, all forecasts will have r = 0.5. In the un-equally likely case, the ith forecast will have an error variance of $r_i = 1/10 + (9/10)i/N_f$, where we have chosen this function to bound the forecast quality between 0.1 and 1 for any number of forecasts. The BMA algorithm will be implemented as described in section 3, and the number of iterations in the EM algorithm is set to 50.

After these calculations have been performed, we calculate the mean of the predicted distribution and then calculate the RMSE with respect to the truth for each technique. We will repeat this entire calculation for 10^5 verification trials for each combination of number of forecasts and training size. For each of these trials, a different truth is drawn from climatology and truth samples are created according to eq. (2.1). The resulting average over the 10^5 verification trials is reported as the difference between the RMSE of the above mentioned Bayesian technique and BMA (Fig. 4). The red line in the upper left corner of the figure denotes the region for which the number of forecasts is equal to or greater than the training size. Experiments in this region were not performed, as we do not believe that either technique can be

expected to deliver sensible results when the number of forecasts (predictors) is greater than the training size. Positive values in Fig. 4 indicate that BMA had larger RMSE averaged over the 10^5 trials than the Bayesian technique. By scanning both panels of Fig. 4, one can see that the Bayesian method has smaller RMSE over a wide range of training sizes and numbers of forecasts. In fact, only one experiment resulted in a very weak negative result, and that is for the equally likely forecast case and a training size of 16 and the number of forecasts equal to 16. Therefore, Fig. 4 shows that this difference between the Bayesian result and BMA's overweighting of climatology is robust in the presence of sampling error from limited training size.

6. An application to wave-forecast post-processing

In this section we apply the previously developed ideas to multi-model forecasts of ocean significant wave heights (units will be in meters) in the North Pacific Ocean. All forecasts are obtained using the Wavewatch III (Tolman 1997, 1999, 2001) global ocean wave model implemented at two U. S. forecasting centers, Fleet Numerical Meteorology and Oceanography Center (FNMOC) and the National Centers for Environmental Prediction (NCEP). Wind data are obtained from the Navy Operational Global Atmospheric Prediction System (NOGAPS) (Rosmond, 1992) model and the Global Forecast System (GFS, Han and Pan 2011), respectively.

We will use two forecasts as in section 2, one from each implementation of Wavewatch III. In the following we consider the +120-h forecast lead time and obtain data at 1×1 degree resolution. The verification data set will be taken as the FNMOC Wavewatch III significant wave height analysis. We will post-process each grid point of these fields that contain a wave height forecast from both models and for grid points between 20° and 60° North latitudes and

between 120° and 270° East longitudes. The training set for this experiment will be the period from 01 December 2011 to 31 January 2012. The climatological mean and variance over this training set is plotted in Fig. 5. We will apply this training set to post-process the forecasts for 01 February to 05 February 2012.

Wave height forecasts have the property that they are positive definite, which leads in these forecasts to highly skewed distributions with long tails. Clearly then both BMA and the Bayesian techniques must be constructed to account for this fact. For BMA we will use a power transformation (Following Yeo and Johnson 2000) such that a variable u is mapped to "log" space through the transform:

$$u \rightarrow log(1+u). \tag{6.1}$$

This transformation into log-space effectively pulls the tails of the distribution inward towards the center of the distribution, which results in a training dataset in log-space that is more accurately fitted to a Gaussian distribution. We apply this to all data required by the algorithm, perform the BMA algorithm, and then map a variable v back to physical space using:

$$v \rightarrow exp(v) - 1. \tag{6.2}$$

It can be shown using the Jacobian of this transformation to be equivalent to using BMA kernels in (3.7) of the form

$$g(x|x_c^n) = \frac{N}{1+x} exp\left[-\frac{1}{2}\frac{(Lx-Lx_c^n)^2}{\sigma^2}\right], \tag{6.3}$$

where $Lx = log(1+x)$ and Lx_c^n is the nth bias corrected forecast in log-space. We will use this representation to plot the predicted distributions from BMA.

For comparison we will also implement the particle-filter inspired method of section 4. Because the training set for climatology only has 60 samples, we choose the version that simply fits the climatological distribution to a function. To begin, we must also account for the positive-

definiteness and big tails of the wave height distributions in both the forecast likelihoods and in the climatological distribution. For the forecast likelihoods we first transform the data as in eq. (6.1) and then apply the equations of section 4 to the transformed data. To map back we simply use the Jacobian of the transformation to obtain forecasts likelihoods of the form

$$p\left(x_f^1|x\right) = \frac{N_1}{1+x_f^1} exp\left[-\frac{1}{2}\frac{\left(Lx_f^1 - H_1(Lx)\right)^2}{R_1}\right], \tag{6.4}$$

where $Lx_f^1 = log\left(1 + x_f^1\right)$. It is important to note that the regression for the function H_1 is done with both the predictor and the predictand in log-space. We choose here to define the function H_1, and H_2 below, as the function that results from linear regression, but note that if the training set was larger we would prefer to use quadratic or even cubic polynomial regression. Similarly, if the training set were larger, we could use a generalized Gaussian (which would fit the first, second, and fourth moments, rather than just the first and second; please see Appendix B) and would allow for a better prediction of the higher moments of the Bayesian posterior. This generalized Gaussian was applied to this dataset but no statistically distinguishable improvement was found.

The second forecast's likelihood is obtained from

$$p\left(x_f^2|x, x_f^1\right) = \frac{N_2}{1+x_f^2} exp\left[-\frac{1}{2}\frac{\left(Lx_f^2 - H_2\left(Lx, Lx_f^1\right)\right)^2}{R_2}\right], \tag{6.5}$$

where $Lx_f^2 = log\left(1 + x_f^2\right)$.

Finally, we must construct the climatological distribution. We do this in the same way as the forecast likelihoods by transforming the wave data into log-space, fitting a Gaussian, and transforming back using the Jacobian to obtain

$$p(x) = \frac{N_c}{1+x} exp\left[-\frac{1}{2}\frac{\left(Lx - \overline{Lx}\right)^2}{P}\right], \tag{6.6}$$

where we again emphasize that $\overline{Lx}$ and P are the climatological mean and variance in log-transformed space.

An example of the resulting functions defined above is plotted in Fig. 6 for the grid point at 50°N, 180°E. At this grid point, the GFS forecasts happened to have less error over the training period and were assigned as f_1. Subsequently, the NOGAPS forecasts were assigned the variable f_2. In Fig. 6a we plot the forecast likelihood describing the distribution of GFS forecasts given the verification [e.g., eq. (6.4)]. By comparing the shape of this distribution to the one-to-one line, we can see that the GFS forecast typically over-forecasts when the true wave height is small and under-forecasts when the true wave height is large. Additionally, note that the width of the distribution increases as the true state being conditioned upon increases. This states that the variance in the forecast increases as the true state being conditioned upon increases and is a common property of positive definite distributions.

Figures 6b and 6c provide the forecast likelihood for the NOGAPS forecasts given the verification and the GFS forecasts [e.g., eq. (6.5)]. In Fig. 6b, we plot f_2 as a function of the verification and evaluated for f_1 equal to the GFS forecast on Feb 01, which is the first day after the training period. In Fig. 6b, we see that the variability in the NOGAPS wave height forecast is significantly larger than that of the GFS forecast, though it too typically over-forecasts when the true wave height is small and under-forecasts when the true wave height is large. We believe this larger variance in the NOGAPS forecast is due to the fact that it is being run at a lower resolution than the GFS model and therefore cannot develop strong-enough winds to force the wave model to large wave heights. In Fig. 6c, we plot f_2 as a function of f_1 and evaluated for x equal to the verification on 00 UTC, 01 Feb 2012. This plot shows the implied correlation between f_2 and f_1 as seen by the training set, and shows that there is generally a weak correlation

between them. In summary, these types of plots of the forecast likelihoods provide a more truly Bayesian way to perform forecast validation by showing the relationship between the forecasts and the truth and the forecasts between each other.

Evaluating Bayes' rule in eq. (4.1) requires a probability density estimation of the climatological distribution. In Fig. 6d we show the pdf for the climatological distribution [e.g., eq. (6.6)]. The histogram in Fig. 6d is composed of the 60 samples at this grid point and shows that the log-normal fit is a reasonable choice for such a small number of samples. In Fig. 6e, we show the resulting Bayesian posterior distribution evaluated for f_2 equal to the GFS forecast on Feb 01. Finally, in Fig. 4f we show the resulting BMA posterior, which we obtained after using eq. (6.3) in eq. (2.5). This example was chosen because it illustrates one of the ways that BMA tends to overweight climatology. By comparing Figs. 4e and 4f, one can see that the BMA posterior is generally significantly wider and has the bulk of its probability mass fixed near the climatological mean of approximately 4 meters. Additionally, one can see that the two kernels in the BMA estimation procedure become bimodal below a ~2-m. wave height, which we believe is unwarranted.

Further evidence that the BMA procedure is overweighting climatology is presented in Fig. 7. In Fig. 7a, we show the mean squared error (MSE) with respect to the verification for the five days after the training period and averaged as a function of latitude. Additionally, we show the MSE for the same period for the climatological mean. Note that the Bayesian procedure of section 4 has a lower MSE than BMA for all latitudes except near 60°N. Note that near 60°N. the climatological variance is much larger than the MSE of the climatological mean. Hence, a procedure that erroneously over-weights climatology will appear to be accurate in the situation where the climatological variance is larger than it should be.

In Fig. 7b, we show the mean squared difference (MSD) between the posterior mean from the procedure in section 4 and its three components (climatological mean, NOGAPS, and GFS forecasts) as discussed in section 2, equation (2.3a). We repeat these calculations for the MSD between the posterior mean from BMA and its three components (climatological mean, NOGAPS, and GFS forecasts) as we know from equation (2.10). The point here is that when the MSD is small this implies that the weight on that component of the posterior mean must be large. Similarly, when the MSD is small for a particular component of the posterior mean then the weight on that component must be small. Figure 7b shows that the posterior mean from BMA has a smaller MSD to the climatological mean than the Bayesian procedure of section 4. Similarly, figure 7b shows that the posterior mean from BMA has a larger MSD to the two forecasts than the Bayesian procedure. We feel that this shows that the theoretical results presented in section 2 can be directly seen in a real application of multi-model post processing.

Lastly, we show in Fig. 7c that the binned-spread diagram for both techniques is very nearly identical. This diagram is created from all the data points across the Pacific and over the five days of verification. We believe that this shows that both techniques have a reasonable relationship between the width of the posterior and their squared errors. The impact of the over-weighting of climatology on the variance prediction appears too small to detect in the moments higher than the first in this example with such a small training set. Further study of the impact of the size of the training set on the quality of the higher moments will be presented in a sequel.

7. Summary and Conclusions

We have shown post-processing techniques like BMA that first apply a state-dependent bias correction and then apply a weighting to ensemble members will systematically over-weight

climatology. This problematic treatment of climatological, or potentially other non-NWP information, results in an increase in the mean-squared error of the resulting state estimate. We demonstrated that this result held for in various parameter spaces, including equally likely and correlated ensemble forecasts. We also showed that this result was independent of the ensemble size and the size of the training set.

We note that the issue of overweighting climatology does not arise because of the kernel-density estimation assumption, but rather because the operation of bias correcting first prematurely fixes the relative weight to climatology. We present an alternative approach based on a direct Bayesian estimation. This approach does not rely on an explicit bias-correction step, but rather relies on accurately fitting the dataset to the appropriate distributions required by Bayes' rule. Our approach can be considered an extension of the Bayesian Processor of Forecasts (BPF) of Krzysztofowicz and Evans (2008), in that this work extends the BPF method to the ensemble case and directly accounts for correlation through the chain rule of probability. In addition, we carefully detail a procedure to construct likelihoods based on a function that maps the true state to the forecast. We also present a particle-filter based formulation of this procedure. We demonstrate the ability of this method to extend to non-Gaussian, distributions through a log transformation, by application to multi-model ocean significant wave heights.

We feel that the most powerful application of the Bayesian technique presented here will be obtained for larger training sizes than we had available. The study of how to increase the training set size as well as what size training set is required to employ higher-order function fitting procedures will be the subject of future work.

Appendix A

Post-processing for the Mean and Variance of a Single-Model Ensemble

In this appendix we generalize Satterfield and Bishop (2014) to the complete Bayesian solution to show how one might incorporate the ideas of this manuscript into the ensemble post-processing problem. In the original work of Satterfield and Bishop (2014) a linear estimate of the posterior mean was used. Here we show how to solve their ensemble post-processing problem using the methods of this manuscript and without making the linear assumption.

We imagine the true state, x_t, to be a scalar and that it is drawn from a climatological distribution whose probability density function (pdf) we label $\rho(x_t)$. We imagine data assimilation has been taking place for some time and the forecasts that we have from this data assimilation procedure have been attempting to incorporate information from observations of the true state. We assume that the forecasts we have available are related in some unknown way to the distribution of possible true states:

$$m_1 = x_t + \sigma\varepsilon, \tag{A.1}$$

$$m_2 = a\sigma^2\delta, \tag{A.2}$$

where m_1 and m_2 are the sample first (mean) and second moments (variance), respectively. The σ^2 denotes the actual error variance given past observations of the first sample moment about the true state and will be assumed to be a random variable in the same sense as x_t, while ε will, for simplicity, be assumed to be drawn from $N(0,1)$.

We note that there are four random variables of importance: $x_t, \sigma^2,\ m_1, m_2$. We begin by writing the joint distribution of these four random variables using the chain rule of probability. Given four random variables there are 4! = 24 different ways to employ the chain rule. Because we are interested in $p(x_t|m_1, m_2)$ we know that we need to marginalize with respect to σ^2. Given this requirement we write:

$$p(x_t, \sigma^2, m_1, m_2) = p(x_t)\rho(m_1|x_t)\, p(m_2|m_1, x_t)\, p(\sigma^2|m_1, m_2, x_t)$$

$$= p(m_1)\rho(m_2|m_1)p(\sigma^2|m_1, m_2)p(x_t|m_1, m_2, \sigma^2). \qquad \text{(A.3)}$$

Equation (A.3) shows that

$$p(x_t|m_1, m_2, \sigma^2) = \frac{p(x_t)\rho(m_1|x_t)\, p(m_2|m_1,x_t)\, p(\sigma^2|m_1,m_2,x_t)}{p(m_1)p(m_2|m_1)p(\sigma^2|m_1,m_2)}. \qquad \text{(A.4)}$$

Note that we must marginalize equation (A.4) to be obtain our sought after density, viz.

$$p(x_t|m_1, m_2) = \int_{-\infty}^{\infty} p(x_t|m_1, m_2, \sigma^2)\ p(\sigma^2|m_1, m_2)d\sigma^2 \qquad \text{(A.5a)}$$

$$= \frac{p(m_1|x_t)p(m_2|m_1,x_t)}{p(m_1)p(m_2|m_1)} p(x_t). \qquad \text{(A.5b)}$$

Note that the integral in (A.5a) is equivalent to a stochastic re-sampling with respect to σ^2 of the posterior density in (A.4). In this sense this integral is no different than the definition of the t-distribution. In any event, equation (5) is the correct version of Bayes' rule for this problem and requires our being able to determine $p(m_1|x_t)$ and $p(m_2|m_1, x_t)$, which we illustrate next.

To determine $p(m_1|x_t)$ we note that the mean of this density is x_t but whose shape is determined from the distribution of the product of σ and ε. Hence, we must determine the distribution of the product of σ and ε. Because σ and ε are independent their joint density is

$p(\sigma, \varepsilon) = p(\sigma)p(\varepsilon)$. We begin by defining a new variable $z = \sigma\varepsilon$. Hence, a reasonable transformation is $(\sigma, \varepsilon) \to (z, \tilde{\varepsilon})$. This implies an inverse transformation of $\left(\frac{z}{\tilde{\varepsilon}}, \tilde{\varepsilon}\right) \to (\sigma, \varepsilon)$, which defines the determinant of the Jacobian of the inverse transformation as

$$\begin{vmatrix} \frac{1}{\tilde{\varepsilon}} & -\frac{z}{\tilde{\varepsilon}^2} \\ 0 & 1 \end{vmatrix} = \frac{1}{|\tilde{\varepsilon}|}. \tag{A.6}$$

This shows that the joint density is therefore $p(z, \tilde{\varepsilon}) = p_\sigma\left(\frac{z}{\tilde{\varepsilon}}\right) p_\varepsilon(\tilde{\varepsilon}) \frac{1}{|\tilde{\varepsilon}|}$ and therefore the sought after density of z is

$$p(z) = \int_{-\infty}^{\infty} p_\sigma\left(\frac{z}{\tilde{\varepsilon}}\right) p_\varepsilon(\tilde{\varepsilon}) \frac{1}{|\tilde{\varepsilon}|}\, d\tilde{\varepsilon}. \tag{A.7}$$

Note however that we know that the distribution of σ^2 is Inverse Gamma, but we do not actually know yet the distribution of σ. To this end we again assert a transformation but this time we define the new variable as $u = \sqrt{\sigma^2}$, which implies that $p_\sigma(u) = \frac{C}{u^{2\alpha+2}} exp\left(-\frac{\beta}{u^2}\right)|2u|$. Because we know α and β from the climatological distribution of σ^2 and we know that $p_\varepsilon(\tilde{\varepsilon})$ is normal with mean zero and variance one we may evaluate equation (A.7), which defines $p(m_1|x_t)$.

To determine $p(m_2|m_1, x_t)$ we first note that

$$p(m_2|m_1, x_t) = \int_{-\infty}^{\infty} p(m_2|\sigma^2)\ p(\sigma^2|m_1, x_t) d\sigma^2. \tag{A.8}$$

Application of Bayes' rule allows the integrand to be re-written using

$$p(\sigma^2|m_1, x_t) = \frac{p(m_1|\sigma^2, x_t)}{p(m_1|x_t)} p(\sigma^2), \tag{A.9}$$

where we repeat that $p(\sigma^2)$ is a known Inverse Gamma distribution and $p(m_1|\sigma^2, x_t)$ is $\mathrm{N}(x_t, \sigma^2)$. Lastly, we fit $p(m_2|\sigma^2)$ as a Gamma distribution under the assumption that sampling

error is the main issue, i.e. it has a mean $a\sigma^2$ and variance equal to $2a^2\sigma^4/(K-1)$, where K is the ensemble size. Given equations (A.7), (A.8), and (A.9) we have now completed construction of equation (A.5).

Appendix B

Fitting to a Generalized Gaussian

We fit a variable u to a generalized Gaussian by first calculating its mean, $\bar{u}$, variance, θ, and fourth moment, F. The generalized Gaussian is defined as

$$p(u) = Nexp\left[-\left(\frac{|u-\bar{u}|}{\sqrt{2}\sqrt{\beta}}\right)^{\alpha}\right] \tag{B.1}$$

where N is simply the normalization and the parameters α and β are to be fit to the variance and fourth moment. We do this using an iterative procedure in which we iterate for α_i by using the previous α_{i-1} in,

$$g_i = \frac{\Gamma\left(\frac{3}{\alpha_{i-1}}\right)^2}{\Gamma\left(\frac{5-\alpha_{i-1}}{\alpha_{i-1}}\right)\Gamma\left(\frac{1}{\alpha_{i-1}}\right)}\frac{F}{\theta^2}, \tag{B.2}$$

to obtain the next update

$$\alpha_i = \frac{5}{1+g_i}. \tag{B.3}$$

Typically, much less than 20 iterations are required for convergence. Once α is known β may be calculated from

$$\beta = \frac{\Gamma\left(\frac{1}{\alpha}\right)}{2\Gamma\left(\frac{3}{\alpha}\right)}\theta. \tag{B.4}$$

Note that if the distribution of u is a standard Gaussian such that $F = 3\theta^2$ then $\alpha = 2$ and $\beta = \theta$, which reduces (B.1) to a standard Gaussian.

Figure Captions

Figure 1: Weights for the first of two members in a forecast from Bayes (top row) and BMA (second row). Weight differences are shown in the third row. Weights and differences are plotted as functions of the ratio of the first forecast's error variance divided by the climatological variance (x), the second forecast's error variance divided by the climatological variance (y), and the correlation between the two forecasts (the three columns).

Figure 2: As in Fig. 1, but here for the weight applied to the climatology. Note that the weight of the second forecast, not shown, is in each case one minus the weight of the first forecast and climatology.

Figure 3: As in Fig. 1, but for RMS errors of the mean for Bayesian and BMA procedures as generated from 1,000 Monte-Carlo experiments, as described in the text.

Figure 4. MSE difference between BMA and the Bayesian particle-filter inspired method of section 4. Panel (a) shows results from the experiment with unequally likely forecasts and panel (b) shows results from the experiment with equally likely forecasts. The diagonal red line in the upper-left corner of each plot denotes the region above which the training size is smaller than the number of predictors, and for which experiments were not performed.

Figure 5. The (a) climatological mean and (b) standard deviation across the training set at each grid point for the period 01 December 2011 to 31 January 2012.

Figure 6. The fitting of the likelihoods and climatology for a grid point in the North Pacific (50° N, 180° E) and evaluated for 01 Feb 2012. In (a) is the forecast likelihood for the GFS model given the truth. In (b) and (c) are the forecast likelihood for the NOGAPS model given the truth

and the GFS forecast. In (d) is the climatological histogram as well as the fit to the log-normal distribution. In (e) and (f) are the posterior distributions for the Bayesian technique and BMA, respectively. The black diagonal line in some panels is the one-to-one line. The red vertical (horizontal) line in panels (e) and (f) are the 01 Feb GFS forecast (truth).

Figure 7. Results for the post-processing of days 1 to 5 Feb 2012. In (a) is shown the mean squared error (MSE) with respect to the NOGAPS analysis as a function of latitude. Solid blue (red) is the MSE of the Bayesian technique (BMA). The green solid line is the MSE of the climatological mean. Dashed lines correspond with the same color choice for the Bayesian technique, BMA, and climatology except are the corresponding variances. The grey bar along the bottom of panel (a) denotes those latitudes for which the difference between the Bayesian technique and BMA is statistically significantly different at 99% using a *t*-test. In (b) is the mean squared difference between the Bayesian technique (BMA) in blue (red). Solid is for the posterior mean difference against the NOGAPS forecast, long dashed is the posterior mean difference against the GFS forecast, and short dashed is the posterior mean difference against the climatological mean. In (c) is the binned spread plot for the Bayesian technique (blue) and BMA(red) using 10 equally populated bins. The linear fit through all the data is also plotted as the red (BMA) and blue (Bayesian) lines.

References

Bellman, R. E., 2003: *Dynamic Programming*. Courier Dover Publications. ISBN 978-0-486-42809-3.

Bishop, C. H., and K. T. Shanley, 2008: Bayesian Model Averaging's problematic treatment of extreme weather and a paradigm shift that fixes it. *Mon. Wea. Rev.*, **136**, 4641–4652. doi: http://dx.doi.org/10.1175/2008MWR2565.1

Delle Monache, L., F. A. Eckel, D. L. Rife, B. Nagarajan, and K. Searight, 2013: Probabilistic weather prediction with an analog ensemble. *Mon. Wea. Rev*., **141**, 3498-3516.

Dempster, A., Laird, N., and Rubin, D., 1977: Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Stat. Soc. Series B: Statistical Methodology* **39**, 1-38.

Doucet, A., S. Godsill, and C. Andrieu, 2000: On sequential Monte Carlo sampling methods for Bayesian filtering. *Statistics and Computing*, **10**, 197-208.

Fraley, C., A. E. Raftery, and T. Gneiting, 2010: Calibrating multimodel forecast ensembles with exchangeable and missing members using Bayesian model averaging. *Mon. Wea. Rev.*, **138**, 190-202.

Fortin, V., A.-C. Favre, and M. Saïd, 2006: Probabilistic forecasting from ensemble prediction systems: Improving upon the best-member method by using a different weight and dressing kernel for each member. *Quart. J. Roy. Meteor. Soc.*, **132,** 1349–1369.

Glahn, B., M. Peroutka, J. Wiedenfeld, J. Wagner, G. Zylstra, B. Schulknecht, and B. Jackson, 2009: MOS uncertainty estimates in an ensemble framework. *Mon. Wea. Rev.*, **137**, 246-268.

Gneiting, T., A. E. Raftery, A. H. Westveld, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, 133, 1098-1118.

Gneiting, T., F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration, and sharpness. *J. Royal Stat. Soc. Series B: Statistical Methodology*, **69**, 243-268.

Golub, G. H., and C. F. Van Loan, 1989: *Matrix Computations (Second Edition)*. Johns Hopkins Press, 642 pp.

Hamill, T.M., 2007: Comments on "Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging." *Mon. Wea. Rev.*, 135, 4226-4236. Han, J. and H.-L. Pan, 2011: Revision of Convection and Vertical Diffusion Schemes in the NCEP Global Forecast System. *Wea. Forecasting*, **26**, 520–533

Hamill, T. M., and J. S.Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: theory and application. *Mon. Wea. Rev*., **134**, 3209–3229.

Hamill, T. M., M. Scheuerer, and G. T. Bates, 2015: Analog probabilistic precipitation forecasts using GEFS reforecasts and Climatology-Calibrated Precipitation Analyses. *Mon. Wea. Rev.*, conditionally accepted.

Houtekamer, P. L., 1993: Global and local skill forecasts. *Mon. Wea. Rev*., **121**, 1834-1846.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the national digital forecast database. *Wea. Forecasting*, **23,** 270–289.

Luo, L., E. F. Wood, and M. Pan (2007), Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.*, **112**, D10102, doi:10.1029/2006JD007655

Raftery, A.E., T. Gneiting, F. Balabdaoui, M. Polakowski, 2005: Using Bayesian Model Averaging to calibrate forecasts. *Mon. Wea. Rev.*, 133, 1155-1174.

Rajagopalan, B., U. Lall, and S. E. Zebiak, 2002: Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles*. *Mon. Wea. Rev.*, **130**, 1792–1811.

Rosmond, T. E., 1992: The design and testing of the Navy Operational Global Atmospheric Prediction System. *Wea. Forecasting*, **7,** 262–272

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles." *Tellus A*, **55**, 16-30.

Satterfield, E.A. and C. H. Bishop, 2014: Heteroscedastic ensemble postprocessing. *Mon. Wea. Rev.*, **142**, 3484–3502. doi: http://dx.doi.org/10.1175/MWR-D-13-00286.1

Sloughter, J. M., A. E. Raftery, T. Gneiting, and C. Fraley, 2007: Probabilistic quantitative precipitation forecasting using Bayesian model averaging. *Mon. Wea. Rev.*, **135**, 3209–3220.**55A**, 16–30.

Sloughter, J. McLean, Tilmann Gneiting, and Adrian E. Raftery., 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J.Amer. Stat. Assoc.*, **105**.489, 25-35.

Tolman, H. L., 1997: User manual and system documentation of WAVEWATCH-III version 1.15. NOAA / NWS / NCEP / OMB Technical Note **151**, 97 pp.

Tolman, H. L., 1999: User manual and system documentation of WAVEWATCH-III version 1.18. NOAA / NWS / NCEP / OMB Technical Note **166**, 110 pp.

Tolman, H. L., 2001: Improving propagation in ocean wave models. *Ocean Wave Measurement and Analysis*, San Francisco, CA, B. L. Edge and J. M. Hemsley, Eds., ASCE, 507-516.

Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev*., **137**, 2365-2379.

Vrugt, J., C. J. H. Diks, and M. P. Clark, 2008: Ensemble Bayesian Model Averaging using Markov-chain Monte Carlo sampling. *Environmental Fluid Mechanics*. **8**, 579-595.

Wang, X., C. H. Bishop, 2003: A comparison of breeding and ensemble transform Kalman filter ensemble forecast schemes. *J. Atmos. Sci.*, **60**, 1140–1158. doi: http://dx.doi.org/10.1175/1520-0469(2003)060<1140:ACOBAE>2.0.CO;2

Wilks, D. S., 2006: Comparison of ensemble-MOS methods in the Lorenz '96 setting. *Meteor. Appl*., **13**, 243-256.

Wilson, L.J., S. Beauregard, A.E. Raftery, and R. Verret, 2007: Calibrated surface temperature forecasts from the Canadian ensemble prediction system using Bayesian Model Averaging. *Mon. Wea. Rev*., **135**, 1364-1385.

Yeo, I.-K., and R.A. Johnson, 2000: A new family of power transformations to improve normality or symmetry. *Biometrika*, **87**, 954-959.

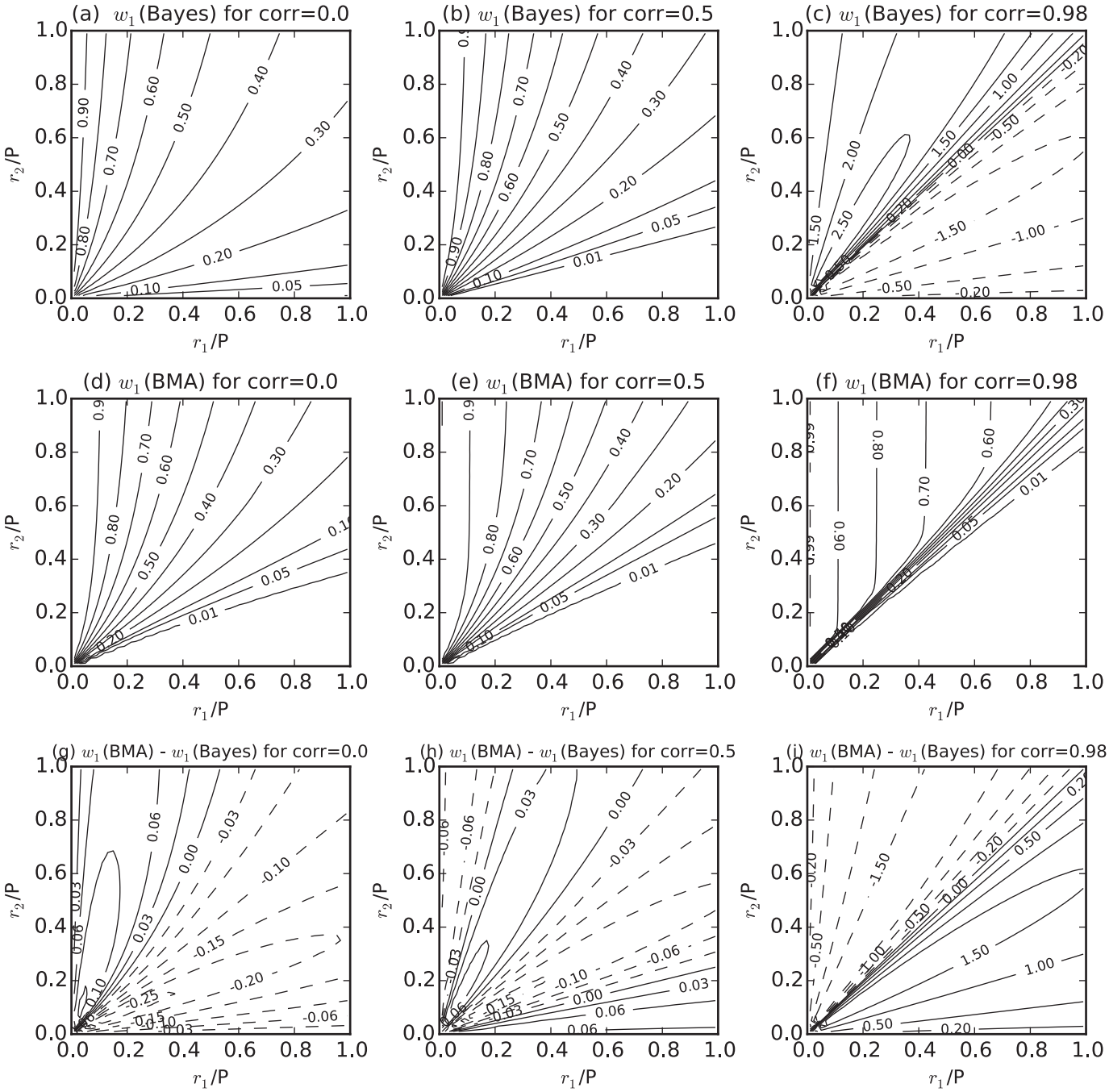

Figure 1: Weights for the first of two members in a forecast from Bayes (top row) and BMA (second row). Weight differences are shown in the third row. Weights and differences are plotted as functions of the ratio of the first forecast's error variance divided by the climatological variance (x), the second forecast's error variance divided by the climatological variance (y), and the correlation between the two forecasts (the three columns).

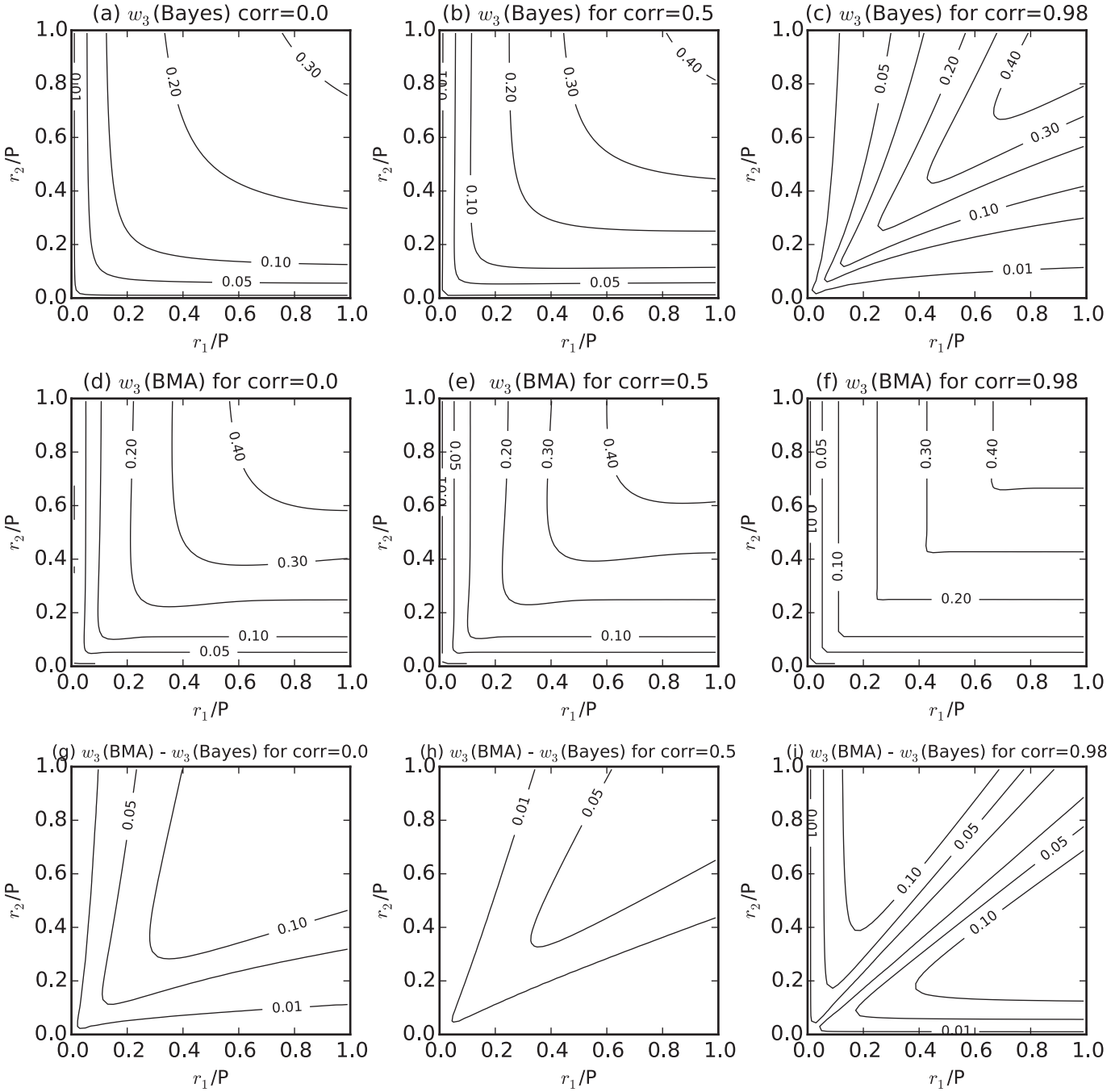

Figure 2: As in Fig. 1, but here for the weight applied to the climatology. Note that the weight of the second forecast, not shown, is in each case one minus the weight of the first forecast and climatology.

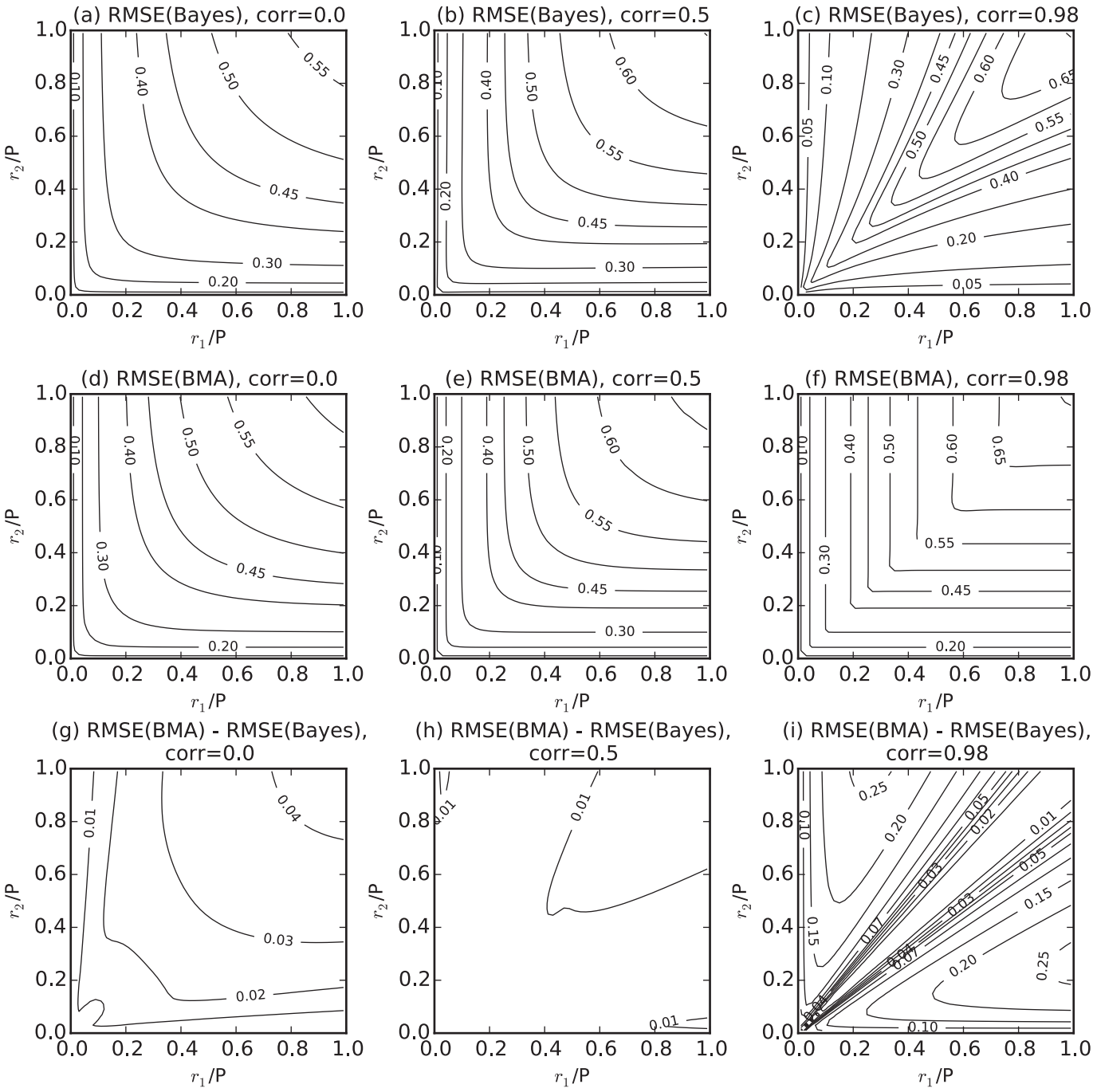

Figure 3: As in Fig. 1, but for RMS errors of the mean for Bayesian and BMA procedures as generated from 1,000 Monte-Carlo experiments, as described in the text.

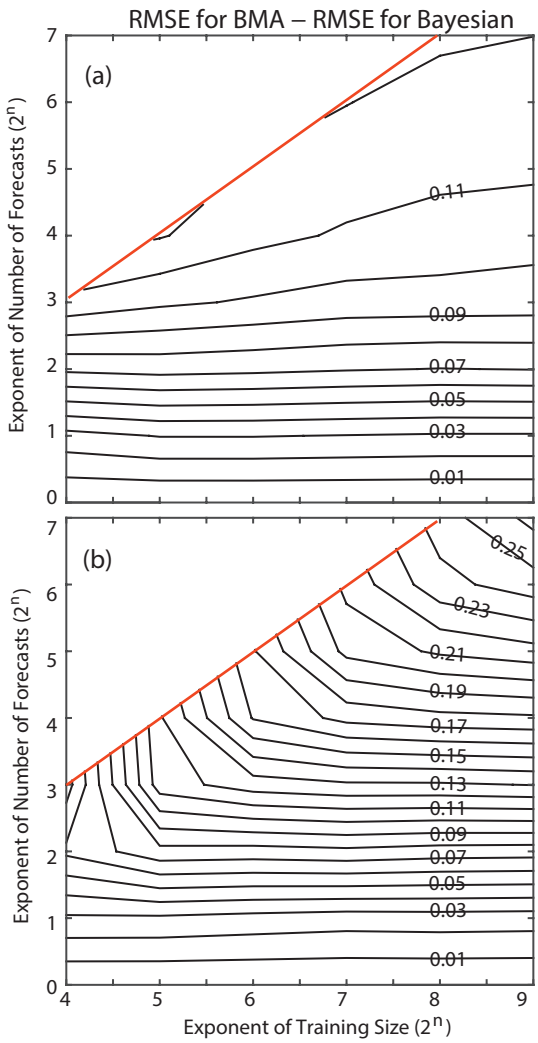

Figure 4. In (a) and (b) is plotted the RMSE difference between BMA and the method presented in section 4. In (a) is the experiment with unequally likely forecasts and in (b) is the experiment with equally likely forecasts. The diagonal red line in the upper left corner of each plot denotes the region above which the training size is equal to or smaller than the number of predictors, and for which experiments were not performed.

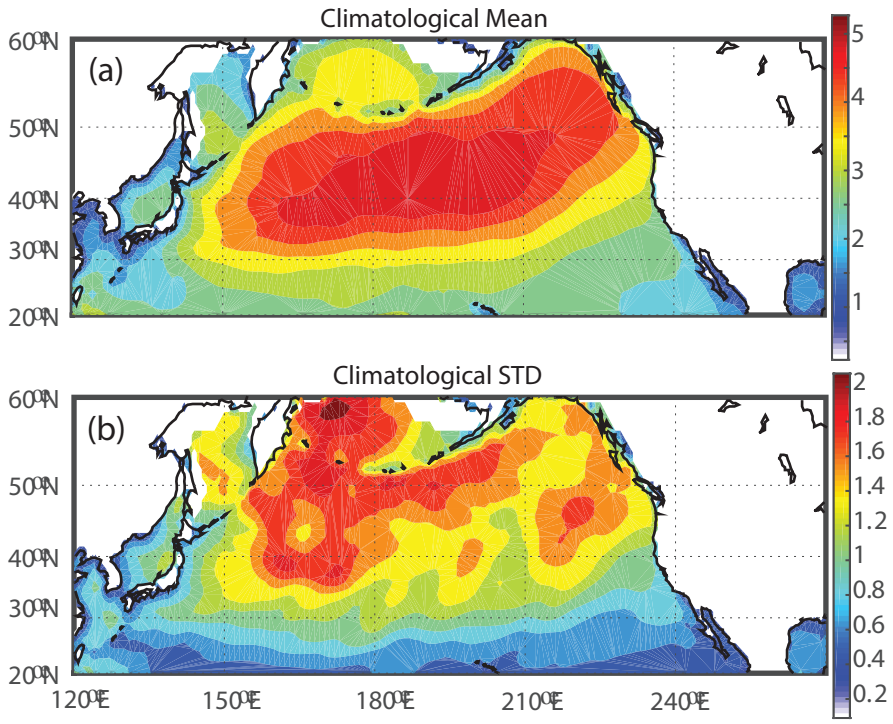

Figure 5. The (a) climatological mean and (b) standard deviation across the training set at each grid point for the period 01 December 2011 to 31 January 2012.

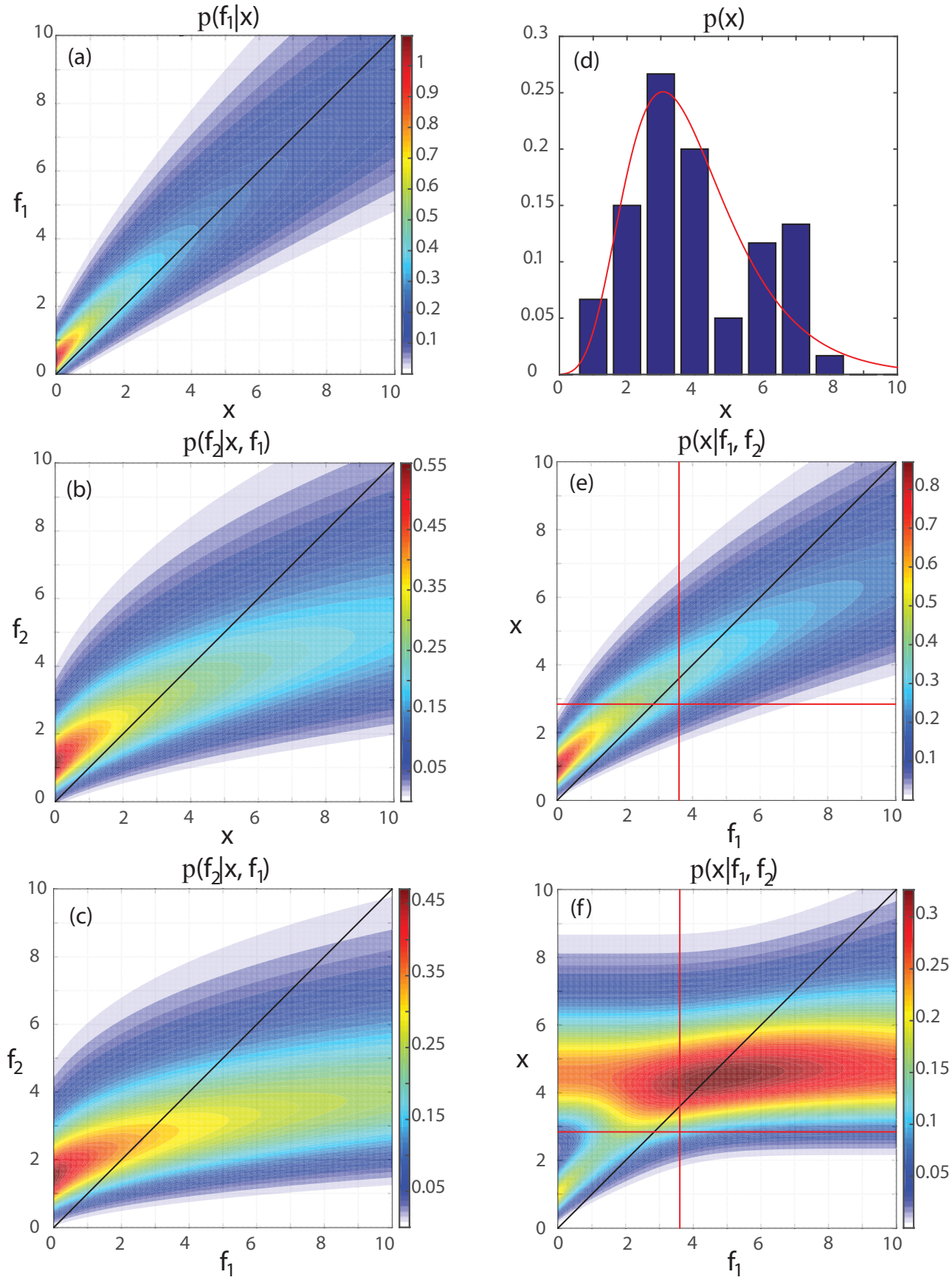

Figure 6. The fitting of the likelihoods and climatology for a grid point in the North Pacific (50 N, 180 E) and evaluated for 01 Feb. In (a) is the forecast likelhood for the GFS model given the truth. In (b) and (c) are the forecast likelihood for the NOGAPS model given the truth and the GFS forecast. In (d) is the climatological histogram as well as the fit to the log-normal distribution. In (e) and (f) are the posterior distributions for the Bayesian terchnqiue and BMA, respectively. The black diagonal line in some panels is the one-to-one line. The red vertical (horizontal) line in panels (e) and (f) are the 01 Feb GFS forecast (truth).

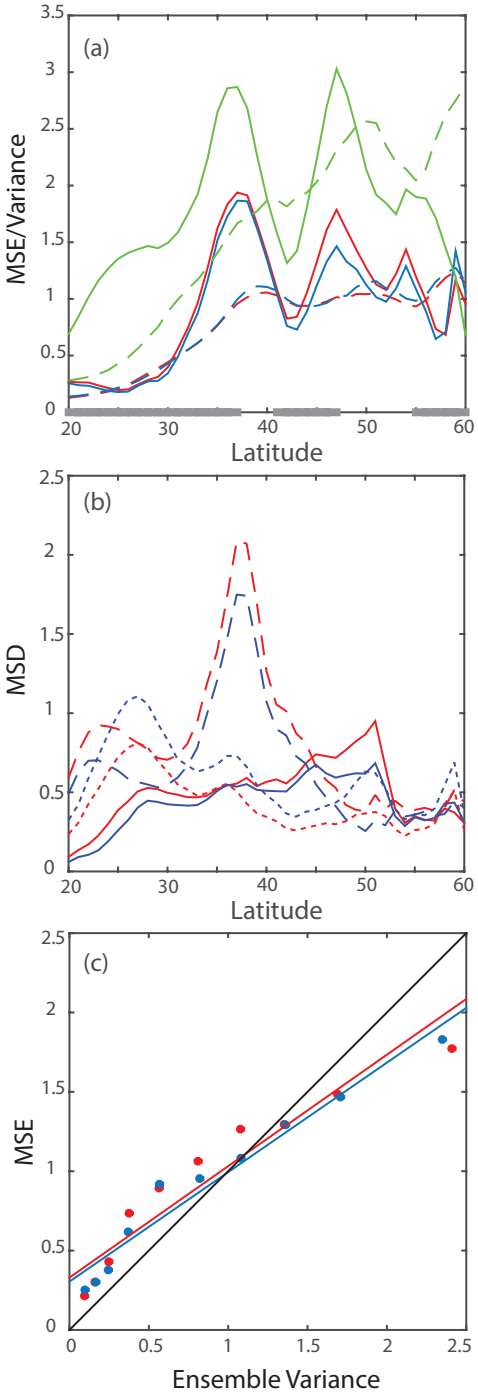

Figure 7. Results for the post-processing of days 01 Feb to 05 Feb. In (a) is shown the mean squared error (MSE) with respect to the NOGAPS analysis as a function of latitude. Solid blue (red) is the MSE of the Bayesian techqnique (BMA). The green solid line is the MSE of the climo mean. Dashed lines correspond with the same color choice for the Bayesian technique, BMA, and climo except are the corresponding variances. The grey bar along the bottom of panel (a) denotes those latitudes for which the difference between the Bayesian technique and BMA is statistically significantly different at 99%. In (b) is the mean squared difference between the Bayesian technique (BMA) in blue (red). Solid is for the Posterior mean difference with respect to the NOGAPS forecast, long dashed is the GFS forecast , and short dashed is against the climatological mean. In (c) is the binned spread plot for the Bayesian technqiue (blue) and BMA(red) using 10 equally populated bins. The linear fit through all the data is also plotted as the red (BMA) and blue (Bayesian) lines.