Heteroscedastic Ensemble Postprocessing

ELIZABETH A. SATTERFIELD AND CRAIG H. BISHOP

Naval Research Laboratory, Monterey, California

(Manuscript received 12 September 2013, in final form 9 April 2014)

ABSTRACT

Ensemble variances provide a prediction of the flow-dependent error variance of the ensemble mean or, possibly, a high-resolution forecast. However, small ensemble size, unaccounted for model error, and imperfections in ensemble generation schemes cause the predictions of error variance to be imperfect. In previous work, the authors developed an analytic approximation to the posterior distribution of true error variances, given an imperfect ensemble prediction, based on parameters recovered from long archives of innovation and ensemble variance pairs. This paper shows how heteroscedastic postprocessing enables climatological information to be blended with ensemble forecast information when information about the distribution of true error variances given an ensemble sample variance is available. A hierarchy of postprocessing methods are described, each graded on the amount of information about the posterior distribution of error variances used in the postprocessing. These homoscedastic methods are used to assess the value of knowledge of the mean and variance of the posterior distribution of error variances to ensemble postprocessing and explore sensitivity to various parameter regimes. Testing was performed using both synthetic data and operational ensemble forecasts of a Gaussian-distributed variable, to provide a proof-of-concept demonstration in a semi-idealized framework. Rank frequency histograms, weather roulette, continuous ranked probability score, and spread-skill diagrams are used to quantify the value of information about the posterior distribution of error variances. It is found that ensemble postprocessing schemes that utilize the full distribution of error variances given the ensemble sample variance outperform those that do not.

1. Introduction

Ensemble perturbations are designed to sample the distribution of analysis and forecast errors. Initial ensemble perturbations that are designed to represent the initial condition error distribution are added to the best available analysis to create the ensemble of initial conditions from which the ensemble forecast is made using one or more nonlinear (possibly stochastic) models (Toth and Kalnay 1997; Toth et al. 2001; Palmer et al. 1998; Houtekamer et al. 1996). Hence, the variance of the ensemble provides a prediction of the error variance of the ensemble mean or, possibly, the high-resolution forecast that is often made in conjunction with a coarser-resolution ensemble forecast. In particular, ensemble forecast variances are designed to predict flow-dependent and observational network dependent aspects of the error variance. Ensemble-based probabilistic forecasts of events are typically based on the relative frequencies of events in the ensemble and have been shown to provide information that can increase the socioeconomic value of weather and climate forecasts by, among others, Richardson (2000), Zhu et al. (2002), Palmer (2002), Mylne (2002), and Cai et al. (2011).

Ensemble forecasts are imperfect: their initial conditions imperfectly sample the distribution of true states given observations, the number of members is typically too few to accurately represent the second and third moments of the distribution of errors, and they imperfectly represent the effects of poorly understood and poorly quantified systematic and stochastic sources of model error. Postprocessing techniques have proven useful in correcting some of the deficiencies of the ensemble. The goal of postprocessing is to describe the distribution of true states given an ensemble of forecasts (including extreme but rare events).

A variety of approaches have been proposed for statistical postprocessing of ensemble forecasts. Nonhomogeneous Gaussian regression (NGR; Gneiting et al. 2005) assumes a Gaussian distribution and performs bias correction of the mean and a variance correction

Corresponding author address: Elizabeth Satterfield, Naval Research Laboratory, Marine Meteorology Division, 7 Grace Hopper Ave., Stop 2, Monterey, CA 93943-5502.
E-mail: elizabeth.satterfield@nrlmry.navy.mil

DOI: 10.1175/MWR-D-13-00286.1

by maximizing a log-likelihood function to obtain coefficients. Dressing methods (Roulston and Smith 2003; Wang and Bishop 2005; Fortin et al. 2006) and Bayesian model averaging (BMA; Raftery et al. 2005) address ensemble dispersion by generating a kernel distribution around each raw ensemble member. The raw ensemble members are typically bias corrected based on historical data and can be weighted, as in BMA, if 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 BMA, the weights and variances are chosen by maximum likelihood from the training data. While BMA was originally formulated to combine forecasts from different models, this method is closely aligned with the ''best-member dressing'' method of Roulston and Smith (2003) and the approach of Wang and Bishop (2005), which ensures that mean ensemble variance agrees with the forecast error variance. These dressing formulations increase the variance of the ensemble and hence are appropriate for underdispersive ensembles but inappropriate for overdispersive ensembles. For an ensemble prediction system with exchangeable members, Wang and Bishop (2005), Roulston and Smith (2003), and BMA only differ in implementation details (Fortin et al. 2006). Fortin et al. (2006) formulated a dressing method that is appropriate for overdispersive ensembles. Krzysztofowicz and Evans (2008) introduced the Bayesian processor of forecasts (BPF) in which climatological information is incorporated via Bayes's rule after distributions have been transformed to be normally distributed. Further, Bishop and Shanley (2008) argue that prior climatological information should be incorporated into BMA ensemble postprocessing through Bayes's theorem.

The method presented in the present study includes climatological information following Bayes's rule as in Krzysztofowicz and Evans (2008). The inclusion of prior climatological information acts to push the deterministic forecast toward the climatological mean in the same way that a regression-based correction would. However, the method presented here features a climatological weighting that changes with day-to-day changes in forecast error variance. In this way, this method accounts directly for the fact that how close we draw to climatology is itself uncertain, and that uncertainty is flow dependent.

Eckel et al. (2012) discussed the uncertainty associated with ensemble variance prediction and described ambiguity as the uncertainty in the prediction of forecast uncertainty (or in the forecast probability of a specific event) arising from finite sampling and deficiencies in simulation of various sources of forecast uncertainty. Eckel et al. (2012) quantified this type of uncertainty using *total ambiguity*, the 90% confidence interval of calibrated ensemble-based forecast probabilities. Eckel et al. (2012) constructed distributions of ensemble-based forecast probabilities through bootstrapping, ensemble of ensembles, and calibrated error sampling. The posterior distribution of error variances in Bishop and Satterfield (2013) and Bishop et al. (2013) show how one can quantify this ambiguity using long archives of innovation, ensemble variance pairs.

In Bishop and Satterfield (2013) an analytic approximation to the posterior distribution of error variances, given an imperfect ensemble prediction and prior climatological information, was introduced and tested using a simple dynamical model. In Bishop et al. (2013), equations were developed that enable all parameters defining the prior, likelihood, and posterior distributions of Bishop and Satterfield's analytical model to be deduced from a long archive of innovation and ensemble variance pairs. These distributions quantify the nonhomogeneous nature of the true error variances or error *heteroscedasticity* given an ensemble variance. In what follows, we develop a method for using error heteroscedasticty information in ensemble postprocessing.

The outline of this paper is as follows: to assess the value of the postprocessing and discover the sensitivity of this value to changes in parameter regimes, in section 2 we apply a hierarchy of ensemble postprocessing techniques to synthetic data. The hierarchy of ensemble postprocessing techniques is ordered according to how much information they incorporate from the known distribution of error variances given an imperfect ensemble variance prediction. In section 3, we apply the same postprocessing techniques and diagnostics to operational ensembles produced by the Fleet Numerical Meteorology and Oceanography Ensemble Forecast System (FNMOC EFS). Results are compared to homoscedastic ensemble postprocessing techniques as well as to Raftery el al.'s (2005) BMA method in section 4. Conclusions follow in section 5.

2. Utilization of error variance distributions in ensemble postprocessing

In this section, a new heteroscedastic postprocessing algorithm is described that incorporates information about (i) the climatological prior distribution of error variances and (ii) the distribution of ensemble sample variances given an error variance. To isolate the value of accounting for heteroscedasticity in postprocessing, we also develop a hierarchy of homoscedastic ensembles that vary in the amount of information used in the ensemble postprocessing.

The hierarchy of postprocessing schemes is then tested using synthetic data consistent with the analytic model described in detail in Bishop et al. (2013). The use of synthetic data serves two purposes. First, it allows us to test our postprocessing algorithm in a framework in which all the assumptions of our theory are satisfied. Second, using synthetic data allows us to easily modify parameter values and document how the benefits of postprocessing depend on underlying parameter values.

a. Distributions of error variance given an ensemble variance

In this section, we briefly summarize the analytical model of imperfect error variance prediction introduced in Bishop et al. (2013). We define a hypothetical "true state" as a random draw from a climatological Gaussian distribution with a specified mean $\langle x^t \rangle_C$ and variance σ_C^2:

$$x_i^t \sim N(\langle x^t \rangle_C, \sigma_C^2). \tag{1}$$

The forecast error is assumed to be a random draw from a Gaussian distribution with variance σ_i^2. To capture the fact that error variances vary from one forecast to the next depending on changes to the flow and observational network, we let the σ_i^2 associated with each forecast be a random draw from an inverse Gamma distribution $\Gamma^{-1}(\alpha_{\text{prior}}, \beta_{\text{prior}})$ of error variances. In terms of mathematical expressions

$$x_i^f = x_i^t + \varepsilon_i^f, \quad \text{where} \quad \varepsilon_i^f \sim N(0, \sigma_i^2) \quad \text{and}$$
$$\sigma_i^2 \sim \Gamma^{-1}(\alpha_{\text{prior}}, \beta_{\text{prior}}). \tag{2}$$

Note that (2) implicitly assumes that forecast error is independent of the actual value of the true state. In practice, a preprocessing step could be added to remove any bias, flow dependent or otherwise. We further assume that the ensemble variance s_i^2 is a random draw from a Gamma distribution with mean $a(\sigma_i^2 - \sigma_{\text{min}}^2) + s_{\text{min}}^2$ such that

$$s_i^2 - s_{\text{min}}^2 \sim \Gamma(k, \theta), \quad k\theta = a(\sigma_i^2 - \sigma_{\text{min}}^2), \tag{3}$$

where a represents a sensitivity parameter defining the mean response of ensemble variance to changes in the error variance and s_{min}^2 is a minimum (possibly zero) value of ensemble variance, due to the technique used to generate the ensemble. Here σ_{min}^2 represents the minimum possible value of error variance. The Gamma distribution parameter k defines the relative variance of the ensemble variances given an error variance, in other words,

$$k^{-1} = \frac{\text{var}(s^2 \mid \sigma^2)}{\langle s^2 \mid \sigma^2 \rangle} = \frac{\text{var}(s^2 \mid \sigma^2)}{a(\sigma^2 - \sigma_{\text{min}}^2)}. \tag{4}$$

As noted in Bishop and Satterfield (2013), since the relative variance of the sample variance of an M member random draw from a Gaussian, normal distribution is given by $k^{-1} = 2/(M-1)$, one can associate an *effective ensemble size M* with any recovered value of k. The parameters M *and* k^{-1} provide a measure of the extent to which ensemble variance tracks error variance that is entirely independent of the accuracy of the model and/or the degree of under- or overdispersion of the ensemble variance. As far as the authors are aware, no other measures of ensemble performance have this property.

Having assumed distributions for the prior climatological distribution of forecast error variances and the distribution of ensemble variances given a true error variance, Bayes's theorem gives the posterior distribution of error variances given s^2. To be specific,

$$\rho_{\text{post}}(\sigma^2 \mid s^2) = \frac{L(s^2 \mid \sigma^2)\rho_{\text{prior}}(\sigma^2)}{\int_0^\infty L(s^2 \mid \sigma^2)\rho_{\text{prior}}(\sigma^2)\, d(\sigma^2)}, \tag{5}$$

where L denotes the pdf of the likelihood of s^2 given a particular σ^2. Bishop and Satterfield (2013) showed that (5) is itself an inverse gamma distribution with posterior parameters $\alpha_{\text{post}} = \alpha_{\text{prior}} + k$ and $\beta_{\text{post}} = \{(s_i^2 - s_{\text{min}}^2)k/a\} + \beta_{\text{prior}}$. Hence, given x_i^f and an inaccurate ensemble variance prediction s_i^2 there is an inverse-gamma posterior distribution of possible error variances. We will denote the jth random draw of this posterior distribution of error variances given an ensemble variance by $(\sigma_s^2)_j$.

Bishop and Satterfield (2013) also showed that the mean of the posterior distribution of variances $\langle \sigma^2 \mid s^2 \rangle$ over all realizations of the error variance given a fixed value of s^2 is given by

$$\begin{aligned}\langle \sigma^2 \mid s^2 \rangle &= \frac{k}{k + (\alpha_{\text{prior}} - 1)} \left(\frac{s_i^2 - s_{\text{min}}^2 + a\sigma_{\text{min}}^2}{a} \right) \\ &\quad + \frac{\alpha - 1}{k + (\alpha_{\text{prior}} - 1)} \langle \sigma^2 \rangle \\ &= w_e \sigma_n^2 + w_c \langle \sigma^2 \rangle, \end{aligned} \tag{6}$$

where

$$\sigma_n^2 = \frac{s_i^2 - s_{\text{min}}^2 + a\sigma_{\text{min}}^2}{a} \tag{7}$$

denotes a "corrected" ensemble prediction of error variance, w_e denotes the weight given to this error variance prediction, and $w_c = 1 - w_e$ denotes the weight given to the prior climatological mean error variance.

The expression in (7) inflates or attenuates the raw ensemble variances according to the sensitivity parameter a and adjusts for minimum values of ensemble variances or error variances. In this way, σ_n^2 represents a debiased ensemble prediction.

b. Heteroscedastic ensemble postprocessing

Among other things, all good postprocessing techniques ensure that information from the climatological distribution of the true state is incorporated into the postprocessed forecast. The degree to which this climatological information adds value to a forecast depends on the accuracy of the forecast. However, ensemble forecasts of the flow-dependent error distribution of the forecast are, in general, inaccurate. *This means that the degree to which climatological information should be incorporated in the probabilistic forecast is itself uncertain.* To account for this uncertainty, we proceed in the following manner. We first express the uncertainty in the true forecast error variance by creating a large (1000 member) ensemble of error variances $(\sigma_S^2)_j$, $j = 1, 2, \ldots, 1000$ by randomly sampling the posterior inverse-gamma distribution of possible error variances given a specific ensemble variance s_i^2 [see our (5); see Bishop and Satterfield (2013) and/or the appendix for equations to compute the parameters defining this posterior distribution]. The well-known formula (Daley 1991, his section 2.2) for optimally blending an estimate x^f (the forecast) with error variance $(\sigma_S^2)_j$ with a climatological estimate $\langle x^t \rangle_C$ (the climatological mean) with error variance σ_C^2 is

$$x_j^{\text{post}} = \left[\left\{ \frac{\dfrac{1}{(\sigma_S^2)_j}}{\dfrac{1}{(\sigma_S^2)_j} + \dfrac{1}{\sigma_C^2}} \right\} x^f + \left\{ \frac{\dfrac{1}{\sigma_C^2}}{\dfrac{1}{(\sigma_S^2)_j} + \dfrac{1}{\sigma_C^2}} \right\} \langle x^t \rangle_C \right] + \zeta_j, \quad j = 1, 2, \ldots, 1000, \tag{8}$$

where ζ_j is a normal random variable with mean zero and variance $(\sigma_p^2)_j$, where $1/(\sigma_p^2)_j = [1/(\sigma_S^2)_j] + (1/\sigma_C^2)$. Note that the variance $(\sigma_p^2)_j$ is the error variance of the state estimate obtained by combining a forecast with error variance $(\sigma_S^2)_j$ with a climatological mean state. Sampling $(\sigma_S^2)_j$ from (5) accounts for uncertainty in the flow-dependent true error variance. Adding the perturbation ζ_j allows us to randomly sample the posterior distribution of states that occur on those occasions when the true error variance is equal to the sampled $(\sigma_S^2)_j$ value. In other words, the above procedure explicitly accounts for and averages over the distribution of possible posterior distributions corresponding to the distribution of possible true error variances given the ensemble variance.

Using (8) to create a 1000-member ensemble ensures that this ensemble optimally combines climatological and forecast information *given that the error variance of the forecast is uncertain.* The climatological mean of the true state, $\langle x^t \rangle_C$ is determined from a long time series of unbiased observations. To remove the signal of observational errors from the variance of the observations, the climatological variance of the true state is determined by subtracting the observation error variance R from the variance σ_o^2 of the time series of observations (i.e., $\sigma_C^2 = \sigma_o^2 - R$). For the case of synthetically generated data, the values for the climatological mean and variance are prescribed: $\sigma_C^2 = 12.25$ and $\langle x^t \rangle_C = 2.5$.

Equation (8) represents a *heteroscedastic* postprocessing method because it assumes a range of possible true error variances given an ensemble variance. Equation (8) shows that the weight given to the climate mean changes for each of the $j = 1, 2, \ldots, 1000$ ensemble members because of its dependence on the value of $(\sigma_S^2)_j$, which is drawn from an inverse-gamma posterior distribution of error variances. Note that, because of this changing mean state and the fact that the value of $(\sigma_S^2)_j$ is drawn from a distribution with a long tail, the distribution of postprocessed ensemble members is not necessarily symmetric. We hereafter refer to the ensemble obtained using (8) and $(\sigma_S^2)_j$ from the inverse-gamma posterior distribution of error variances as *the fully postprocessed ensemble or FP ensemble.*

c. "Non-FP" homoscedastic ensembles

To isolate the value of heteroscedastic postprocessing in a simplified context, we now generate partially postprocessed non-FP homoscedastic ensembles using three different methods, designed to gradually increase the amount of information the ensemble has about the posterior distribution of error variances. In the first method, we completely ignore the ensemble variance s_i^2 and set $(\sigma_S^2)_j = \langle \sigma_{\text{prior}}^2 \rangle$, where $\langle \sigma_{\text{prior}}^2 \rangle$ is the mean of the climatological distribution of error variances, for all j in (8) to obtain 1000 postprocessed ensemble members. The weighting of climatological information in this ensemble is only informed by the climatological mean of error variances and ignores the ensemble variance. We refer to this first ensemble as the *invariant ensemble* because its variance is the same for every single forecast. This type of ensemble postprocessing could be done without the analysis tools given in Bishop et al. (2013) and is in no sense a new method. Note that via (8), it incorporates an aspect of the BPF approach.

In the second non-FP experiment, we set the posterior error variance equal to the corrected error variance

prediction σ_n^2 by using $(\sigma_S^2)_j = \sigma_n^2$ for all j in (8). The parameter a represents a sensitivity parameter defining the mean response of ensemble variance to changes in the error variance. Among other things, the inverse of a in the expression for σ_n^2 inflates or attenuates σ_n^2 relative to the raw ensemble variance. The flow-dependent error variance prediction given by σ_n^2 is free of systematic bias in that for a given true value of error variance σ^2, its mean error $\langle \sigma_n^2 - \sigma^2 \,|\, \sigma^2 \rangle$ is equal to zero. The parameters $s_{\min}^2$ and $\sigma_{\min}^2$ describe a minimum (possibly zero) value of the ensemble variance and error variance respectively. We note that if the parameters $s_{\min}^2$ and $\sigma_{\min}^2$ were equal to zero, the corrected error variance prediction σ_n^2 would simply remove the mean over- or underdispersion of the raw ensemble. In Eckel et al.'s (2012) ground-breaking work on ambiguity in ensemble forecasting, he utilized a ''shift and stretch'' calibration, a simplified version of the method described by Johnson and Bowler (2009). This method entailed a first and second moment correction, which could be applied as either a bulk or conditional correction. The bulk correction involves removing a mean bias and inflating perturbations by the inverse of the fractional error (square root of the mean ensemble variance divided by the mean squared error). For cases in which $s_{\min}^2 = \sigma_{\min}^2 = 0$ a similar correction is achieved by the parameter a. In this paper, we have formalized this correction in order to account for the contribution of observation error variance and to allow for nonzero minimum values of ensemble variance and forecast error variance. We refer to the ensemble obtained from this method as the modified shift and stretch (MSS) ensemble because, as in Eckel et al. (2012) it includes a shift toward climatology through (8) and a stretch through the parameter a. Unlike Eckel et al. (2012), the MSS method is like BPF in that climatological information is incorporated via (8).

For the third non-FP ensemble, we set $(\sigma_s^2)_j = \langle \sigma^2 \,|\, s^2 \rangle$ for all j in (8) thus giving the ensemble variance the correct posterior mean error variance for each particular forecast, while ignoring any possible deviation of error variance about the mean error variance. This ensemble ignores the variable nature of the possible error variances given the ensemble prediction and prior climatological information. We shall refer to this ensemble as the *informed-Gaussian* ensemble because while it falsely assumes that the posterior distribution of truth is Gaussian, it has the correct posterior mean error variance. One can view this ensemble as a variant of NGR. NGR uses a maximum likelihood estimator to determine coefficients a, b, c, and d and assumes that a calibrated forecast is drawn from a normal distribution with mean $a + b\overline{x}$ and variance $c + ds_i^2$. Since we recenter our forecasts on the higher-resolution deterministic and found the bias to be small we can assume that $a = 0$ and $b = 1$. Equation (6) shows that the posterior mean error variance can be written as $w_e\sigma_n^2 + w_c\langle\sigma^2\rangle$. Hence, by setting $c = w_c\langle\sigma^2\rangle$ and $ds_i^2 = w_e\sigma_n^2$, one can see the similarity between the ''informed Gaussian'' method and the NGR method. Unlike the NGR method, the informed-Gaussian method is like BPF in that it incorporates climatological information via (8).

The FP method arose directly from theoretical considerations outlined in Bishop and Shanley (2008), Bishop and Satterfield (2013), and Bishop et al. (2013). These considerations incorporated ideas from the ''shift and stretch,'' NGR, and BPF methods. In addition, they lent an entirely new aspect to ensemble postprocessing; namely, the random sampling of the distribution of hidden error variances *given an imperfect ensemble variance*. The non-FP ensemble postprocessing approaches (invariant, MSS, and informed Gaussian) are *homoscedastic* in that they all assume that there is just a single forecast error variance associated with each forecast x^f. Only the *heteroscedastic* FP method accounts for the fact that there is a distribution of true error variances associated with each ensemble forecast.

d. *Rank frequency histograms for FP and non-FP ensembles*

To create rank frequency histograms (RFHs) (also known as ''Talagrand diagrams;'' Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997) one takes an M-member ensemble forecast and the observed verification state and orders the M forecasts and observed state from lowest to highest value. If the verification state is lower than all the forecasts, it is assigned rank 1, if it is lower than all but one of the ensemble members, it is said to have rank 2, and so on. The RFH plots the frequency with which the verifying state takes each possible rank over a large number of forecasting trials. If the verification state is drawn from the same distribution as the ensemble members then it would take each rank with equal frequency. This is a sufficient but not necessary condition for a flat histogram [see Bishop and Shanley (2008) for an example of a flat RFH from an ensemble that is not drawn from the distribution of truth given the forecast]. Since our experiments have a known ''true'' state, we simply rank the true state among the ordered ensemble members.

Figure 1 compares the results of the RFHs for the four methods used to postprocess the ensemble. For these experiments, the prescribed parameters were $k = 0.5$ ($M = 2$), $R = 0.05$, $\mathrm{var}(\sigma^2) = 0.01$, $\langle\sigma^2\rangle = 0.036$, and $a = 0.75$. As discussed in Bishop and Satterfield (2013), one can associate an *effective ensemble size* M with

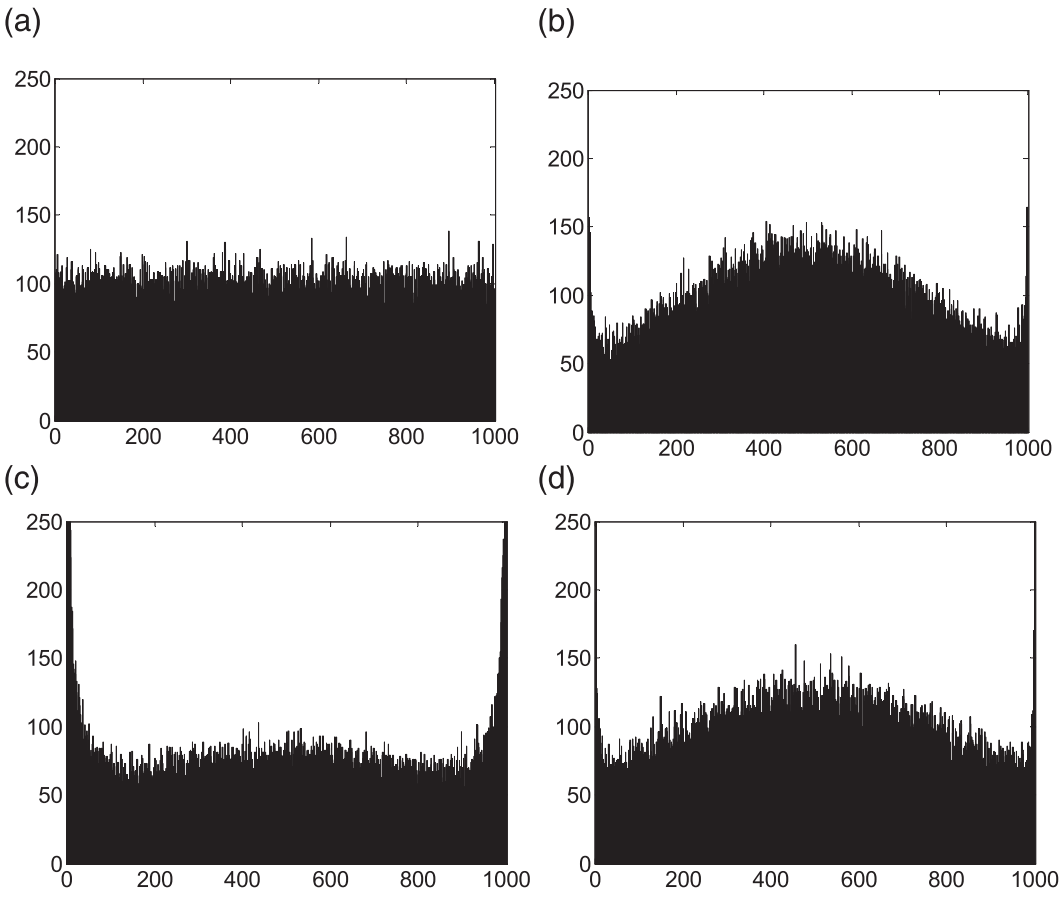

FIG. 1. Rank frequency histogram obtained from an $M = 1000$ member postprocessed ensemble. In this figure, the known true state is ranked among the ordered ensemble members. The true state is taken from a normal distribution with mean $\langle x^t \rangle_C = 2.5$ and standard deviation $\sigma_C = 3.5$. Results are shown for (a) FP ensemble, (b) invariant ensemble, (c) MSS ensemble, and (d) informed-Gaussian ensemble. This experiment has prescribed parameter values of $k = 0.5$, $\text{var}(\sigma^2) = 0.01$, and $a = 0.75$.

any recovered value of k. This value of k corresponds to an effective ensemble size of $M = 2$ members, where $M = 2k+1$. As previously mentioned, the true state was generated by taking a random draw from a Gaussian distribution with mean $\langle x^t \rangle_C = 2.5$ and standard deviation $\sigma_C = 3.5$ (variance $\sigma_C^2 = 12.25$); $M_p = 1000$ postprocessed ensembles were generated for 100 000 forecast events. If the postprocessed ensemble had exactly the same distribution of the true state, each of the 1001 possible ranks would occur 99.9 times for every 100 000 trials (on average). We find that the FP ensemble shows a near-optimal frequency for each bin—as one would expect if the postprocessed ensemble were sampling the distribution of truth given the forecast. The MSS ensemble gives a RFH that appears underdispersive, while the invariant ensemble and the informed-Gaussian ensemble both look overpopulated in the center bins, dip toward the tails, and are overpopulated in the outermost bins. The postprocessed ensemble variances averaged over 100 000 forecast events was found to be 0.0359, 0.0347, 0.0357, and 0.0357 for the invariant, MSS, informed-Gaussian, and fully postprocessed ensembles, respectively, very close to the forecast error variances of the postprocessed ensemble means of 0.0354, 0.0357, 0.0353, and 0.0353. This means that even though the invariant, informed-Gaussian, and fully postprocessed ensembles have the correct variance on average, only the heteroscedastic FP ensemble has a flat RFH because it is the only method that accounts for the heteroscedastic nature of true error variances given an ensemble variance.

e. *Weather roulette*

To translate the enhanced RFH flatness due to heteroscedastic ensemble postprocessing to a recognizable "value" to users of ensemble forecasts, we implement a version of Hagedorn and Smith's (2009) weather roulette diagnostic. Weather roulette is a hypothetical

gambling game in which two players, A and B, each open a weather roulette casino in which one can bet that a verifying observation (or for the case of synthetic data, a known true state) will fall into one of $i = 1, 2, \ldots, N_b$ climatological bins. Players A and B use their respective ensemble-based probabilistic forecasts for each bin to set the odds or payout ratio in their own casino, denoted $1/P_A(i)$ and $1/P_B(i)$, respectively.

To be clear about what we mean by "payout ratio", suppose that player B bets \$10 in player A's casino that the verifying observation will fall into the third climatological bin. If the verifying observation actually does fall into the third climatological bin, player B will receive $\$[10/P_A(i)]$. If the verifying observation does not fall in the third bin player B will receive \$0 and the \$10 bet will stay with player A.

We also assume that when the players play in the other's casino they distribute the respective fractions $P_A(i)$ and $P_B(i)$ of their own money in the $i = 1, 2, \ldots, N_b$ climatological bins of the other casino. In what follows, we consider $N_b = 100$ percentile based climatological bins. We want to ensure that this gambling game can be played over N forecast events (for the synthetic data $N = 100\,000$ forecasts), so we must avoid the possibility that a gambler loses all his/her winnings and can therefore no longer place bets. To achieve this property, we guarantee that the smallest probability assigned to a bin is greater than zero by defining the probability $P(i)$ of truth falling into the ith climatological bin, by

$$P(i) = \frac{b_i + 1}{M + N_b}, \tag{9}$$

where b_i is the number of members that fall in the ith climatological bin, N_b is the number of climatological bins ($N_b = 100$ in our experiments), and M is the total number of ensemble members ($M = 1000$ for these experiments). Under the aforementioned scenario, it can be shown that if player A is betting his money in player B's casino and i_n denotes the index of the climatological bin that the verifying observation falls into then the average interest rate earned per trial is given by

$$I = \left\{ \sqrt[N]{\prod_{n=1}^{N} \left[\frac{P_A(i_n)}{P_B(i_n)} \right]} \right\} - 1. \tag{10}$$

Thus, if in the geometric average $\sqrt[N]{\prod_{n=1}^{N}[P_A(i_n)/P_B(i_n)]}$, player A manages to set higher probabilities than the casino run by player B for the bins that that the verifying observation falls into then player A earns a positive return on her or his bets. Following Hagedorn and Smith, we shall hereafter refer to the "per trial" return given by (10) as the *effective daily interest rate* and report values in terms of percent.

f. Parameter regime dependence

To study the sensitivity of the value of heteroscedastic postprocessing to error regime, we first perform our tests using synthetic data in which the parameter values can be easily controlled and modified. The models and methods used to generate the synthetic data for these tests are described in detail in Bishop and Satterfield (2013) and Bishop et al. (2013). For our control experiment, we define the climatological distribution of innovation variances given by setting $\langle \sigma^2 \rangle = 0.036$, $\mathrm{var}(\sigma^2) = 0.01$, $\sigma^2_{\min} = 0$, $s^2_{\min} = 0$, and consider ensemble variance accuracies consistent with raw ensemble sizes of $M_e = 2$, $M_e = 4$, $M_e = 6$, $M_e = 8$, and $M_e = 10$. Bishop and Satterfield (2013) showed that these *effective ensemble sizes* correspond to gamma distribution parameters $k = 0.5$, $k = 1.5$, $k = 2.5$, $k = 3.5$, and $k = 4.5$, respectively. The inverse k^{-1} of k gives the relative error variance of the ensemble variance prediction; hence, larger values of k correspond to a more accurate error variance prediction. For each raw ensemble size, we let the gambler use the heteroscedastic FP ensemble to place bets in casinos with odds set by (9) using (i) invariant, (ii) MSS, and (iii) informed-Gaussian homoscedastic ensemble postprocessing techniques. Each casino creates odds for the same $N_b = 100$ equally likely climatological bins. Each ensemble has $M = 1000$ postprocessed ensemble members. The gambler places bets using the strategy outlined above for $N = 100\,000$ forecasts in each casino and the resulting effective daily interest rates are calculated. For these synthetic data experiments, the known true state is used as the verifying observation. The results of these calculations are shown in Fig. 2.

In considering Fig. 2, note that each of the compared ensemble forecasts are based on the same deterministic forecast x^f and the same ensemble variance prediction s^2. The ensembles only differ in their ability to characterize the flow dependent error variance of this forecast. The invariant ensemble bases its probabilistic forecast solely on the climatological error variance of the forecast, while the MSS ensemble ignores climatological information about error variance and solely relies on a bias corrected form of s^2. Figure 2 shows that the gambler can make more money in the MSS ensemble casino than in the invariant ensemble casino for very small effective ensemble sizes (illustrated by the $M_e = 2$ and $M_e = 4$ cases). This is because the stochastic variations of the MSS prediction of ensemble variance about the actual variance are very large when M_e is small. For $M_e \geq 6$, the MSS ensemble outperforms the

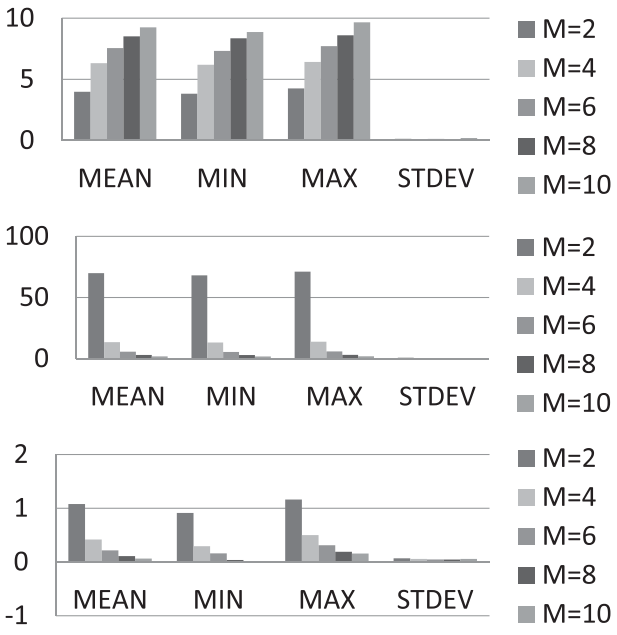

FIG. 2. The mean, maximum, minimum, and standard deviation of the effective daily interest rate (in percent) earned by the FP ensemble when the (top) invariant, (middle) MSS, and (bottom) informed Gaussian are played for 10 independent trials, each consisting of 100 000 forecast events. Effective ensemble sizes of $M_e = 2, M_e = 4, M_e = 6, M_e = 8$ and $M_e = 10$ are shown.

invariant ensemble because its ability to track the significant variations in error variance outweighs the deleterious effects of its stochastic fluctuations. The informed-Gaussian ensemble optimally combines the error variance information used in the MSS and invariant ensembles. Because of this, the gamblers winnings from the informed-Gaussian casino are considerably less than that from the MSS and invariant casinos. Nevertheless, Fig. 2 shows that the gambler using the FP ensemble still wins money in the informed-Gaussian casino. This is because the FP ensemble accounts for the fact that, for small M_e values, there is a broad distribution of error variances given the sample variance s^2 whereas the informed-Gaussian ensemble does not.

We see that the gamblers positive effective daily interest rate per trial in the informed-Gaussian casino decreases markedly as M_e increases from 2 to 10. This is a direct consequence of the fact that distribution of error variances given s^2 narrows as M_e increases and hence in the limit of infinite M_e there is no difference between the informed-Gaussian ensemble and the FP ensemble. Note that the negative minimum values seen when the gambler using the FP ensemble plays the informed-Gaussian ensemble are simply due to the effective daily interest rate tending toward zero, and the sample size not being large enough to discern this fact. As the effective ensemble size M tends toward infinity (increasing the parameter k), nonzero earnings are solely due to random effects. Conversely, as the effective ensemble size M tends toward zero, the accuracy of the ensemble-based variance prediction goes to zero and the posterior distribution of error variances tends toward the climatological distribution of error variances. In this case, if the climatological distribution of error variances had a variance of zero, the FP, the informed-Gaussian, and the invariant ensembles would converge. However, if the variance of the climatological distribution of error variances is nonzero and the effective ensemble size tends to zero, the FP becomes superior to the informed Gaussian because it accounts for the variance of the climatological distribution of error variances.

The relative variance of the assumed prior climatological inverse-Gamma distribution of error variances is given by $\mathrm{var}(\sigma^2)/(\langle\sigma^2\rangle - \sigma^2_{\min})^2 = 1/(\alpha_{\mathrm{prior}} - 2)$. We control this parameter by simply adjusting $\mathrm{var}(\sigma^2)$. Figure 3 shows the results of weather roulette when a gambler using the FP ensemble plays the invariant, MSS, and informed-Gaussian ensembles for cases in which $1/(\alpha_{\mathrm{prior}} - 2)$ has been increased by a factor of 10 (left panels) and decreased by a factor of 10 (right panels). The α_{prior} values for the small, medium, and large $\mathrm{var}(\sigma^2)$ cases are 3.30, 2.13, and 2.01, respectively. In the limit that $\mathrm{var}(\sigma^2)/(\langle\sigma^2\rangle - \sigma^2_{\min})^2$ becomes small (i.e., we have the same forecast error variance every day), the informed-Gaussian, the invariant, and the FP ensembles converge. Comparing the left and the right panels, we see that a gambler using the FP ensembles wins less (more) money against the informed-Gaussian ensemble when $\mathrm{var}(\sigma^2)/(\langle\sigma^2\rangle - \sigma^2_{\min})^2$ is decreased (increased) by a factor of 10. However, even in the small $\mathrm{var}(\sigma^2)$ case, the gambler using the FP ensemble is still able to win against the informed-Gaussian ensemble for all cases except $M_e = 8$ and $M_e = 10$. Again, note that the negative minimum values seen for the $M_e = 8$ and $M_e = 10$ case are due to the effective daily interest rate tending toward zero as the posterior distribution narrows, and the sample size not being large enough to reduce the standard deviation. We also note that in the reduced $\mathrm{var}(\sigma^2)$ case, the MSS ensemble does not outperform the invariant ensemble until $M_e \geq 8$, effective ensemble members are used. In the case of large $\mathrm{var}(\sigma^2)$ (left panels of Fig. 3), we see that the relative performance of the MSS ensemble improves. At the same time, with large $\mathrm{var}(\sigma^2)$, the performance of the invariant and the informed Gaussian is worse because only the FP accounts for the large possible values of error variances that can occur in this large $\mathrm{var}(\sigma^2)$ case.

Aside from the parameters that define the prior and posterior distributions, the results are also sensitive to

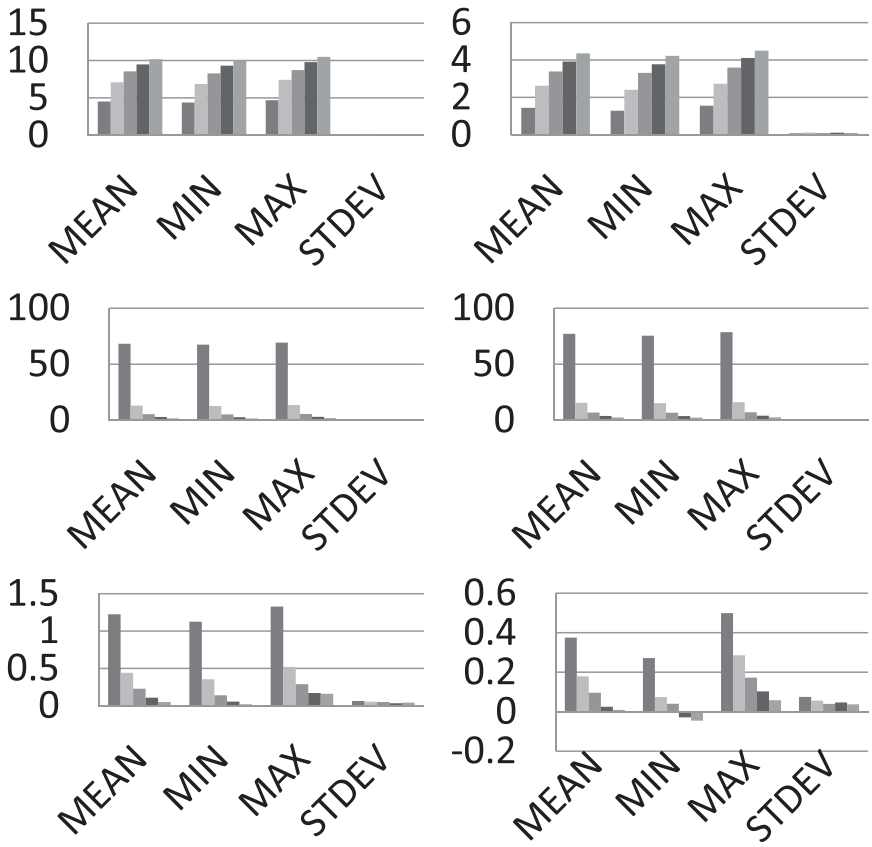

FIG. 3. As in Fig. 2, but (left) $1/(\alpha^2)$ increased by a factor of 10 and (right) $1/(\alpha^2)$ decreased by a factor of 10.

the bin width used in weather roulette. Recall that we had set $N_b = 100$ to define our hypothetical roulette wheel. If we were to set the bin size large enough to average out any differences seen in the RFH of Fig. 1, the superiority of the FP ensemble over the informed-Gaussian ensemble might be harder to detect. To test this, we decreased the number of bins to $N_b = 10$ while using the same parameters which were used in the control experiment shown in Fig. 2. The results are shown in Fig. 4. While, even with the increased bin width, the FP ensemble still wins against the informed-Gaussian ensemble sizes of $M_e = 2$, 4 we start to see small negative minimum values for $M_e = 6$ indicating that the weather roulette diagnostic can no longer detect the superiority of the FP ensemble at this sample size. Figure 4 shows that the daily effective interest rate obtained by a gambler using the FP ensemble decreases as number of bins decrease (increasing the bin width) regardless of whether the invariant, the MSS, or the informed-Gaussian ensemble is used to define the odds. This decrease indicates that when an increased bin width is used, the weather roulette diagnostic has more difficulty distinguishing which ensemble is superior. For this reason, in the following section we choose $N_b = 100$, where the synthetic data results indicate that the weather roulette diagnostic can easily distinguish ensemble performance.

3. Application to 500-hPa virtual temperature fields

a. *Datasets*

We now test our postprocessing algorithms on the raw ensemble forecasts produced by the FNMOC EFS. This ensemble employs the ensemble transform (ET) technique (Bishop and Toth 1999; McLay et al. 2008) to produce initial conditions for 20 forecasts made using the Navy Operational Global Atmospheric Prediction System (NOGAPS). The NOGAPS model used for the ensemble forecasts has a spectral resolution of T159 and 42 vertical levels. For this study, we focus on 500-hPa virtual temperatures. We apply postprocessing techniques to ensemble forecasts made at 0000 and 1200 UTC for lead times at 24-h increments from 24 to 168 forecast hours and consider the time period of

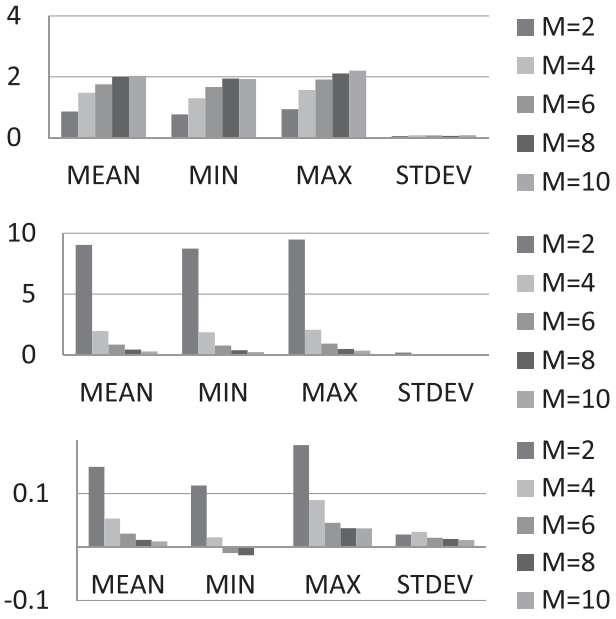

FIG. 4. As in Fig. 2, but only $N_b = 10$ weather roulette bins are used.

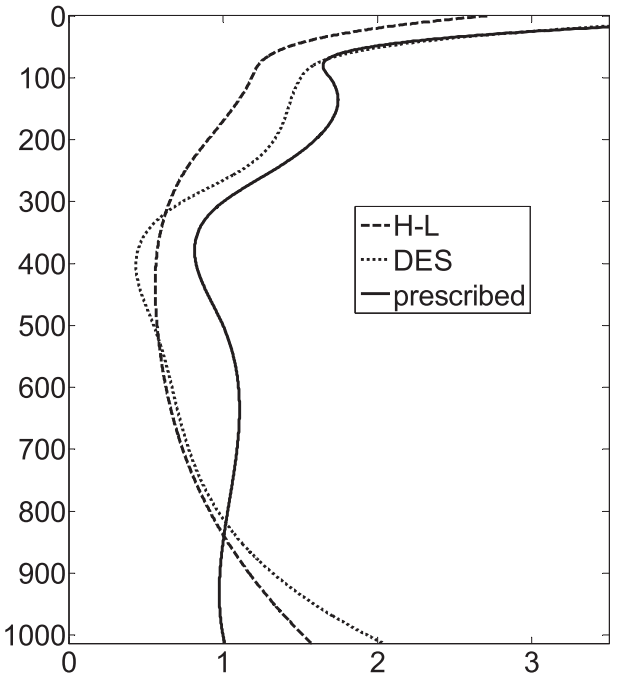

FIG. 5. A comparison of the recovered observation error variance using the Derozier method (dotted line) and the Hollingsworth–Lönnberg method (dashed line) data used from temperature measurements of radiosondes and NOGAPS analyses at T319 spectral resolution for 1 Feb–1 Apr 2011. The prescribed value is also shown (solid line). Values were recovered at eight pressure levels and a cubic spline interpolation was used to plot the vertical profile.

0000 UTC 1 March–1200 UTC 31 May 2012. For this time period, the resolution of the deterministic forecast was T319 with 42 vertical levels. For the results presented in this paper, a 1° × 1° gridpoint representation of these fields was used.

b. *Recovering observation error variance*

To compute the mean of the climatological distribution of error variances $\langle \sigma^2_{\text{prior}} \rangle$, we simply subtract the observation error variance, denoted R, from the innovation variance [see (A1) in the appendix]. Often, the specification of these observation error variances within the data assimilation system is out dated or ad hoc. For the recovery of hidden error variance parameters, it is necessary to estimate the actual value of R as accurately as possible. To achieve this estimate, we compared two commonly used methods to estimate the observation error variance: the Hollingsworth–Lönnberg method (Hollingsworth and Lönnberg 1986) and the Desroziers method (Desroziers et al. 2005). The Hollingsworth–Lönnberg method involves using innovation statistics from a dense observing network to calculate a histogram of innovation covariances binned by horizontal separation. By assuming uncorrelated observation errors and using an isotropic correlation model to extrapolate to zero separation one can obtain an estimate of the background error variance. Subtracting this value from the innovation variance provides the observation error variance.

The Desroziers method is based on computing the expected value of the analysis residual (observation minus analysis) and innovation (observation minus background). It can be shown that, if the data assimilation system properly specifies the background and observation error variance, this expected value will result in the observation error variance. Desroziers et al. (2005) suggested an iterative method for converging on the true value of the observation error variance if the observation error covariance matrix is inappropriately specified. In the present study, however, we do not apply an iterative procedure, since this would be computationally costly. Instead, we simply compare the result from the Desroziers diagnostic to those recovered from the Hollingsworth–Lönnberg method. Figure 5 shows a comparison of the recovered values of observation error variance for radiosonde observations of temperature in the Northern Hemisphere. The data for use in Fig. 5 came from NOGAPS analyses at T319 spectral resolution for the period of 1 February–1 April 2011. Both methods were applied at eight vertical levels (1000, 850, 500, 300, 250, 100, 50, and 10 hPa) and the recovered values were interpolated to all model levels using a cubic spline interpolation. The value of R prescribed within the data assimilation system is also

TABLE 1. Hidden error variance parameters recovered from time series of innovation and ensemble variance pairs. Forecasts with lead times ranging from 24 to 168 h, valid during the period (top) 1 Mar–30 Apr 2012 and (bottom) 1 Apr–31 May 2012, were considered.

	24	48	72	96	120	144	168
1 Mar–30 Apr 2012							
No. of points	46 362	45 542	45 120	44 308	43 927	43 567	43 598
$\langle\sigma^2\rangle$	0.806	1.88	3.678	6.352	9.862	13.74	17.66
Var(σ^2)	0.866	3.658	15.34	40.66	88.69	136.9	189.9
σ^2_{min}	0	0	0	0	0	0	0
a	0.572	0.766	0.68	0.612	0.563	0.585	0.547
k	0.66	1.731	1.946	2.243	2.738	5.012	5.408
Mean(s^2)	0.936	1.734	2.991	4.653	6.585	8.609	10.56
α_{prior}	2.75	2.967	2.882	2.992	3.097	3.38	3.642
$1/(\alpha_{\text{prior}} - 2)$	1.333	1.035	1.134	1.008	0.912	0.725	0.609
1 Apr–31 May 2012							
No. of points	45 716	44 896	44 887	44 903	44 502	44 510	44 116
$\langle\sigma^2\rangle$	0.72	1.684	3.142	5.488	8.707	11.92	15.26
Var(σ^2)	0.774	3.033	8.917	32.41	72.88	104.6	136.4
σ^2_{min}	0	0	0	0	0	0	0.06
a	0.66	0.845	0.935	0.691	0.65	0.663	0.626
k	0.663	1.542	2.686	2.585	4.937	7.338	6.721
Mean(s^2)	0.948	1.751	2.993	4.526	6.221	7.982	9.636
α_{prior}	2.67	2.935	3.107	2.929	3.04	3.358	3.694
$1/(\alpha_{\text{prior}} - 2)$	1.493	1.07	0.903	1.076	0.961	0.736	0.59

shown. At lower and middle pressure levels we see a better agreement between both recovered values of R than the value prescribed, the exception is the stratospheric levels where data are sparse. For the 500-hPa level, which is the focus of this study we find good agreement: 0.5518 for Desroziers and 0.5735 from Hollingsworth–Lönnberg. In what follows, we simply use the value obtained from Hollingsworth–Lönnberg method.

c. *Application of parameter recovery equations and heteroscedastic postprocessing*

To recover the parameters [$\langle\sigma^2\rangle$, var(σ^2), σ^2_{min}] defining the climatological distribution of true error variances and the parameters (k, a, s^2_{min}) defining the distribution of ensemble variances given a true error variance, the method of Bishop et al. (2013) requires a time series of innovation and ensemble variance pairs in a region where ensemble performance can be considered homogeneous in terms of dispersion and ability to track changes in forecast error variance. One may consider applying these parameter recovery equations separately for different geographic regions, or seasons. For the results presented in this paper, we restrict our application of postprocessing algorithms to the Northern Hemisphere (30°–90°N) and we restrict the observation type to radiosondes.

The parameter recovery equations, derived in Bishop et al. (2013) and summarized in the appendix, assume that the forecasts and observations are unbiased and that observation error and forecast error are uncorrelated. If a global bias exists (i.e., the model has the wrong climatology), that bias should be removed prior to the application of the recovery equations and postprocessing. In the present dataset, the mean bias was found to be small enough (maximum of −0.17 K at a 24-h lead time), that no improvement was shown when an out of sample bulk bias correction was applied. Therefore, in what follows no systematic bias correction is included. Equation (8) acts to push the deterministic forecast toward the climatological mean in the same way that a regression-based correction would. However, the climatological weighting changes with day-to-day changes in forecast error variance.

The values of parameters recovered using (A1)–(A6) from two time periods, 1 March–30 April 2012 and 1 April–3 May, are summarized in Table 1. Values of recovered parameters are shown for lead times between 24 and 168 h at 24-h increments. The value of a is less than 1 for all lead times and for both time periods, indicating that the ensemble is typically underdispersive. The nonzero values of k corresponds to effective ensemble size ($M = 2k + 1$) larger than one, which indicates that the raw FNMOC ensemble has at least some skill in predicting the flow-dependent error variance. The value of k, and hence the effective ensemble size increases with lead time. This behavior is consistent with the findings of Satterfield and Szunyogh (2010), who showed that ensembles span a greater proportion of the vector space of forecast errors as the lead time increases. The recovered values are in fairly good agreement between the two time periods. This agreement

makes the idea of recovering parameters during a training period of the previous two months plausible. In what follows, we use recovered parameters from the training period of 1 March–30 April 2012 and apply those parameters to postprocess ensemble data for the month of May 2012.

Section 2f discussed the sensitivity of the benefits associated with the use of the FP ensemble to the parameter regime. Guidance from synthetically generated data indicated that the FP ensemble outperformed the informed-Gaussian ensemble up to values of $1/(\alpha_{\text{prior}} - 2) = 0.77$ with an effective ensemble size of $M_e = 8$. Although we see that $1/(\alpha_{\text{prior}} - 2)$ tends to decrease with lead time, we expect to see benefits from the use of the FP ensemble for lead times out to ~144 h.

4. Comparison to non-FP homoscedastic ensembles and Bayesian model averaging

To create an $M = 1000$ postprocessed ensemble from the FNMOC ensembles we use the same approach as that for the synthetic data following (8) to obtain a 1000-member sample from our estimate of the distribution of truth given the forecast. Then, since we use observations for verification we add, to each of the 1000 ensemble members, random normal numbers with mean zero and variance equal to the observation error variance R, following Bowler (2006). This second step yields an estimate of the distribution of observations given the forecast. To define the climatological mean and standard deviation, $\langle x^t \rangle_C$ and σ_C, needed to apply (8), we simply compute their respective values, in 5° latitude bands centered at the latitude of interest, for the months of March and April and linearly extrapolate through time to approximate the values for May. We note that the values of σ_C obtained through this extrapolation are significantly smaller than the actual values calculated from the May data, especially in the 40°–55° latitude range, where many of the radiosonde observations are located. This underestimation of σ_C can cause the postprocessed ensemble to overweight the climatology at lead times where σ_C^2 and $\langle \sigma^2 \rangle$ have comparable magnitudes. In practice, a more suitable climatology could be formed by calculating $\langle x^t \rangle_C$ and σ_C from, for example, a 10-yr dataset of monthly data.

a. Comparison to homoscedastic ensembles

We compare the results of our heteroscedastic postprocessing methods with the three homoscedastic methods described in section 2b. Again, we first assess the performance of these postprocessing methods using RFHs. Figure 6 shows the rank frequency histograms for the raw ensemble and the FP ensemble, as well as the ensembles postprocessed using homoscedastic methods. For this figure, all postprocessing parameters were derived out of sample. This figure considers all 48-h ensemble forecasts for the Northern Hemisphere (30°–90°N) valid during the month of May 2012 and uses 500-hPa radiosonde observations of temperature as the verifying observation. For the month of May, there are $N = 23\,927$ observations, to achieve a perfectly flat rank frequency histogram with $M = 1000$ ensemble forecasts each bin would have a frequency of 23–24. For the raw ensemble with 20 members, the optimal frequency would be 1139–1140. The raw ensemble shows a high bias, with observations falling below the minimum ensemble value more frequently. When the raw ensemble is recentered on the higher-resolution deterministic forecast (shown in Fig. 6b) the bias is largely removed.

The FP ensemble gives a RHF with overpopulated center bins. The overpopulation in the center bins is most likely due to the recovered value of the ensemble sensitivity parameter a for the training period 1 March–30 April being slightly lower than the actual value for the month of May (as indicated by Table 1). The invariant ensemble is markedly peaked in the center bins and also has overpopulated outer bins, very similar to the results achieved when synthetic data were used (shown in Fig. 1b). The MSS ensemble shows overpopulated outer bins again similar to the corresponding result (Fig. 1c) for the synthetic data. As with the synthetic data, the RFH produced when the informed-Gaussian ensemble is used, is the closest to that of the FP ensemble.

Figure 7 shows the mean, minimum, maximum, and standard deviation of the effective daily interest rate (in terms of percent) achieved over five trials when the FP ensemble plays the homoscedastic non-FP ensembles. Each trial has the same original FNMOC EFS ensemble and recovered parameters, but different random draws from the posterior and a different set of 10 000 randomly drawn verifying observations. Equally probable weather roulette bins were determined from the full set of observations. At the 24-h lead time the invariant ensemble performs very similarly to the informed Gaussian due to small values of effective ensemble size and that $\text{var}(\sigma^2)/(\langle \sigma^2 \rangle - \sigma^2_{\text{min}})^2 = 1/(\alpha_{\text{prior}} - 2)$ is not large enough to allow the ensemble variance prediction to influence the posterior distribution (Table 1). As the forecast lead time increases, so does $\text{var}(\sigma^2)$, and the FP ensemble wins increasing amounts of money when the invariant ensemble is played. The MSS ensemble begins to outperform the invariant ensemble around the 72-h lead time, this lead time corresponds to an effective ensemble size of $M_e \approx 5$, comparable to the results found when synthetic data were used. When the informed-Gaussian ensemble is played, the highest maximum

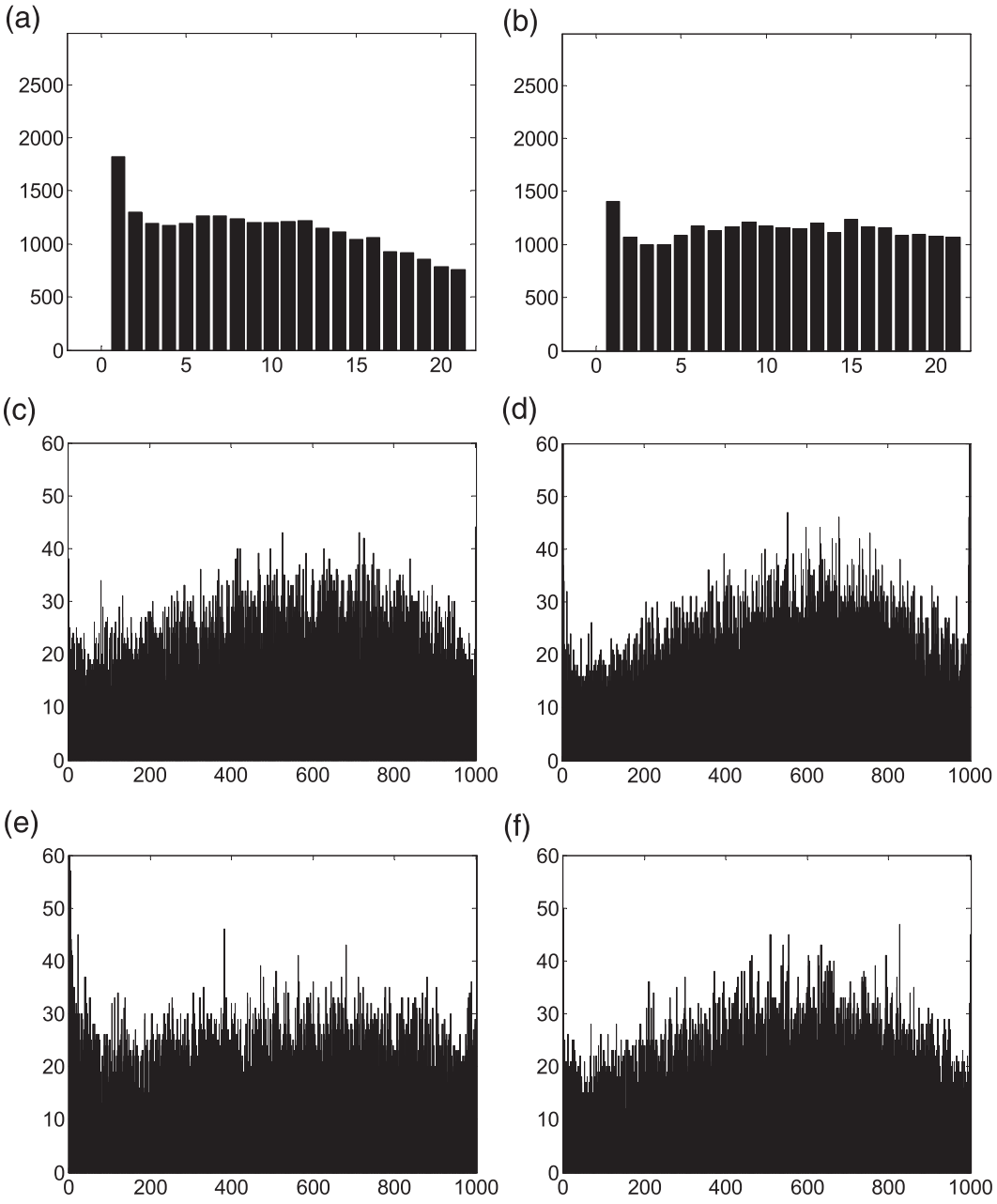

FIG. 6. RFH obtained from an $M = 1000$ member postprocessed FNMOC EFS ensemble. Results are shown for (a) the raw ensemble, (b) the raw ensemble recentered on the higher-resolution deterministic forecast, (c) FP ensemble, (d) invariant ensemble, (e) MSS ensemble, and (f) informed-Gaussian ensemble. Shown are 48-h forecasts valid for May 2012 postprocessed using hidden error variance parameters derived from the training period 1 Mar–30 Apr 2012.

values achieved are 1.39 and 1.57 at the 96- and 120-h lead time, respectively. Even though the effective ensemble size grows with lead time, the FP ensemble still earns money when the informed-Gaussian ensemble is played at all lead times. Again, this result is consistent with the synthetic data findings that showed the benefit of the FP ensemble up to $1/(\alpha_{\text{prior}} - 2) = 0.77$ and an effective ensemble size of $M_e = 8$.

b. *Comparison to Bayesian model averaging*

Here we compare our results with the fairly well-established BMA ensemble postprocessing method of Raftery et a1. (2005). We tried to follow the method described in Raftery et al. (2005) as closely as possible while preserving consistency, where possible, with the other methods used in this paper.

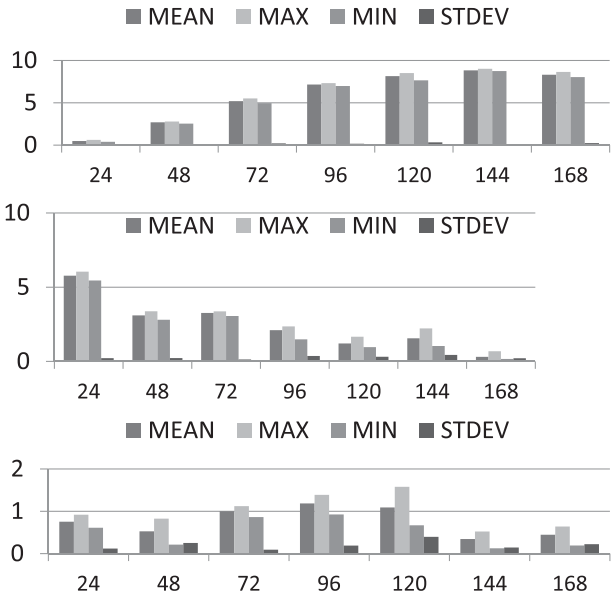

FIG. 7. The mean, maximum, minimum, and standard deviation of the effective daily interest rate (in percent) earned by the FP ensemble when the (top) invariant, (middle) MSS, and (bottom) informed Gaussian are played for five independent trials, each consisting of 10 000 randomly chosen forecast events. All forecasts are valid for May 2012 and postprocessed using hidden error variance parameters derived from the training period 1 Mar–30 Apr 2012.

1) We use a linear regression–based correction, where the regression coefficients are determined from the previous two months and vary latitudinally based on 5° latitude bands centered on the latitude of interest, to correct the 20 raw ensemble members following, $a_\ell x_i^f + b_\ell$, where ℓ represents latitude. Since all ensemble forecasts are run using the same model, the regression coefficients are determined for the first member and the same coefficients are applied to all members. Note that the regression step implicitly incorporates climatological information into the forecast.
2) Following the same procedure as in step 1, regression coefficients are determined for the deterministic forecast.
3) The 20 regression-corrected ensemble members combined with the single regression-corrected deterministic forecast from a 21-member ensemble are used as input to the expectation and maximization algorithm.
4) We calculate the weights w_k, for the $k = 1, \ldots, 21$ ensemble members, and variance σ using the expectation and maximization algorithm [Eq. (6) of Raftery et a1. 2005]. While, Raftery et al. (2005) found the optimal training period to be a sliding 25-day window, to be consistent with the other methods presented in this paper, we simply train on the months of March and April 2012 and show results for postprocessed forecasts valid for the month of May 2012. We also omit refinement of σ values using CRPS and simply set the convergence criterion to a change of less than 0.02 from the previous iteration.
5) We create a $M = 1000$ member postprocessed ensembles by drawing a random number between $[1, \ldots, k = 21]$ with probabilities $[w_1, \ldots, w_k]$ and drawing a value from the pdf $N(f_k, \sigma^2)$, where f_k represents the kth regression corrected ensemble forecast and σ^2 is the converged value obtained from step 4.

We follow the method of Raftery et al. (2005) and do not include climatological information using (8). However, as noted above, climatological information is included through the regression correction. Additionally, the observation error perturbation is omitted from the BMA ensemble, since it is included as part of the formulation. The RFH produced from the BMA method for the 48-h lead time is shown in the top panel of Fig. 8. The RFH shows that the ensemble is overdispersive, with the observations falling in the center bins more frequently. The bottom panel of Fig. 8 shows the weather roulette interest rate obtained when the FP ensemble plays the BMA ensemble. The FP ensemble wins a statistically significant amount of money against the BMA ensemble for all lead times greater than 48 h. We speculate that the lack of a statistically significant difference at earlier times is because of benefits conferred by the linear regression included in the BMA technique, but which we omitted from the postprocessing approaches discussed in detail here. As lead time increases, so does the amount of money won by the FP ensemble. At the 48-h lead time the BMA ensemble outperforms the informed-Gaussian ensemble; however, the informed Gaussian overtakes the BMA ensemble by the 72-h lead time. By the 144-h lead time the BMA ensemble performs similarly to the invariant ensemble. This may be because either (i) the method of incorporating climatological information given by (8) is better than that associated with BMA regression coefficients [Bishop and Shanley (2008) discussed this issue in detail] or (ii) the statistical differences between the March–April training data period and the May test period may, by random chance, have a more deleterious effect on the BMA approach than the invariant approach. A detailed analysis of this issue is beyond the scope of this paper.

Another score of interest is the binned spread-skill plot (e.g., Wang and Bishop 2003). We divide all data points into $N_b = 10$ equally populated bins, arranged in order of increasing ensemble variance. Each point shows

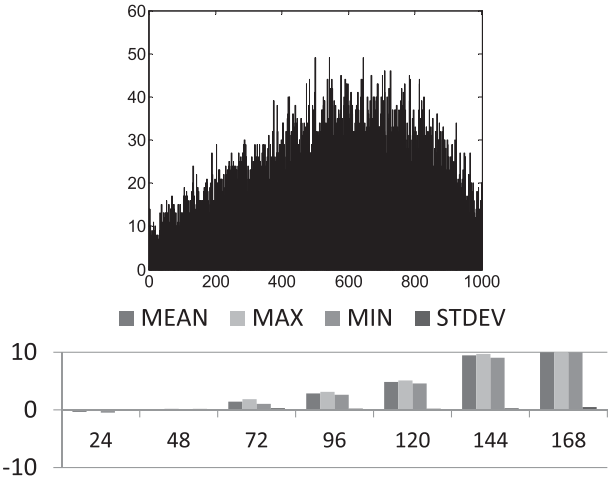

FIG. 8. (top) RFH obtained from an $M = 1000$ member FNMOC EFS ensemble postprocessed using the BMA method shown for the 48-h forecast lead time. (bottom) Weather roulette interest rate earned by the FP ensemble when the BMA ensemble is played. All forecasts are valid for May 2012 and postprocessed using hidden error variance parameters derived from the training period 1 Mar–30 Apr 2012.

the bin mean ensemble variance and the bin-averaged squared innovation. Figure 9 shows the results of the binned spread-skill plot for the 48-h lead time. Accurate predictions of forecast error variance are indicated by plotted points and the corresponding regression line falling along the diagonal (black lines). The positive slope in Fig. 9a shows that, on average, when the raw ensemble variances are larger, innovation variance is larger and hence forecast error variance is larger. However, the raw ensemble gives underdispersive lower bins and overdispersive higher bins. The FP ensemble gives a regression line fairly close to the diagonal, slightly overdispersive, indicating reasonably good calibration. The BMA ensemble is increasingly overdispersive as the forecast error variance increases. The invariant ensemble, as one would expect, performs poorly under this measure. While the informed-Gaussian ensemble performs similarly to the FP ensemble, the MSS ensemble shows underdispersive lower bins and overdispersive higher bins. At later lead times, when accurate information about May's climatology is more important, none of the methods did a particularly good job of tracking the true error variance. We attribute this to the fact that, in this particular year, the climatological variance of March and April were not particularly good linear predictors of the climatological variance of May. Presumably, the use of more relevant training datasets would deal with this problem.

We also compute the continuous ranked probability score (CRPS) for the postprocessed ensemble, as well as the decomposition into reliability, resolution, and uncertainty following Hersbach (2000). The CRPS compares the cumulative distribution functions between forecasts and observations, having an optimal value of zero. A perfectly reliable forecast is indicated by a reliability value of zero and this computation of reliability is closely related to the rank histogram (Hersbach 2000). Positive values of resolution indicate that the ensemble performs better than the climatological probabilistic forecast.

This score was chosen since we want to focus on the full distribution, not on a particular event. For our computation, we computed the CRPS for five independent trails, each trial randomly choosing $N = 10\,000$ observations as verification. Shown in Fig. 10 are the CRPS (top panels), reliability (middle panels), and resolution (bottom panels) computed as the mean values over those five trials. The standard deviation computed over the five independent trials is also shown. In terms of CRPS, initially, the BMA ensemble performs similarly to the FP ensemble; however, beyond the 48-h lead time a benefit is seen from the FP ensemble. Beyond the 48-h lead time the FP ensemble provides the best performance, converging with the MSS and informed Gaussian around the 144-h lead time. Although the BMA ensemble shows the highest values for resolution (bottom panels of Fig. 9), beyond the 24-h lead BMA shows the poorest performance for reliability. The most reliable forecasts are given by the FP ensemble for all lead times. In terms of resolution, the FP ensemble performs similarly to invariant ensemble and informed-Gaussian ensembles at shorter lead times, and similarly to the MSS and informed-Gaussian ensembles at longer lead times.

5. Conclusions

A heteroscedastic ensemble postprocessing technique has been introduced that incorporates aspects of the BPA, shift and stretch, and the NGR postprocessing methods. Unlike previous approaches, it allows climatological information to be incorporated in the forecast in a way that takes advantage of the best available estimate of the distribution of true error variances given an ensemble variance. Here, we employed estimates of the distribution of true error variances given an ensemble variance that are based on Bishop and Satterfield (2013) and Bishop et al.'s (2013) theory of hidden error variances. We also introduced a hierarchy of homoscedastic ensemble postprocessing techniques in which a single error variance is assigned to each forecast that, in differing ways, partially captured the information in Bishop and Satterfield's

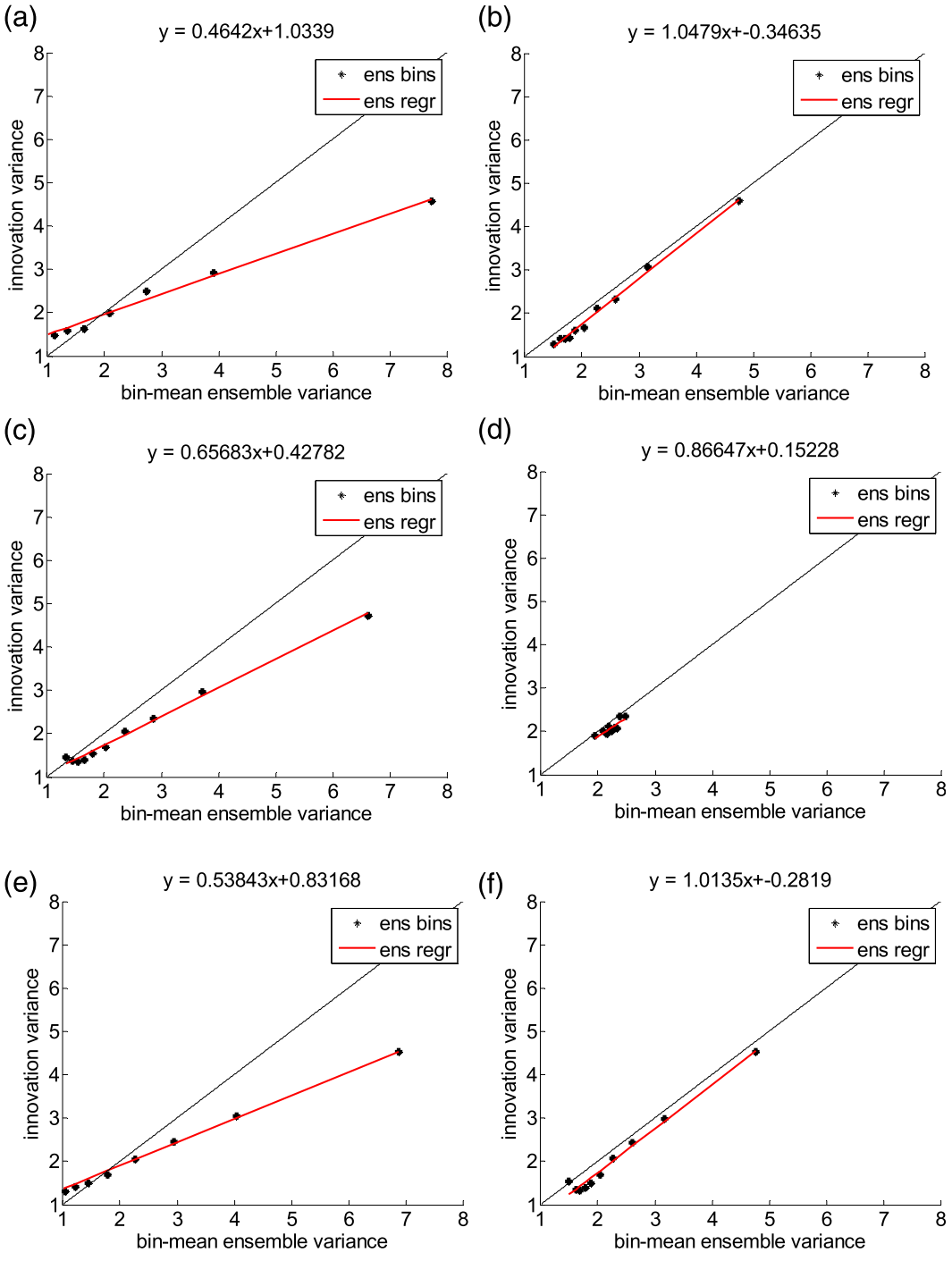

FIG. 9. Binned spread-skill diagram shown for the 48-h lead time for (a) the raw ensemble, (b) FP, (c) BMA, (d) invariant, (e) MSS, and (f) informed Gaussian. All forecasts are valid for May 2012 and postprocessed using hidden error variance parameters derived from the training period.

analytical model of the distribution of true error variances given an ensemble variance. The homoscedastic hierarchy included techniques very similar to the previously proposed BPA, shift and stretch, and NGR postprocessing techniques. Our experiments with synthetic data demonstrated that ignoring *any* aspect of the distribution of true error variances given an imperfect error variance prediction degrades the quality of probabilistic forecasts. Using real data and real ensemble forecasts from FNMOC, we again found

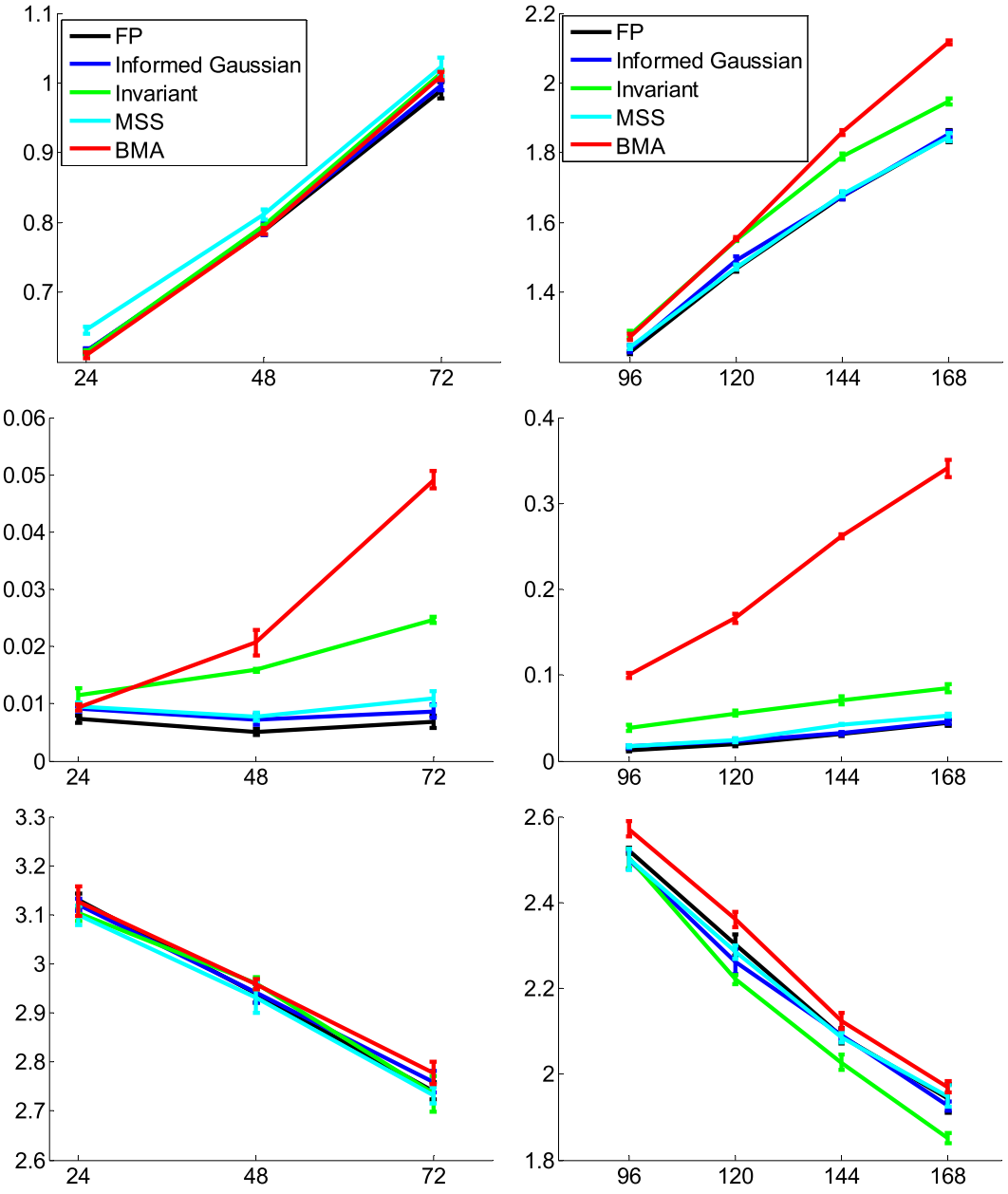

FIG. 10. (top) CRPS and the corresponding decomposition into (middle) reliability and (bottom) resolution shown for the FP ensemble (black), informed-Gaussian ensemble (blue), invariant ensemble (green), MSS ensemble (cyan), and the BMA ensemble (red). Lead times are shown at 24-h increments (left) between 24 and 72 h and (right) between 96 and 168 h. All forecasts are valid for May 2012 and postprocessed using hidden error variance parameters derived from the training period 1 Mar–30 Apr 2012.

that the more extensively the postprocessing technique incorporated information about the *estimated* posterior distribution of hidden error variances given an ensemble variance, the greater the improvement to probabilistic skill scores. In the real data test, heteroscedastic postprocessing outperformed our implementation of the BMA method.

In this paper, heteroscedastic postprocessing was applied to virtual temperatures at 500 hPa. This variable was chosen because 1) the distribution is expected to be Gaussian and 2) the bias is lower than that of variables closer to the surface. These simplifications allowed for a proof-of-concept demonstration of this method in a semi-idealized framework. For nonnormally distributed

variables, one would have to apply appropriate transformations as in Krzysztofowicz and Evans (2008) and Hamill (2008). In a companion paper, we will focus on applying this algorithm to the less idealized situations where the distribution of errors is nonnormal, as well as to cases where bias correction must be applied as a first step. In addition, this study was based on a three-month period, in practice an optimal running or historical training set would need to be defined. Finally, we note that various aspects of the postprocessing methods presented in this paper could be combined, for example BPF and BMA could be combined by including climatology in BMA through (8) as opposed to using a linear regression–based correction. Conversely, a linear regression–based correction could be applied to heteroscedastic methods prior to parameter recovery. One aspect of BMA and similar kernel dressing methods that could potentially offer a benefit for bimodal distribution is that the original structure of the ensemble is retained; for this reason in a companion paper, we will also present a kernel formulation of heteroscedastic postprocessing.

Acknowledgments. We thank the three anonymous reviewers for their valuable comments that helped to improve this manuscript. EAS gratefully acknowledges support from the Jerome and Isabella Karle Fellowship at NRL and from the NRL Base Program through BE033-03-45. CHB gratefully acknowledges support from the U.S. Office of Naval Research Grants 4304-D-O-5 and 6681-0-4-5.

APPENDIX

Estimating Parameters of Error Variance Distributions

To sample variances $(\sigma_s^2)_j$ from (5) we simply need to define the parameters $\alpha_{\text{post}} = \alpha_{\text{prior}} + k$ and $\beta_{\text{post}} = \{(s_i^2 - s_{\text{min}}^2)k/a\} + \beta_{\text{prior}}$. The equations to recover these parameters were derived in Bishop et al. (2013), we summarize them here:

$$\langle \sigma^2 \rangle = \langle v^2 - R \rangle, \tag{A1}$$

$$\begin{aligned} \text{var}(\sigma^2) &= \frac{\langle v^4 \rangle}{3} - (\langle \sigma^2 \rangle + \langle R \rangle)^2 - \text{var}(R) \\ &= (\langle \sigma^2 \rangle + \langle R \rangle)^2 \left[\frac{\text{kurtosis}(v) - 3}{3}\right] - \text{var}(R), \end{aligned} \tag{A2}$$

$$a = \frac{\text{covar}(v^2, s^2)}{\text{var}(\sigma^2)}, \tag{A3}$$

$$\sigma_{\text{min}}^2 = \langle \sigma^2 \rangle - \frac{\langle s^2 \rangle - s_{\text{min}}^2}{a}, \quad \text{and} \tag{A4}$$

$$k^{-1} = \frac{\text{var}(s^2) - a^2 \,\text{var}(\sigma^2)}{a^2[(\langle \sigma^2 \rangle - \sigma_{\text{min}}^2)^2 + \text{var}(\sigma^2)]}. \tag{A5}$$

Once these values are known, α_{prior} and β_{prior} are given by

$$\begin{aligned} \alpha_{\text{prior}} &= \frac{(\langle \sigma^2 \rangle - \sigma_{\text{min}}^2)^2}{\text{var}(\sigma^2)} + 2 \quad \text{and} \\ \beta_{\text{prior}} &= (\langle \sigma^2 \rangle - \sigma_{\text{min}}^2)\left[\frac{(\langle \sigma^2 \rangle - \sigma_{\text{min}}^2)^2 + \text{var}(\sigma^2)}{\text{var}(\sigma^2)}\right]. \end{aligned} \tag{A6}$$

When using (A4) to recover a value of σ_{min}^2, negative values are likely when σ_{min}^2 is close to zero and the sample size is small (see Bishop et al. 2013). For this reason, we modify the approach to set $\sigma_{\text{min}}^2 = 0$ when negative values are recovered. Equation (A4) then becomes

$$\sigma_{\text{min}}^2 = \begin{cases} \langle \sigma^2 \rangle - \dfrac{\langle s^2 \rangle - s_{\text{min}}^2}{a} & \text{if} \quad \langle \sigma^2 \rangle - \dfrac{\langle s^2 \rangle - s_{\text{min}}^2}{a} \geq 0 \\ 0 & \text{if} \quad \langle \sigma^2 \rangle - \dfrac{\langle s^2 \rangle - s_{\text{min}}^2}{a} < 0 \end{cases}. \tag{A7}$$

One could obtain this same result more formally by first assuming a prior uniform distribution of σ_{min}^2 going from 0 to positive infinity and then use the recovered value of σ_{min}^2 to govern some likelihood distribution of recovered σ_{min}^2 values given a true σ_{min}^2. With this setup, the most likely value of σ_{min}^2 when the recovered value of σ_{min}^2 was negative would always be zero.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530, doi:10.1175/1520-0442(1996)009<1518:AMFPAE>2.0.CO;2.

Bishop, C. H., and Z. Toth, 1999: Ensemble transformation and adaptive observations. *J. Atmos. Sci.,* **56,** 1748–1765, doi:10.1175/1520-0469(1999)056<1748:ETAAO>2.0.CO;2.

——, 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:10.1175/2008MWR2565.1.

——, and E. A. Satterfield, 2013: Hidden error variance theory. Part I: Exposition and analytic model. *Mon. Wea. Rev.,* **141,** 1454–1468, doi:10.1175/MWR-D-12-00118.1.

——, ——, and K. T. Shanley, 2013: Hidden error variance theory. Part II: An instrument that reveals hidden error variance distributions from ensemble forecasts and observations. *Mon. Wea. Rev.,* **141,** 1469–1483, doi:10.1175/MWR-D-12-00119.1.

Bowler, N. E., 2006: Explicitly accounting for observation error in categorical verification of forecasts. *Mon. Wea. Rev.,* **134,** 1600–1606, doi:10.1175/MWR3138.1.

Cai, X., M. Hejazi, and D. Wang, 2011: Value of probabilistic weather forecasts: Assessment by real-time optimization of irrigation scheduling. *J. Water Resour. Plann. Manage.,* **137,** 391–403, doi:10.1061/(ASCE)WR.1943-5452.0000126.

Daley, R., 1991: *Atmospheric Data Analysis.* Cambridge University Press, 457 pp.

Desroziers, G., L. Berre, B. Chapnik, and P. Poli, 2005: Diagnosis of observation, background and analysis-error statistics in observation space. *Quart. J. Roy. Meteor. Soc.,* **131,** 3385–3396, doi:10.1256/qj.05.108.

Eckel, F. A., M. S. Allen, and M. C. Sittel, 2012: Estimation of ambiguity in ensemble forecasts. *Wea. Forecasting,* **27,** 50–69, doi:10.1175/WAF-D-11-00015.1.

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, doi:10.1256/qj.05.167.

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, doi:10.1175/MWR2904.1.

Hagedorn, R., and L. A. Smith, 2009: Communicating the value of probabilistic forecasts with weather roulette. *Meteor. Appl.,* **16,** 143–155, doi:10.1002/met.92.

Hamill, T. M., cited 2008: Krzysztofowicz and Evans' ''Bayesian processing of forecasts'' evaluation with GFS reforecasts. National Oceanic and Atmospheric Administration. [Available online at http://www.esrl.noaa.gov/psd/people/tom.hamill/BPF_review_hamill.pdf.]

——, and S. J. Colucci, 1997: Verification of Eta–RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327, doi:10.1175/1520-0493(1997)125<1312:VOERSR>2.0.CO;2.

Hersbach, H., 2000: Decomposition of the continuous ranked probability score for ensemble prediction systems. *Wea. Forecasting,* **15,** 559–570, doi:10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2.

Hollingsworth, A., and P. Lönnberg, 1986: The statistical structure of short-range forecast errors as determined from radiosonde data. Part I: The wind field. *Tellus,* **38A,** 111–136, doi:10.1111/j.1600-0870.1986.tb00460.x.

Houtekamer, P. L., L. Lefaivre, J. Derome, H. Ritchie, and H. L. Mitchell, 1996: A system simulation approach to ensemble prediction. *Mon. Wea. Rev.,* **124,** 1225–1242, doi:10.1175/1520-0493(1996)124<1225:ASSATE>2.0.CO;2.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.,* **137,** 1717–1720, doi:10.1175/2009MWR2715.1.

Krzysztofowicz, R., and W. B. Evans, 2008: Probabilistic forecasts from the National Digital Forecast Database. *Wea. Forecasting,* **23,** 270–289, doi:10.1175/2007WAF2007029.1.

McLay, J. G., C. H. Bishop, and C. A. Reynolds, 2008: Evaluation of the ensemble transform analysis perturbation scheme at NRL. *Mon. Wea. Rev.,* **136,** 1093–1108, doi:10.1175/2007MWR2010.1.

Mylne, K. R., 2002: Decision-making from probability forecasts based on forecast value. *Meteor. Appl.,* **9,** 307–315, doi:10.1017/S1350482702003043.

Palmer, T. N., 2002: The economic value of ensemble forecasts as a tool for risk assessment: From days to decades. *Quart. J. Roy. Meteor. Soc.,* **128,** 747–774, doi:10.1256/0035900021643593.

——, R. Gelaro, J. Barkmeijer, and R. Buizza, 1998: Singular vectors, metrics, and adaptive observations. *J. Atmos. Sci.,* **55,** 633–653, doi:10.1175/1520-0469(1998)055<0633:SVMAAO>2.0.CO;2.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using Bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.,* **133,** 1155–1174, doi:10.1175/MWR2906.1.

Richardson, D., 2000: Skill and relative economic value of the ECMWF ensemble prediction system. *Quart. J. Roy. Meteor. Soc.,* **126,** 649–667, doi:10.1002/qj.49712656313.

Roulston, M. S., and L. A. Smith, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30, doi:10.1034/j.1600-0870.2003.201378.x.

Satterfield, E., and I. Szunyogh, 2010: Predictability of the performance of an ensemble forecast system: Predictability of the space of uncertainties. *Mon. Wea. Rev.,* **138,** 962–981, doi:10.1175/2009MWR3049.1.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. ECMWF Workshop on Predictability,* Reading, United Kingdom, ECMWF, 1–26.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319, doi:10.1175/1520-0493(1997)125<3297:EFANAT>2.0.CO;2.

——, Y. Zhu, and T. Marchok, 2001: The use of ensembles to identify forecasts with small and large uncertainty. *Wea. Forecasting,* **16,** 463–477, doi:10.1175/1520-0434(2001)016<0463:TUOETI>2.0.CO;2.

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

——, and ——, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986, doi:10.1256/qj.04.120.

Zhu, Y., Z. Toth, R. Wobus, D. Richardson, and K. Mylne, 2002: The economic value of ensemble-based weather forecasts. *Bull. Amer. Meteor. Soc.,* **83,** 73–83, doi:10.1175/1520-0477(2002)083<0073:TEVOEB>2.3.CO;2.