Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation

TILMANN GNEITING, ADRIAN E. RAFTERY, ANTON H. WESTVELD III, AND TOM GOLDMAN

Department of Statistics, University of Washington, Seattle, Washington

(Manuscript received 14 May 2004, in final form 21 September 2004)

ABSTRACT

Ensemble prediction systems typically show positive spread-error correlation, but they are subject to forecast bias and dispersion errors, and are therefore uncalibrated. This work proposes the use of ensemble model output statistics (EMOS), an easy-to-implement postprocessing technique that addresses both forecast bias and underdispersion and takes into account the spread-skill relationship. The technique is based on multiple linear regression and is akin to the superensemble approach that has traditionally been used for deterministic-style forecasts. The EMOS technique yields probabilistic forecasts that take the form of Gaussian predictive probability density functions (PDFs) for continuous weather variables and can be applied to gridded model output. The EMOS predictive mean is a bias-corrected weighted average of the ensemble member forecasts, with coefficients that can be interpreted in terms of the relative contributions of the member models to the ensemble, and provides a highly competitive deterministic-style forecast. The EMOS predictive variance is a linear function of the ensemble variance. For fitting the EMOS coefficients, the method of minimum continuous ranked probability score (CRPS) estimation is introduced. This technique finds the coefficient values that optimize the CRPS for the training data. The EMOS technique was applied to 48-h forecasts of sea level pressure and surface temperature over the North American Pacific Northwest in spring 2000, using the University of Washington mesoscale ensemble. When compared to the bias-corrected ensemble, deterministic-style EMOS forecasts of sea level pressure had root-mean-square error 9% less and mean absolute error 7% less. The EMOS predictive PDFs were sharp, and much better calibrated than the raw ensemble or the bias-corrected ensemble.

1. Introduction

During the past decade, the use of forecast ensembles for assessing the uncertainty of numerical weather predictions has become routine. Three operational methods for the generation of synoptic-scale ensembles have been developed: the breeding growing modes method used by the National Centers for Environmental Prediction (NCEP; Toth and Kalnay 1997), the singular vector method used by the European Centre for Medium-Range Weather Forecasts (ECMWF; Molteni et al. 1996), and the perturbed observations method used by the Canadian Meteorological Centre (CMC; Houtekamer et al. 1996). More recently, mesoscale short-range ensembles have been developed, such as the University of Washington ensemble system over the North American Pacific Northwest (Grimit and Mass 2002; Eckel 2003). The ability of ensemble systems to improve deterministic-style forecasts and to predict forecast skill has been convincingly established. Statistically significant spread-error correlations suggest that ensemble variance and related measures of ensemble spread are skillful indicators of the accuracy of the ensemble mean forecast (Eckel and Walters 1998; Stensrud and Yussouf 2003; Scherrer et al. 2004).

Case studies in probabilistic weather forecasting have typically focused on the prediction of categorical events. Ensembles also allow for probabilistic forecasts of continuous weather variables, such as air pressure and temperature, which are ideally expressed in terms of predictive cumulative distribution functions (CDFs) or predictive probability density functions (PDFs). However, ensemble systems are finite and typically include of 5 to 50 member models. Hence, raw ensemble output does not provide predictive PDFs, and some form of postprocessing is required (Hamill and Colucci 1998; Richardson 2001). Various challenges in the statistical postprocessing of ensemble output have been described. Systematic biases are substantial in current modeling systems (Atger 2003; Mass 2003) and might disguise probabilistic forecast skill. Furthermore, forecast ensembles are typically underdispersive (Hamill and Colucci 1997; Eckel and Walters 1998).

Corresponding author address: Tilmann Gneiting, Department of Statistics, University of Washington, Box 354320, Seattle, WA 98195-4322.
E-mail: tilmann@stat.washington.edu

In this paper, we propose the use of ensemble model output statistics (EMOS), an easy to implement statistical postprocessing technique that addresses the aforementioned issues. Our method is a variant of multiple linear regression or model output statistics (MOS) techniques that have traditionally been used for deterministic-style and probability of precipitation forecasts (Glahn and Lowry 1972; Wilks 1995). Specifically, suppose that $X_1, \ldots, X_m$ denotes an ensemble of individually distinguishable forecasts for a univariate weather quantity Y. A multiple linear regression equation for Y in terms of the ensemble member forecasts can be written as

$$Y = a + b_1X_1 + \cdots + b_mX_m + \varepsilon, \quad (1)$$

where a and $b_1, \ldots, b_m$ are regression coefficients, and where ε is an error term that averages to zero. Regression approaches of this type have been shown to improve the deterministic-style forecast accuracy of synoptic weather and seasonal climate ensembles (Krishnamurti et al. 1999, 2000; Kharin and Zwiers 2002), and the associated forecast systems have been referred to as superensembles.

The use of regression techniques for probabilistic forecasting has not received much attention in the literature, with the exception of forecasts of binary events (Glahn and Lowry 1972; Stefanova and Krishnamurti 2002). In this work, we obtain full predictive PDFs and CDFs from ensemble forecasts of a continuous weather variable. Standard regression theory suggests a straightforward way of constructing predictive PDFs and CDFs from a regression equation, by taking them to be Gaussian with predictive mean equal to the regression estimate, and predictive variance equal to the mean squared prediction error for the training data. This approach corrects for model biases and takes account of dispersion errors. However, the resulting assessment of uncertainty is static, in that the predictive variance is independent of the ensemble spread, thereby negating the spread-skill relationship (Whitaker and Loughe 1998). Hence, we model the variance of the error term in Eq. (1) as a linear function of the ensemble spread, that is,

$$\mathrm{Var}(\varepsilon) = c + dS^2, \quad (2)$$

where S^2 is the ensemble variance, and where c and d are nonnegative coefficients. Combining (1) and (2) yields the Gaussian predictive distribution

$$\mathcal{N}(a + b_1X_1 + \cdots + b_mX_m, c + dS^2)$$

whose mean derives from the regression equation and forms a bias-corrected weighted average of the ensemble member forecasts, and whose variance depends linearly on the ensemble variance. We refer to the resulting predictive PDFs and CDFs as ensemble model output statistics or EMOS forecasts. Negative regression weights can, and frequently do, occur in this type of formulation, as in Tables 2, 4, 5, and 6 of Van den Dool and Rukhovets (1994). This effect is typically caused by collinearities of the ensemble member forecasts, and the negative weights seem hard to interpret. They imply, all else being equal, that sea level pressure, say, is predicted to be lower when the forecast with the negative weight is higher. To address this issue, we propose an alternative implementation of the EMOS technique, which constrains the coefficients $b_1, \ldots, b_m$ to be nonnegative. We call this variant of the EMOS technique EMOS$^+$. In our experiments, EMOS and EMOS$^+$ gave equally skillful probabilistic forecasts, but the EMOS$^+$ coefficients were easier to interpret.

TABLE 1. Phase I of the University of Washington mesoscale short-range ensemble, Jan–Jun 2000. Initial conditions (ICs) and lateral boundary conditions (LBCs) were obtained from the AVN, the NGM data assimilation system, and the Eta data assimilation system, all run by NCEP; the GEM analysis run by the CMC; and the NOGAPS analysis run by Fleet Numerical Meteorology and Oceanography Center (FNMOC). See Grimit and Mass (2002) for details.

No.	Ensemble member	IC/LBC source
1	AVN-MM5	NCEP
2	GEM-MM5	CMC
3	ETA-MM5	NCEP
4	NGM-MM5	NCEP
5	NOGAPS-MM5	FNMOC

We applied the EMOS and EMOS$^+$ techniques to the University of Washington mesoscale short-range ensemble described by Grimit and Mass (2002). This is a multianalysis, single-model [fifth-generation Pennsylvania State University–National Center for Atmospheric Research Mesoscale Model (MM5)] ensemble driven by initial conditions and lateral boundary conditions obtained from major operational weather centers worldwide. Table 1 provides an overview of the phase I University of Washington ensemble system. Figure 1 illustrates the spread-skill relationship for sea level pressure forecasts, using the same period, January–June 2000, on which the study of Grimit and Mass (2002) was based. The ensemble spread provides useful information about the error of the ensemble mean forecast. Figure 2 gives an example of a 48-h EMOS forecast of sea level pressure. This forecast was initialized at 0000 UTC on 25 May 2000 and was valid at Hope Airport, British Columbia, Canada. Both the EMOS predictive PDF and the EMOS predictive CDF are shown. The construction of prediction intervals from the predictive CDF, say F, is straightforward. For instance, the 16⅔rd and 83⅓rd percentile of F form the lower and upper endpoints of the 66⅔% central prediction interval, respectively. In the Hope Airport example, and using the millibar as unit, this interval was [1007.3, 1011.9]. The ensemble range of the University of Washington ensemble was [1003.7, 1016.8]. For a five-member ensemble, this is also a nominal 66⅔% prediction interval, but it is much wider. Perhaps surprisingly,

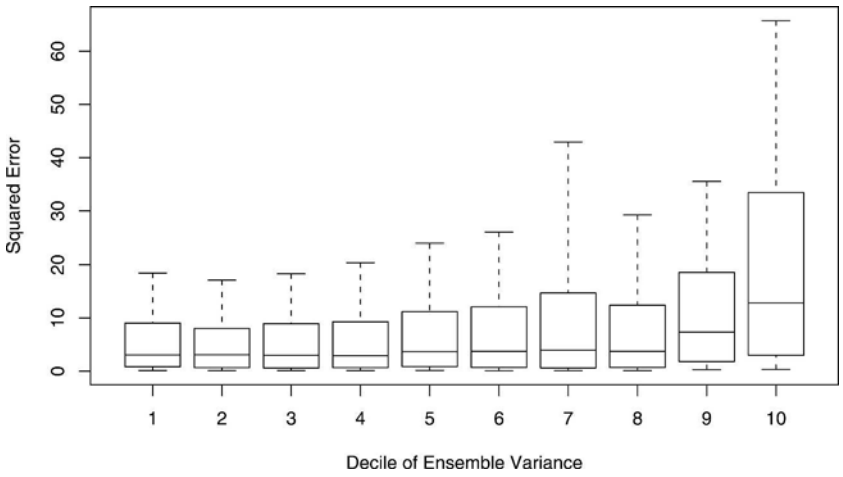

FIG. 1. Spread-skill relationship for ensemble mean forecasts of sea level pressure over the Pacific Northwest, Jan–Jun 2000. For each decile of the ensemble variance, the box plot shows the 10th, 25th, 50th, 75th, and 90th percentiles of the squared forecast error. The correlation between the ensemble variance and the squared forecast error was 0.33 for individual forecasts and 0.52 for daily averages aggregated across the Pacific Northwest.

this situation—EMOS prediction intervals that were shorter than their ensemble counterparts—was not uncommon. In our case study this occurred in about 28% of the sea level pressure forecasts.

The paper is organized as follows. In section 2 we describe the EMOS and EMOS$^+$ techniques in detail, and we explain how we go about verifying probabilistic forecasts. In assessing forecast PDFs, we are guided by the principle that probabilistic forecasts strive to maximize sharpness subject to calibration (Gneiting et al. 2003). We apply diagnostic tools, such as the verification rank histogram and the probability integral transform (PIT) histogram, as well as scoring rules, among them the continuous ranked probability score (CRPS) and the ignorance score. For estimating the EMOS and EMOS$^+$ coefficients, we introduce the novel approach of minimum CRPS estimation, which forms a particular case of minimum contrast estimation. Specifically, we find the coefficient values that minimize the continuous ranked probability score for the training data. In our experiments, this method gave better results than classical maximum likelihood estimation, which is nonrobust and tends to favor overdispersive forecast PDFs.

Section 3 provides a case study of EMOS and EMOS$^+$ forecasts for sea level pressure and surface temperature in spring 2000 over the Pacific Northwest, using the University of Washington ensemble. We explain how we find a suitable training period, and we describe and verify the EMOS and EMOS$^+$ forecasts. The EMOS and EMOS$^+$ forecast PDFs were much better calibrated than the raw ensemble or the bias-corrected ensemble, and the mean absolute error

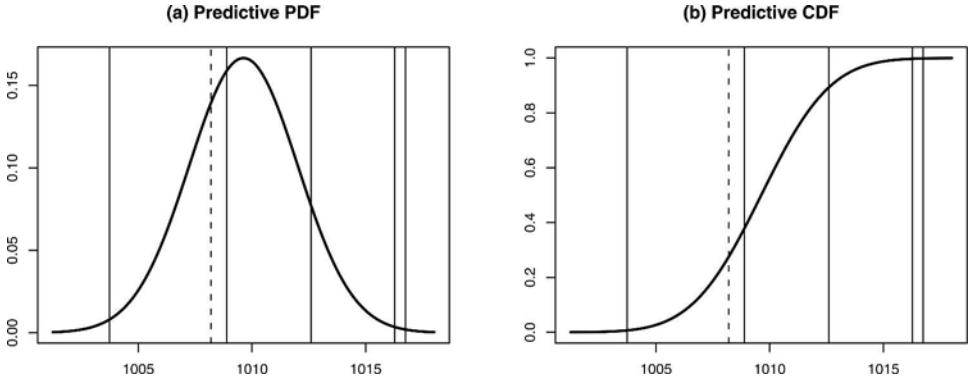

FIG. 2. EMOS 48-h forecast of sea level pressure at Hope Airport, British Columbia, initialized at 0000 UTC on 25 May 2000: (a) predictive PDF and (b) predictive CDF. Also shown are the five ensemble member forecasts (solid vertical lines) and the verifying observation (broken vertical line). The ensemble member forecasts are, from lowest to highest, the NOGAPS-MM5, AVN-MM5, GEM-MM5, ETA-MM5, and NGM-MM5 forecasts.

(MAE), root-mean-square error (RMSE), continuous ranked probability score (CRPS), and ignorance score (IGN) for the EMOS and EMOS$^+$ forecasts were consistently, and substantially, better than the corresponding quantities for the raw ensemble or the bias-corrected ensemble. The paper closes with a discussion in section 4.

2. Methods

We now explain our approach to verifying probabilistic forecasts and describe the EMOS technique in detail. For estimating the EMOS coefficients we use the novel approach of minimum CRPS estimation, which forms a special case of minimum contrast estimation (MCE). This method is best explained in terms of verification measures, so we describe these first.

a. Assessing sharpness and calibration

The goal of probabilistic forecasting is to maximize the sharpness of the forecast PDFs subject to calibration (Gneiting et al. 2003). Calibration refers to the statistical consistency between the forecast PDFs and the verifications and is a joint property of the predictions and the verifications. Briefly, a forecast technique is calibrated if meteorological events declared to have probability p occur a proportion p of the time on average. Sharpness refers to the spread of the forecast PDFs and is a property of the predictions only. A forecast technique is sharp if prediction intervals are shorter on average than prediction intervals derived from naive methods, such as climatology or persistence. The more concentrated the forecast PDF, the sharper the forecast, and the sharper the better, subject to calibration.

The principal tool for assessing the calibration of ensemble forecasts is the verification rank histogram or Talagrand diagram (Anderson 1996; Hamill and Colucci 1997; Talagrand et al. 1997; Hamill 2001). To obtain a verification rank histogram, find the rank of the verifying observation when pooled within the ordered ensemble values and plot the histogram of the ranks. Anderson (1996) and Saetra et al. (2004) suggested that when observations are used to assess an ensemble system, normally distributed noise based on the observation error could be added to the ensemble member models before the rank histogram is computed. In our experiments with the University of Washington system, the rank histograms computed from the ensembles with and without noise looked almost identical, and we retained the former.

The analogous tool for PDF forecasts is the PIT histogram. If F denotes the predictive CDF, the probability integral transform is simply the value $F(x)$ at the verification x, a number between 0 and 1. For the Hope Airport forecast in Fig. 2b, for instance, the PIT value was 0.28. Rosenblatt (1952) studied the probability integral transform, and Dawid (1984) proposed its use in the assessment of probabilistic forecasts. The PIT histogram—that is, the histogram of the PIT values—is a commonly used tool in the econometric literature on probabilistic forecasting (see, e.g., Weigend and Shi 2000). Its interpretation is the same as that of the verification rank histogram: calibrated probabilistic forecasts yield PIT histograms that are close to uniform, while underdispersive forecasts result in U-shaped PIT histograms.

How can ensembles and PDF forecasts be fairly compared? An ensemble provides a finite, typically small, number of values only, while PDF forecasts give continuous statements of uncertainty, so this seems difficult. There are two natural approaches to a fair comparison, using either the verification rank histogram or the PIT histogram. To obtain an m-member ensemble from a PDF forecast, take the CDF quantiles at levels $i/(m + 1)$, for $i = 1, \ldots, m$. The verification rank histogram can then be formed in the usual way. To obtain a PIT histogram from an ensemble, fit a PDF to each ensemble forecast, as proposed by Déqué et al. (1994), Wilson et al. (1999), and Grimit and Mass (2004). The standard ensemble smoothing approach of Grimit and Mass (2004) fits a normal distribution with mean equal to the ensemble mean and variance equal to the ensemble variance. The PIT value is then computed on the basis of the fitted Gaussian CDF. Wilks (2002) proposed to smooth forecast ensembles by fitting mixtures of Gaussian distributions, an approach that allows for multimodal forecast PDFs. Multimodality may indeed be an issue for larger ensembles. For smaller ensembles, such as the University of Washington ensemble, standard ensemble smoothing using a single normal PDF suffices.

In addition to showing verification rank histograms and PIT histograms, we report the coverage of the 66⅔% central prediction interval; we chose this interval because the range of a five-member ensemble provides such. Finally, to assess sharpness, we consider the average width of the 66⅔% prediction intervals. For a five-member ensemble, this is just the average ensemble range.

For Gaussian predictive PDFs, the average width of the $100 \times (1 - \alpha)$% central prediction intervals is

$$2z_{1-\alpha/2}\overline{S}, \tag{3}$$

where $z_{1-\alpha/2}$ denotes the $1 - \alpha/2$ quantile of the normal distribution with mean 0 and variance 1, and where $\overline{S}$ stands for the average standard deviation of the predictive PDFs. For instance, Table 2 shows that the average width of the central 66⅔% prediction intervals for MCE–EMOS forecasts of sea level pressure is 4.71. From (3) with $\alpha = ⅓$ we find that $\overline{S} = 2.44$. Using again (3), the average width of the 50% and 90% central prediction intervals is 3.29 and 8.01, respectively.

TABLE 2. Comparison of EMOS predictive PDFs obtained by maximum likelihood estimation (MLE-EMOS) and minimum CRPS estimation (MCE-EMOS). The results are for the test data, region, and 40-day sliding training period described in section 3.

	Score		Score		66⅔% prediction interval	
	MAE	RMSE	CRPS	IGN	Coverage	Average width
			Sea level pressure			
MLE-EMOS	1.969	2.489	1.394	2.327	68.09	4.921
MCE-EMOS	1.966	2.484	1.393	2.326	65.91	4.712
			Surface temperature			
MLE-EMOS	2.241	2.917	1.615	2.490	72.71	5.920
MCE-EMOS	2.231	2.907	1.606	2.487	68.58	5.427

b. Scoring rules

Scoring rules for the verification of deterministic-style forecasts are well known and have been widely used in forecast assessment. If μ_i denotes a deterministic-style forecast and y_i is the verification, the MAE is defined as

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - \mu_i|,$$

where the sum is taken over the test data. A related error measure is the mean-square error (MSE), defined by

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (y_i - \mu_i)^2.$$

The RMSE is the square root of the MSE and has the advantage of being recorded in the same unit as the verifications.

We also consider two scoring rules for the assessment of predictive PDFs: the continuous ranked probability score (Unger 1985; Hersbach 2000; Gneiting and Raftery 2004) and the ignorance score (Good 1952; Roulston and Smith 2002). These scoring rules are attractive in that they address calibration as well as sharpness.

The CRPS is the integral of the Brier scores at all possible threshold values t for the continuous predictand (Hersbach 2000; Toth et al. 2003, section 7.5.2). Specifically, if F is the predictive CDF and y is the verifying observation, the continuous ranked probability score is defined as

$$\text{crps}(F, y) = \int_{-\infty}^{\infty} [F(t) - H(t - y)]^2 \, \mathrm{d}t, \qquad (4)$$

where $H(t - y)$ denotes the Heaviside function and takes the value 0 when $t < y$ and the value 1 otherwise. Applications of the continuous ranked probability score have been hampered by a lack of closed-form expressions for the associated integral. However, when F is the CDF of a normal distribution with mean μ and variance σ^2, repeated partial integration in (4) shows that

$$\text{crps}[\mathcal{N}(\mu, \sigma^2), y] = \sigma\left\{\frac{y-\mu}{\sigma}\left[2\Phi\left(\frac{y-\mu}{\sigma}\right) - 1\right] + 2\varphi\left(\frac{y-\mu}{\sigma}\right) - \frac{1}{\sqrt{\pi}}\right\}, \qquad (5)$$

where $\varphi\left(\frac{y-\mu}{\sigma}\right)$ and $\Phi\left(\frac{y-\mu}{\sigma}\right)$ denote the PDF and the CDF, respectively, of the normal distribution with mean 0 and variance 1 evaluated at the normalized prediction error, $(y - \mu)/\sigma$. We note from (4) that the average score,

$$\text{CRPS} = \frac{1}{n}\sum_{i=1}^{n} \text{crps}(F_i, y_i), \qquad (6)$$

reduces to the MAE if each F_i is a deterministic-style forecast. For this and other reasons, the CRPS can be interpreted as a generalized version of the MAE (Gneiting and Raftery 2004).

The ignorance score is the negative of the logarithm of the predictive density f at the verifying value y, that is, for a single PDF forecast,

$$\text{ign}(f, y) = -\log f(y). \qquad (7)$$

In the case of a normal predictive PDF with mean μ and variance σ^2, we have

$$\text{ign}[\mathcal{N}(\mu, \sigma^2), y] = \frac{1}{2}\ln(2\pi\sigma^2) + \frac{(y-\mu)^2}{2\sigma^2} \qquad (8)$$

and the average ignorance is

$$\text{IGN} = \frac{1}{n}\sum_{i=1}^{n} \text{ign}(F_i, y_i) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{2}\ln(2\pi\sigma_i^2) + \frac{(y_i-\mu_i)^2}{2\sigma_i^2}\right]. \qquad (9)$$

When interpreting improvements in the IGN score, it is absolute rather than relative changes that are relevant (Roulston and Smith 2002).

Both CRPS and IGN are negatively oriented scores, in that a smaller value is better, and both scores are proper, meaning that they reward honest assessments. However, a key difference between the ignorance score and the continuous ranked probability score is that (5) grows linearly in the normalized prediction error, $z = (y - \mu)/\sigma$, while (8) grows quadratically in z. Hence, the ignorance score assigns harsh penalties to particularly poor probabilistic forecasts, and can be exceedingly sensitive to outliers and extreme events (Weigend and Shi 2000; Gneiting and Raftery 2004). (This will become apparent in Tables 5 and 7 below.) We report both scores, but in view of the lack of robustness of the ignorance score, we prefer the continuous ranked probability score. A more detailed discussion of scoring rules is given in Gneiting and Raftery (2004).

c. *Ensemble model output statistics and minimum CRPS estimation*

We now describe the standard version of the EMOS method. Suppose that $X_1, \ldots, X_m$ denotes an ensemble of forecasts for a univariate weather quantity Y, and that S^2 is the ensemble variance. The EMOS predictive PDF is that of the normal distribution

$$\mathcal{N}(a + b_1X_1 + \cdots + b_mX_m, c + dS^2). \qquad (10)$$

The EMOS predictive mean, $a + b_1X_1 + \cdots + b_mX_m$, is a bias-corrected weighted average of the ensemble member forecasts; it provides a deterministic-style forecast. The EMOS predictive variance, $c + dS^2$, is a linear function of the ensemble variance. The regression coefficients $b_1, \ldots, b_m$ in (10) reflect the overall performance of the ensemble member models over the training period, relative to the other members. They tend to be ordered in the inverse order of the forecast RMSEs, but this is not a direct relationship; they also reflect the correlations between the ensemble member forecasts. The variance coefficients c and d can be interpreted in terms of the ensemble spread and the performance of the ensemble mean forecast. All else being equal, larger values of the coefficient d suggest a more pronounced spread-error relationship. If spread and error are independent of each other, the coefficient d will be estimated as negligibly small. Hence, EMOS is robust, in the sense that it adapts to the presence as well as to the absence of significant spread-error correlation.

A classical technique for estimating the coefficients a, $b_1, \ldots, b_m$, c, and d from training data is maximum likelihood (Wilks 1995, section 4.7). The likelihood function is defined as the probability of the training data given the coefficients, viewed as a function of the coefficients. In practice, it is more convenient to maximize the logarithm of the likelihood function, for reasons of both algebraic simplicity and numerical stability. The log-likelihood function for the statistical model (10) is

$$\begin{aligned}\ell(a; b_1, \ldots, b_m; c; d) &= -\frac{1}{2}\left\{k \log(2\pi) \right. \\ &\quad + \sum_{i=1}^{k} \frac{[Y_i - (a + b_1X_{i1} + \cdots + b_mX_{im})]^2}{c + dS_i^2} \\ &\quad \left. + \sum_{i=1}^{k} \log(c + dS_i^2)\right\}, \qquad (11)\end{aligned}$$

where the sum is taken over the training data. Here $X_{i1}, \ldots, X_{im}$ denote the ith ensemble forecast in the training set, S_i^2 denotes its variance, and Y_i denotes the ith verification. Strictly speaking, (11) is the log-likelihood function under the assumption of independence. Note that the log-likelihood (11) is essentially the negative of the ignorance score (9) but is applied to the training data rather than the test data. Hence, maximum likelihood estimation is equivalent to minimizing the ignorance score for the training data.

This observation suggests a general estimation strategy: pick a scoring rule that is relevant to the problem at hand, express the score for the training data as a function of the coefficients, and optimize that function with respect to the coefficient values. We take scoring rules to be negatively oriented, so a smaller value is better, and we minimize the training score. For positively oriented scoring rules, we would maximize the training score. Such an approach is formally equivalent to MCE, a technique that has been studied in the theoretical statistics literature (Pfanzagl 1969; Birgé and Massart 1993). The minimum score approach can also be interpreted within the framework of robust M estimation (Huber 1964; Huber 1981, section 3.2) and forms a special case thereof, in that the function to be optimized derives from a strictly proper scoring rule (Gneiting and Raftery 2004). A more detailed methodological and theoretical discussion is beyond the scope of this paper.

However, we compared EMOS PDF forecasts estimated by MCE with the continuous ranked probability score, as described below, to EMOS PDF forecasts estimated by maximum likelihood. The MCE-EMOS approach clearly performed better: the predictive PDFs were sharper, and they were better calibrated. This comparison is summarized in Table 2. As a rule of thumb, it seems that predictive PDFs estimated by maximum likelihood tend to be overdispersive, resulting in unnecessarily wide prediction intervals that have higher than nominal coverage, and in inverted U-shaped PIT histograms. This latter shape is also seen in Figs. 4 and 5 of Weigend and Shi (2000), who estimate predictive densities by the maximum likelihood method in the form of the expectation maximization (EM) algorithm.

We argued in section 2b that the CRPS is a more

TABLE 3. Minimum CRPS estimation of the EMOS$^+$ coefficients for the Hope Airport forecast PDF in Fig. 2. The regression coefficients $b_1, \ldots, b_5$ correspond to the AVN-MM5, GEM-MM5, ETA-MM5, NGM-MM5, and NOGAPS-MM5 forecasts, respectively.

	a	b_1	b_2	b_3	b_4	b_5	c	d
EMOS coefficients	135.61	0.38	0.36	−0.16	0.01	0.29	5.74	0.00
1st Iteration	143.23	0.35	0.33	0.00	−0.09	0.28	5.81	0.00
EMOS$^+$ coefficients	130.34	0.31	0.31	0.00	0.00	0.25	5.88	0.00

robust and therefore more appropriate scoring rule than the ignorance score. This suggests the use of the continuous ranked probability score in minimum contrast estimation; we call this minimum CRPS estimation. The minimum CRPS estimator finds the coefficients $a, b_1, \ldots, b_m, c$, and d in the statistical model (10) that minimize the CRPS value for the training data. Using (5) and (6), we express the training CRPS as an analytic function of the coefficients, namely

$$\Gamma(a; b_1, \ldots, b_m; c; d) = \frac{1}{k}\sum_{i=1}^{k}(c + dS_i^2)^{1/2}\left\{Z_i[2\Phi(Z_i) - 1] + 2\varphi(Z_i) - \frac{1}{\sqrt{\pi}}\right\}, \quad (12)$$

where

$$Z_i = \frac{Y_i - (a + b_1X_{i1} + \cdots + b_mX_{im})}{(c + dS_i^2)^{1/2}}$$

is a standardized forecast error, and where φ and Φ denote the PDF and the CDF, respectively, of a normal distribution with mean 0 and variance 1. We find the coefficient values that minimize (12) numerically, using the Broyden–Fletcher–Goldfarb–Shanno algorithm (Press et al. 1992, section 10.7) as implemented in the R language and environment (www.cran.r-project.org). The variance coefficients c and d are constrained to be nonnegative, which is not an issue for the parameter c. To enforce the nonnegativity of the variance coefficient d, we set $d = \delta^2$ and optimize over δ. The optimization algorithm requires initial values, and starting values based on past experience usually give good solutions. However, convergence to a global maximum cannot be guaranteed, and the solution reached can be sensitive to the initial values.

The upper row in Table 3 shows the estimated EMOS coefficients for predictions on 25 May 2000, the day on which the Hope Airport forecast in Fig. 2 was issued. The training set included the forecasts and observations from the most recent 40 days that were available on 25 May 2000. The EMOS weights, b_k, reflect the overall performance of the ensemble member models over the training period, relative to the other members. They tend to be ordered in the inverse order of the forecast RMSEs, but this is not a direct relationship; they also reflect the correlations between the ensemble member forecasts, and bias effects are taken account of, too. In this example, the Aviation (AVN)-MM5, Global Environmental Multiscale Model (GEM)-MM5, and the U.S. Navy Operational Global Atmospheric Prediction System (NOGAPS)-MM5 forecasts received the highest EMOS weights and, indeed, the (linearly) bias-corrected AVN-MM5, GEM-MM5, and NOGAPS-MM5 forecasts had a smaller training RMSE than the bias-corrected Eta Model (ETA)-MM5 and Nested-Grid Model (NGM)-MM5 forecasts. The AVN-MM5, ETA-MM5, and NGM-MM5 forecasts all used NCEP initializations, and they were highly collinear. For the training period, the pairwise correlations within this group ranged from 0.93 to 0.97. EMOS assigned a substantial weight to the most skillful of these three collinear forecasts, the AVN-MM5 forecast; the ETA-MM5 and NGM-MM5 forecasts received negative and negligible weights, respectively. The correlation coefficients between NCEP- and non-NCEP-initialized member model forecasts were also high, but they reached at most 0.92. The estimated variance coefficient d turned out to be negligibly small, thereby suggesting a weak spread-skill relationship during the training period. Indeed, the correlation between the ensemble variance and the squared error of the ensemble mean forecast was only 0.11 for this particular 40-day training period, as compared to 0.33 for the entire period, January–June 2000.

It is straightforward to draw random samples from the EMOS predictive PDF. An alternative, and likely preferable, way of forming an m-member ensemble from the predictive PDF is by taking the CDF quantiles at level $i/(m + 1)$, for $i = 1, \ldots, m$, respectively. In this way, ensembles of any size can be obtained, and in this sense, EMOS can be viewed as a dressing method (Roulston and Smith 2002). We note that EMOS requires the ensemble member models to have individually distinguishable characteristics. This is true for the University of Washington ensemble, a multianalysis, mesoscale, short-range ensemble, and also for poor person's and multimodel ensembles. In other types of ensembles, the member models might be exchangeable. In this type of situation, the linear regression Eq. (1) can be based on the ensemble mean forecast only, which constrains the regression coefficients $b_1 = \cdots = b_m$ to be equal. In this implementation, EMOS can be applied to essentially all ensemble systems, including perturbed observations, bred, or singular vector ensembles. Jewson et al. (2004) applied such an approach to the synoptic ECMWF ensemble, using maximum likelihood estimation. However, they did not report out of sample

forecasts, and consequently neither verification scores nor rank histograms. For the University of Washington ensemble, the general formulation seems preferable. In the situation of Table 2, constraining the regression coefficients in (10) to be equal, that is, using the ensemble mean only, results in MAE, RMSE, CRPS, and IGN scores up to 7% worse, as compared to the full formulation.

d. *Enforcing nonnegative regression coefficients: EMOS$^+$*

The regression Eq. (1) allows for negative EMOS weights $b_1, \ldots, b_m$, and in many applications of multiple linear regression negative regression coefficients are crucial. Consider, for instance, predictions of wind speed based on time series of past values. Recent changes in wind speed may have predictive power, and to take account of the changes, negative regression coefficients are required. In the context of ensemble forecasts, the negative weights seem much harder to interpret. They imply, all else being equal, that sea level pressure, say, is predicted to be lower when the forecast with the negative weight is higher. In our experiments with the EMOS technique, negative regression coefficients occurred frequently, and they were caused by collinearities between the ensemble member model forecasts. Indeed, it is well known that highly correlated predictor variables in a regression model lead to coefficient estimates that are unstable under small changes in the training data. In this sense, the negative EMOS weights could be viewed as an artifact.

To address this issue, we propose an implementation of the EMOS technique that constrains the coefficients $b_1, \ldots, b_m$ in the regression Eq. (1) to be nonnegative. We call this variant of the EMOS technique EMOS$^+$. The plus sign stands for the nonnegative EMOS weights, and the idea relates to Tibshirani's (1996) least absolute shrinkage and selection operator (lasso) regression, which imposes a penalty on the absolute size of the regression coefficients and results in estimates that tend to be exactly zero for some of the coefficients, thereby improving forecast accuracy and interpretability. In our case, EMOS weights that are exactly zero can be interpreted in terms of reduced ensembles. To fit the EMOS$^+$ model we proceed stepwise, as follows. We first find the unconstrained minimum of the CRPS value (12), that is, we estimate the coefficients of the standard EMOS model. If all estimated regression coefficients are nonnegative, the EMOS$^+$ estimates are the same as the EMOS estimates, and the EMOS$^+$ model is complete. If one or more of the regression coefficients are negative, we set them to zero and minimize the CRPS value (12) under this constraint. We also recompute the ensemble variance, using only the ensemble members that remain in the regression equation, and subsequently use the recomputed ensemble spread. This procedure is iterated until all estimated regression coefficients are nonnegative.

Table 3 illustrates this algorithm for predictions on 25 May 2000, the day on which the Hope Airport forecast in Fig. 2 was issued. The upper row shows the parameter estimates for the standard EMOS model, which include a negative coefficient for the ETA-MM5 forecast. We set this coefficient to zero and proceed with the constrained minimization, resulting in a negative weight for the NGM-MM5 forecast. The final EMOS$^+$ equation uses only one of the three ensemble members initialized with NCEP models, namely the AVN-MM5 forecast, along with the GEM-MM5 and the NOGAPS-MM5 forecasts.

In our experiments with the University of Washington ensemble, which are summarized below, the EMOS and EMOS$^+$ techniques had equal forecast skill. However, we found the EMOS$^+$ parameters to be easier to interpret, in that the EMOS$^+$ forecasts correspond to reduced ensembles. The member models with vanishing EMOS$^+$ coefficients were judged not to be useful during the training period, relative to all the others, and therefore were removed from the regression equation. That said, it is important to distinguish the usefulness and the skill of an ensemble member model. Consider, for instance, a three-member ensemble. Model A has a lower RMSE than model B, and model B has a lower RMSE than model C; all three models are unbiased. If A and B are highly collinear and both are independent of C, then model C may be a more useful but less skillful ensemble member than model B.

3. Results for the University of Washington ensemble over the Pacific Northwest

We now give the results of applying the EMOS and EMOS$^+$ techniques to 48-h forecasts of sea level pressure and surface temperature over the northwestern United States and British Columbia, using phase I of the University of Washington ensemble described by Grimit and Mass (2002). The University of Washington ensemble system is a mesoscale, short-range ensemble based on MM5 and forms an integral part of the Pacific Northwest regional environmental prediction effort (Mass et al. 2003). The ensemble system was in operation on 102 days between 12 January 2000 and 30 June 2000 using initializations at 0000 UTC; it is described in Table 1. During this period, there were 16 015 and 56 489 verifying observations of sea level pressure and surface temperature, respectively. Model forecast data at the four grid points surrounding each observation were bilinearly interpolated to the observation site (Grimit and Mass 2002). When we talk of a 40-day training period, say, we refer to a sliding training period that consists of the 40 most recent days prior to the forecast for which ensemble output and verifying observations were available. In terms of calendar days, this period typically corresponds to more than 40 days.

a. Length of training period

What training period should be used for estimating the EMOS and EMOS$^+$ regression coefficients and variance parameters? There is a trade-off in selecting the length of the sliding training period. Shorter training periods can adapt rapidly to seasonally varying model biases, changes in the performance of the ensemble member models, and changes in environmental conditions. On the other hand, longer training periods reduce the statistical variability in the estimation of the EMOS and EMOS$^+$ coefficients. We considered sliding training periods of 19, 20, . . . , 62 days for EMOS forecasts of sea level pressure; the results for the EMOS$^+$ forecasts were similar. For comparability, the same test set was used in assessing all the training periods; that is, the first 63 days on which the ensemble was operating were not included in the test period. This results in a 39-day test period, consisting of all the days between 24 April and 30 June 2000 on which the ensemble system was operational. The unit used for the sea level pressure forecasts is the millibar (mb).

The results of this experiment are summarized in Fig. 3. Figures 3a and 3b show the MAE and RMSE of the deterministic-style EMOS forecasts, respectively. These decrease sharply for training periods less than 30 days, stay about constant for training periods between 30 and 45 days, and increase thereafter. Figures 3c and 3d show the CRPS and the IGN. The patterns are similar to those for the MAE and the RMSE. The empirical coverage of the EMOS 66⅔% prediction intervals is shown in Fig. 3e along with the nominal coverage, which is represented by the horizontal line. Training periods under 30 days seem to result in underdispersive PDFs, but training periods between 30 and 60 days show close to nominal coverage. Figure 3f shows the average width of the 66⅔% prediction intervals. The average width increases with the length of the training period, but is about constant for training periods between 30 and 40 days.

To summarize these results, there appear to be substantial gains in increasing the training period beyond 30 days. As the training period increases beyond 45 days, the skill of the probabilistic forecasts declines slowly but steadily, presumably as a result of seasonally varying model biases. In view of our goal of maximizing sharpness subject to calibration, we chose a sliding 40-day training period for the EMOS and EMOS$^+$ forecasts of sea level pressure. For instance, the predictive PDFs for forecasts initialized on 30 June 2000 were trained on the 40 most recent ensemble runs that had verified by this date. The earliest forecasts in this particular training set were initialized on 15 April 2000 and verified on 17 April 2000; the latest were initialized on 28 June 2000 and verified on 30 June 2000. The sliding 40-day training period worked well for temperature forecasts, too. However, distinct training periods might work best for distinct variables, forecast horizons, time periods, and regions. Ideally, we would include training data from previous years to address seasonal effects. Further research in this direction is desirable as multiyear runs of stable mesoscale ensembles become available. Ensemble reforecasting (Hamill et al. 2004) provides an attractive yet, in many cases, computationally prohibitive alternative.

b. Sea level pressure forecasts

We now give the results for EMOS and EMOS$^+$ forecasts of sea level pressure, using a 40-day sliding training period and the same test set that was used to compare the different training periods. We also summarize the results for the bias-corrected ensemble member forecasts and for a climatological forecast. The bias-corrected ensemble member forecasts used least squares regression fits, $\alpha_k + \beta_k X_k$, of the ensemble member models, X_k, on the observations, and the regression parameters were estimated on the same 40-day sliding training period. The deterministic-style climatological forecast was the average sea level pressure among the verifying observations in the training set, and the climatological predictive PDF was obtained by fitting a normal PDF to these observations.

Figure 4 shows the estimates of the EMOS coefficients for the 39 days in the test period. The estimated intercept in the multiple linear regression equation is shown in Fig. 4a. Figures 4b–f show the EMOS weights for the five ensemble member models, respectively. The weights for the AVN-MM5, CMC-MM5, and NOGAPS-MM5 forecasts were consistently positive and substantial, the weights for the ETA-MM5 forecast were consistently negative, and the weights for the NGM-MM5 forecast decreased from positive to negative. The negative regression coefficients are caused by collinearities between the ensemble member model forecasts, as discussed in sections 2c and 2d. Figures 4g and 4h show the estimated variance coefficients c and d, respectively. The estimates of the variance parameter c decreased toward the end of the test period, thereby indicating improved ensemble skill or improved atmospheric predictability, or both. The estimated values of d were mostly negligible.

The corresponding estimates of the EMOS$^+$ coefficients are shown in Figs. 5a–h, respectively. The EMOS$^+$ weights for the AVN-MM5, CMC-MM5, and NOGAPS-MM5 forecasts were consistently substantial, and the weights for the ETA-MM5 and NGM-MM5 forecasts were consistently zero. Hence, EMOS$^+$ retained only one of the three highly collinear ensemble member models that were initialized by NCEP analyses and picked the most skillful of them, namely the AVN-MM5 forecast. The EMOS$^+$ weights for this forecast were consistently higher than the respective EMOS weights. The EMOS$^+$ estimates of the variance parameter c decreased during the test period, and the estimates of the variance coefficient d, shown in Fig. 5h, were mostly nonzero. The increase toward the end of

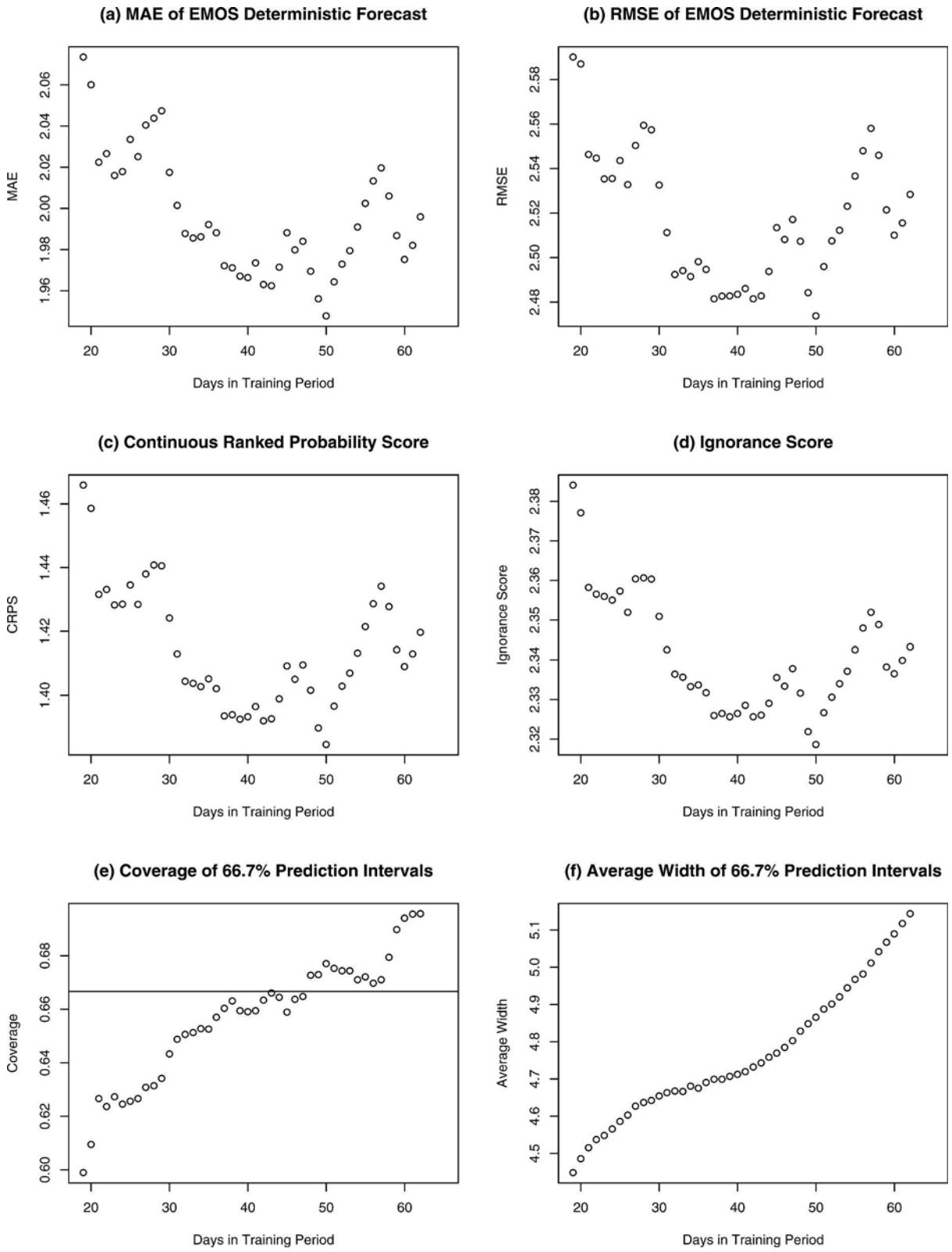

FIG. 3. Comparison of training period lengths for forecasts of sea level pressure over the Pacific Northwest: (a) MAE of EMOS deterministic-style forecasts; (b) RMSE of EMOS deterministic-style forecasts; (c) continuous ranked probability score; (d) ignorance score; (e) coverage of 66⅔% prediction intervals; and (f) average width of 66⅔% prediction intervals.

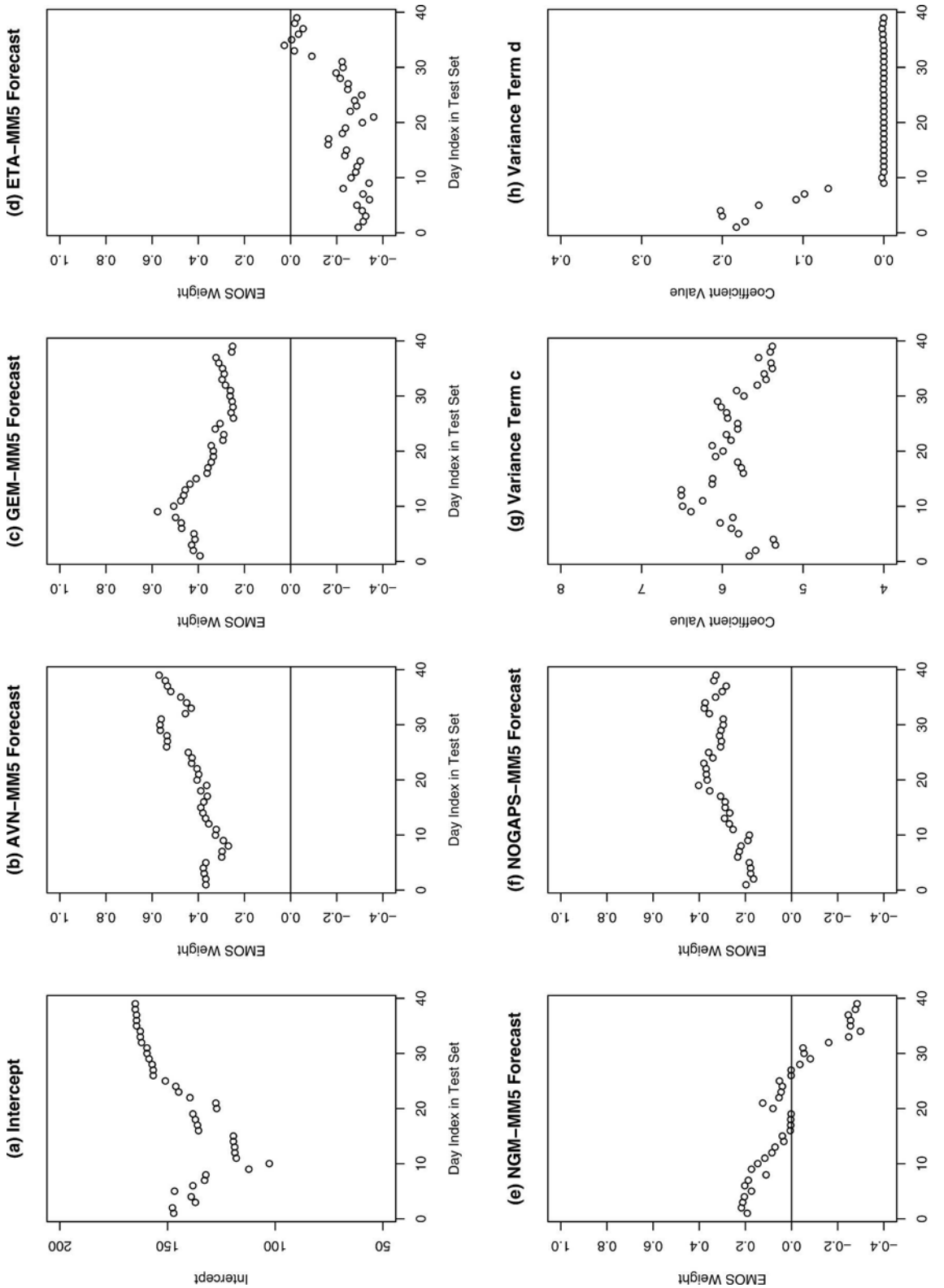

FIG. 4. Coefficient estimates for EMOS forecasts of sea level pressure over the Pacific Northwest, for each of the 39 days in the test period: (a) intercept, (b)–(f) member model weights, and (g) and (h) variance terms c and d.

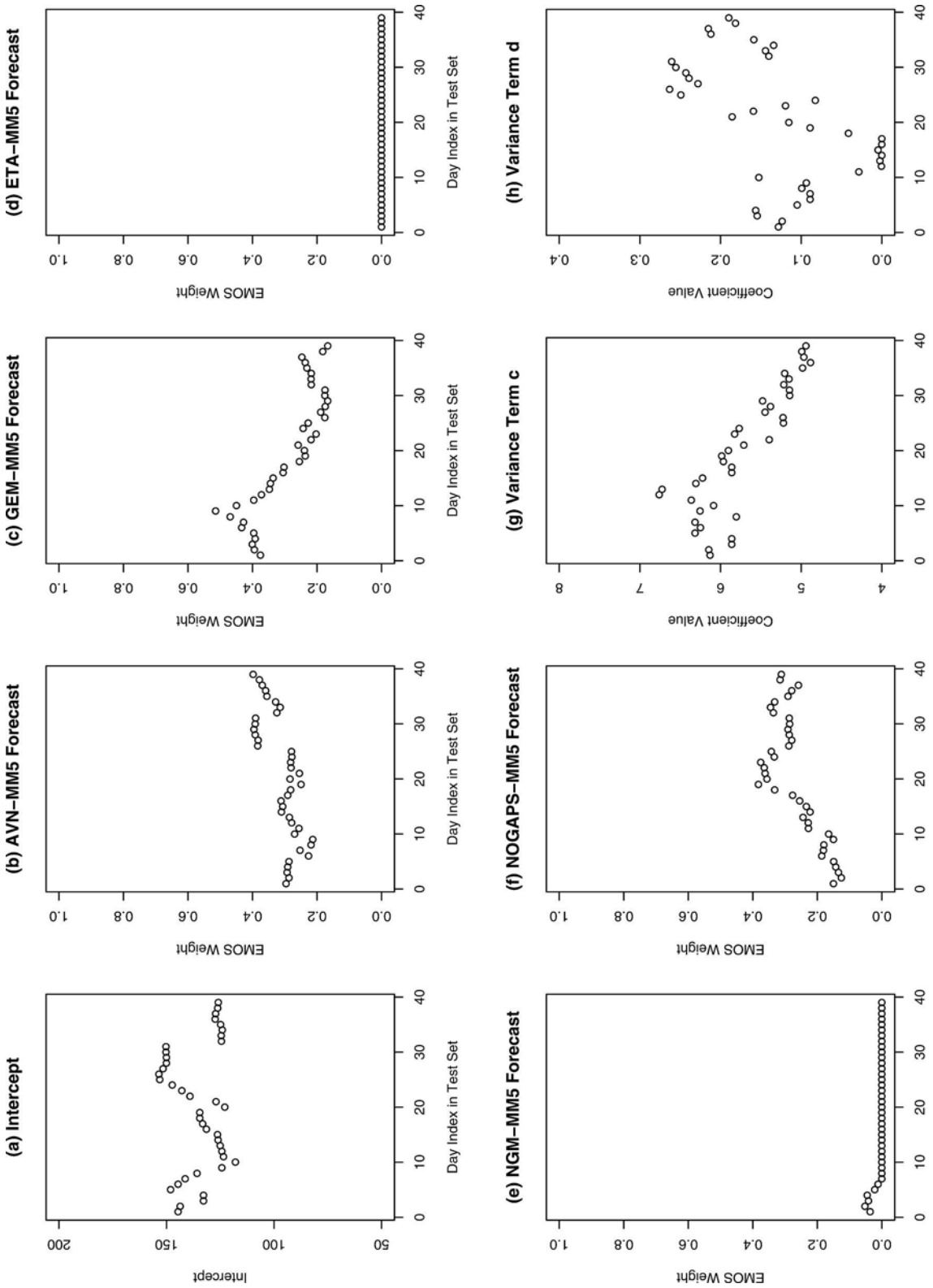

FIG. 5. Coefficient estimates for EMOS$^+$ forecasts of sea level pressure over the Pacific Northwest, for each of the 39 days in the test period: (a) intercept, (b)–(f) member model weights, and (g) and (h) variance terms c and d.

TABLE 4. Comparison of deterministic-style forecasts of sea level pressure over the Pacific Northwest. The climatological, bias-corrected, EMOS, and EMOS$^+$ forecasts were trained on a sliding 40-day period.

	MAE	RMSE
Climatological forecast	4.72	5.83
AVN-MM5	2.20	2.90
GEM-MM5	2.35	3.00
ETA-MM5	2.50	3.25
NGM-MM5	2.70	3.40
NOGAPS-MM5	2.50	3.21
AVN-MM5 bias-corrected	2.10	2.68
GEM-MM5 bias-corrected	2.24	2.88
ETA-MM5 bias-corrected	2.37	3.14
NGM-MM5 bias-corrected	2.48	3.23
NOGAPS-MM5 bias-corrected	2.10	2.66
Mean of raw ensemble	2.11	2.73
Mean of bias-corrected ensemble	2.08	2.69
EMOS forecast	1.97	2.48
EMOS$^+$ forecast	1.95	2.49

the test period indicates a strengthening of the spread-error relationship or an increase in the temporal variability of the forecast skill, or both. The comparison to the EMOS estimates of the variance coefficient d, shown in Fig. 4h, seems interesting; it suggests that the spread-error relationship can be masked by collinearities between the ensemble member models.

Table 4 provides summary measures of deterministic-style forecast accuracy. Among the raw ensemble member models, the AVN-MM5 forecast performed best. Bias correction resulted in a reduction of the RSME for the ensemble member model forecasts between 4% and 17%. The improvement was most pronounced for the NOGAPS-MM5 forecast. The AVN-MM5, CMC-MM5, and NOGAPS-MM5 forecasts were more accurate than the ETA-MM5 and NGM-MM5 forecasts. The ensemble mean forecast performed considerably better than any of the ensemble member models, but the bias-corrected AVN-MM5 and NOGAPS-MM5 forecasts were more accurate than the mean of the bias-corrected ensemble. The deterministic-style EMOS and EMOS$^+$ forecasts performed about equally well and were much better than any of the other forecasts. They had RMSEs between 7% and 9% less when compared to the mean of the raw ensemble and to the mean of the bias-corrected ensemble, respectively. The results in terms of the MAE were similar.

Table 5 turns to summary measures of probabilistic forecast skill. The climatological predictive PDFs showed the correct coverage, but they were too spread out to be competitive. The bias-corrected ensemble showed reduced ensemble spread, but was even more underdispersive than the raw ensemble. The EMOS and EMOS$^+$ prediction intervals showed accurate coverage. The CRPS and the IGN were computed as described in section 2b, using standard ensemble smoothing for the raw and bias-corrected ensemble. The CRPS can also be computed directly, by using the empirical

TABLE 5. Comparison of predictive PDFs for sea level pressure over the Pacific Northwest. The bias-corrected ensemble, the EMOS, and the EMOS$^+$ forecasts were trained on a sliding 40-day period.

	66⅔% prediction interval		Score	
	Coverage	Average width	CRPS	IGN
Climatological forecast	67.0	11.83	3.32	3.19
Raw ensemble	53.9	3.93	1.61	4.84
Bias-corrected ensemble	40.7	2.77	1.66	6.01
EMOS forecast	65.9	4.71	1.39	2.33
EMOS$^+$ forecast	67.6	4.75	1.39	2.33

ensemble CDF, which takes the values 0, ⅕, . . . , ⅘, 1, with jumps at the ensemble member forecasts. This gave somewhat higher CRPS values of 1.69 and 1.72 for the raw ensemble and for the bias-corrected ensemble, respectively. The EMOS and EMOS$^+$ predictive PDFs performed equally well and had by far the best scores among the forecasts. When compared to the bias-corrected ensemble, EMOS and EMOS$^+$ reduced the CRPS score by 16%, and the IGN score was 3.68 points lower. The EMOS and EMOS$^+$ prediction intervals were not much wider than the prediction intervals obtained from the raw ensemble. A more detailed analysis shows, perhaps surprisingly, that in 28% of the forecasts the EMOS 66⅔% prediction interval was shorter than the range of the five-member raw ensemble. In 10% of the forecasts, the EMOS 66⅔% prediction interval was shorter than the range of the bias-corrected ensemble.

The verification rank histograms for the various ensembles are shown in Fig. 6. The EMOS and the EMOS$^+$ ensemble were much better calibrated than the raw ensemble or the bias-corrected ensemble and showed rank histograms that were close to being uniform but not quite uniform. Indeed, the latter was not to be expected. Sea level pressure is a synoptic variable with strong spatial correlation throughout the ensemble domain, and there were only 39 days in the evaluation period. The PIT histograms in Fig. 7 accentuate the underdispersion in the raw ensemble and in the bias-corrected ensemble.

c. *Temperature forecasts*

We now summarize the results for forecasts of surface temperature, a case of primary interest to the public (Murphy and Winkler 1979). The 2-m temperature forecasts were obtained as an average of the predicted lowest sigma level temperature and the predicted ground temperature. Similar to the sea level pressure forecasts, we used a sliding 40-day training period, and we considered the same region and the same test period. We omit the results for the climatological forecast, which was even less competitive than for sea level pressure, given seasonal and topographic effects. The unit used for the temperature forecasts is degrees kelvin.

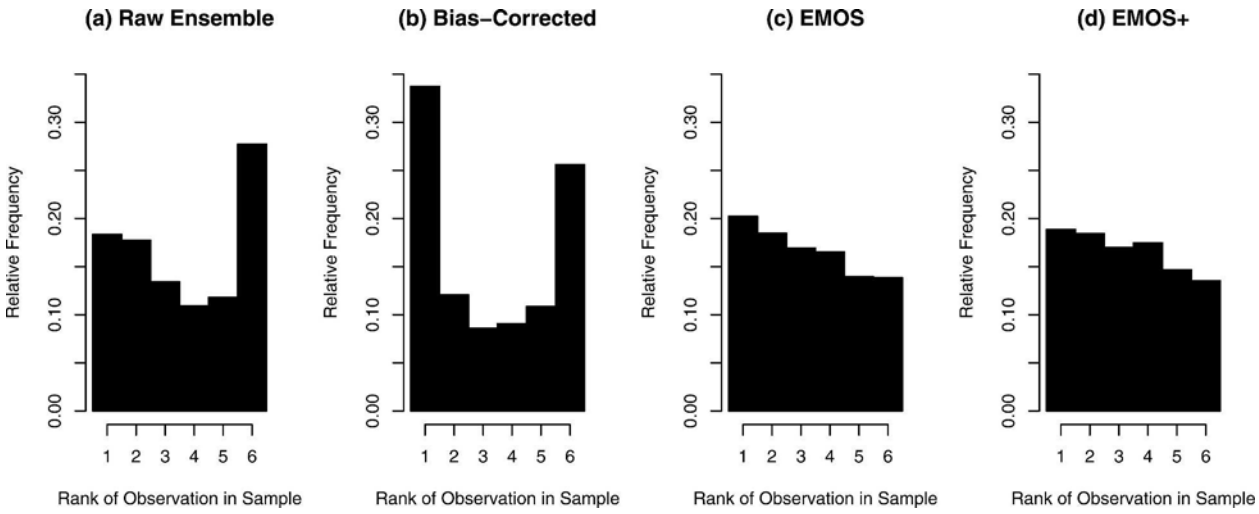

FIG. 6. Verification rank histograms for ensemble forecasts of sea level pressure over the Pacific Northwest: (a) raw ensemble, (b) bias-corrected ensemble, (c) EMOS ensemble, and (d) EMOS$^+$ ensemble.

Figure 8 displays the estimates of the EMOS coefficients for the 39-day test period. Figure 8a shows the estimates of the intercepts, which were consistently negative. Figures 8b–f show the estimated EMOS weights. The weights for the AVN-MM5 forecast reached a maximum of 0.57 and were consistently the highest among the five ensemble member models. The weights for the ETA-MM5 and NOGAPS-MM5 forecasts were smaller but still positive and substantial, those for the NGM-MM5 forecast oscillated about zero, and those for the GEM-MM5 forecast were initially negative, before increasing to substantially positive levels. Figures 8g and 8h show the estimated variance parameters c and d.

Figures 9a–h turn to the corresponding estimates of the EMOS$^+$ coefficients. These were very similar to the EMOS estimates, except that the weights for the NGM-MM5 forecast and, initially, for the GEM-MM5 forecast, were zero. These results can, again, be interpreted in terms of the relative contributions of the ensemble member models. The correlation between the ETA-MM5 and the NGM-MM5 forecasts was the highest among the forecast pairs. To avoid collinearity, EMOS$^+$ retained only one of them. The AVN-MM5 forecast was the most accurate member model and received the highest EMOS and EMOS$^+$ weights.

Table 6 confirms that the AVN-MM5 forecast had the lowest RMSE among the ensemble member forecasts, both before and after bias correction. Bias correction resulted in percentage improvements in the

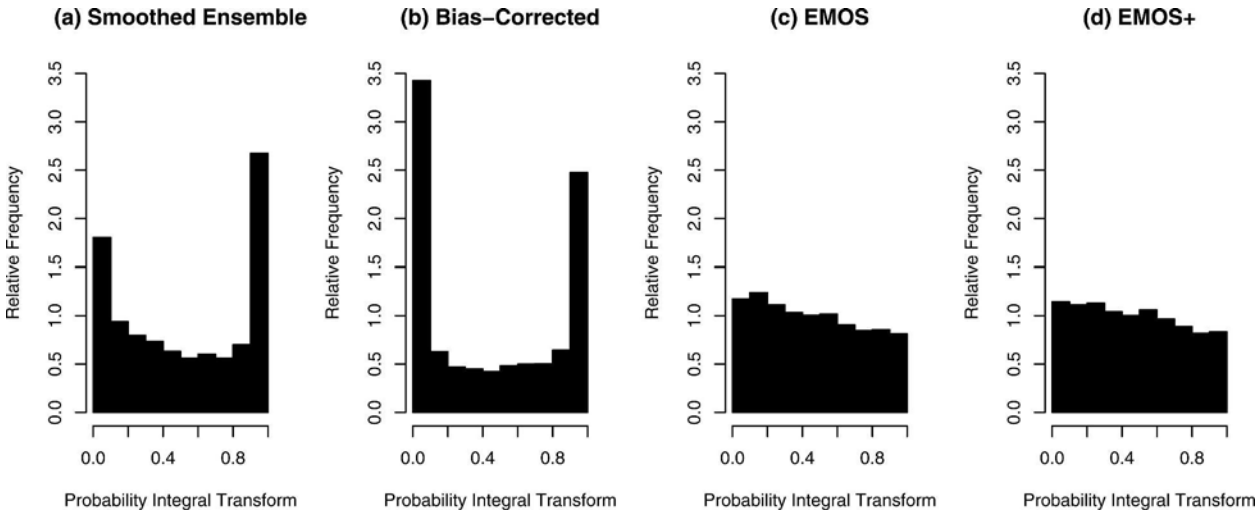

FIG. 7. PIT histograms for PDF forecasts of sea level pressure over the Pacific Northwest: (a) smoothed ensemble forecast, (b) smoothed bias-corrected ensemble forecast, (c) EMOS forecast, and (d) EMOS$^+$ forecast.

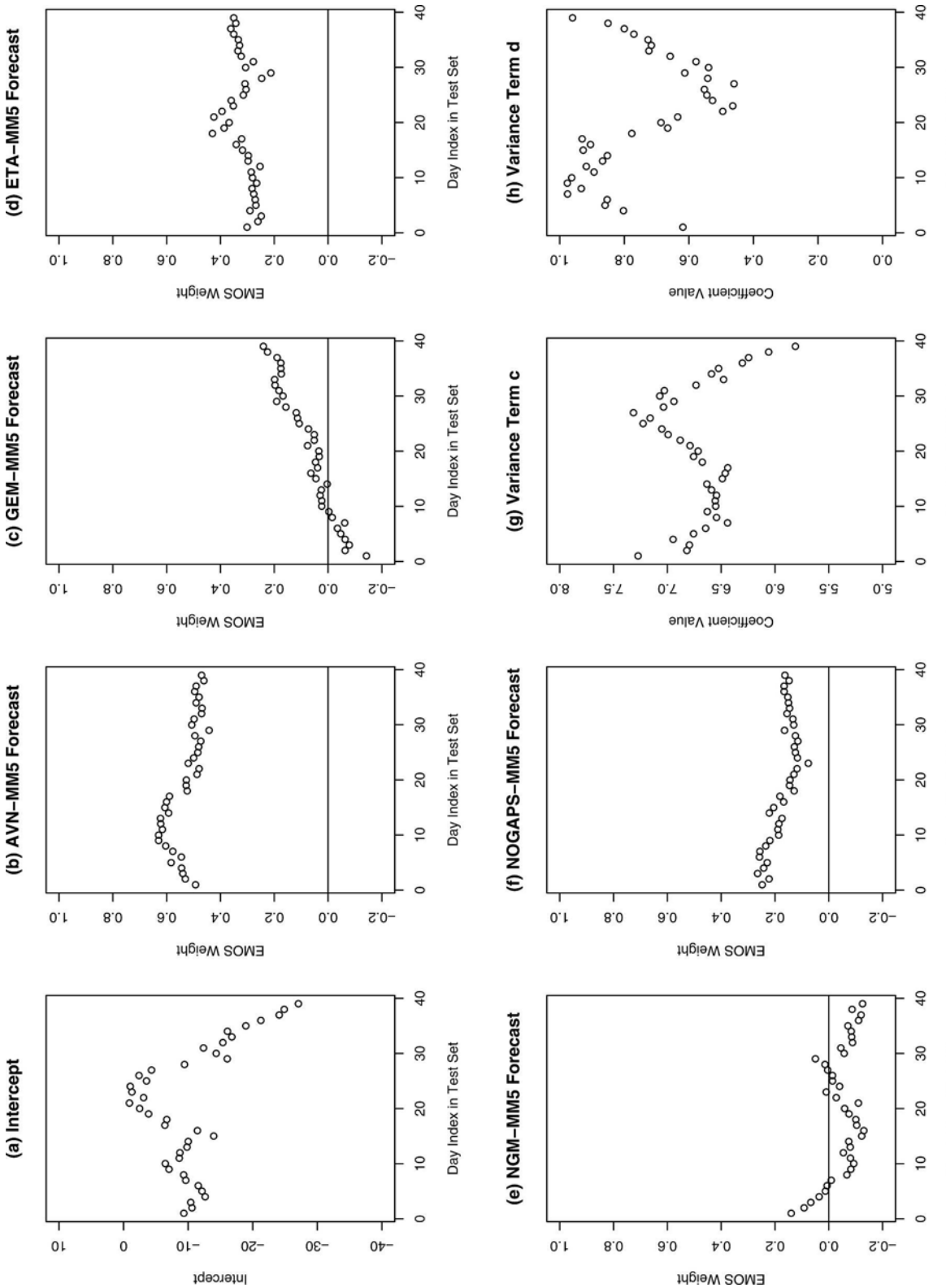

FIG. 8. Coefficient estimates for EMOS forecasts of surface temperature over the Pacific Northwest, for each of the 39 days in the test period: (a) intercept, (b)–(f) member model weights, (g) and (h) variance terms c and d.

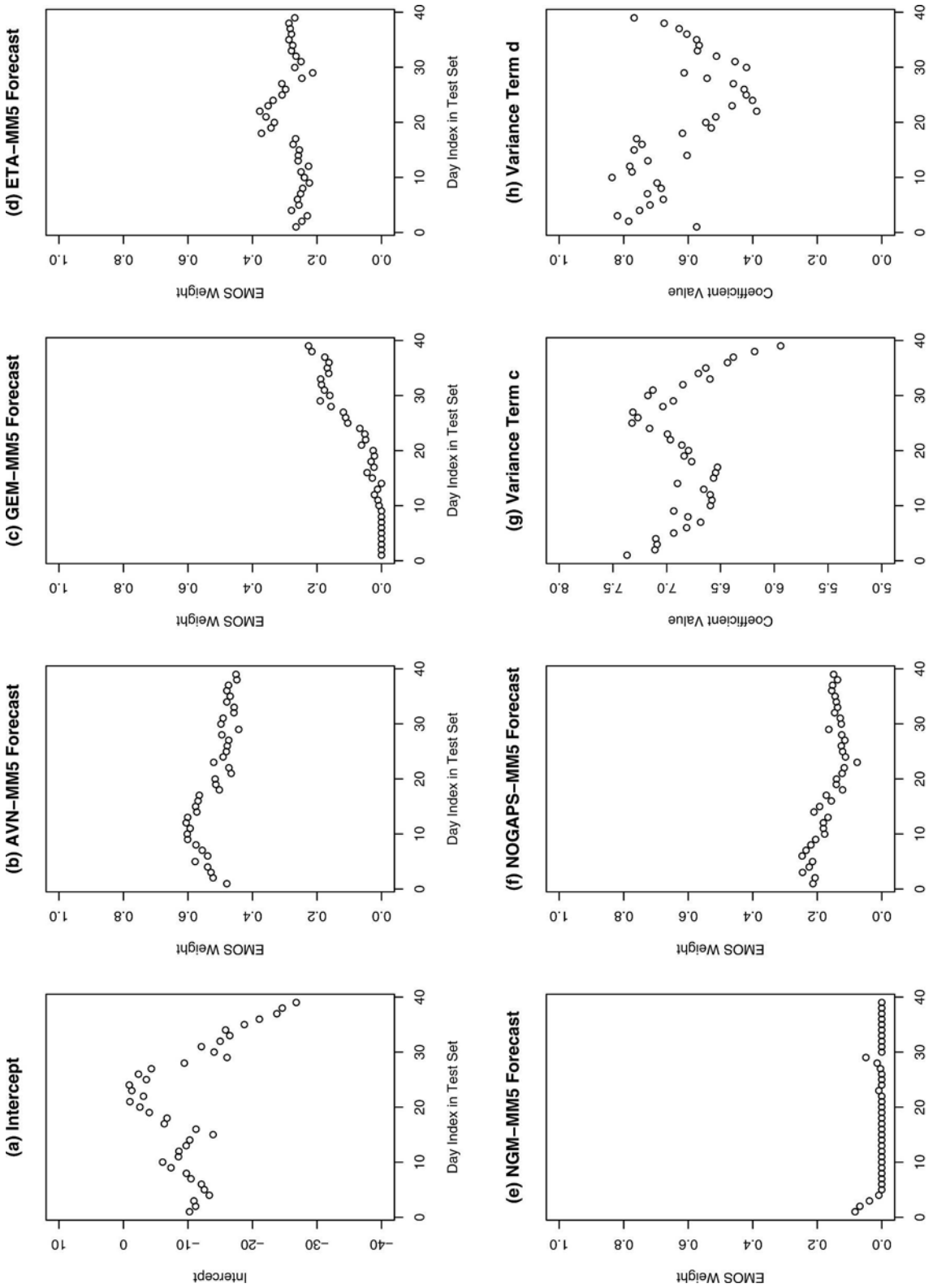

FIG. 9. Coefficient estimates for EMOS$^+$ forecasts of surface temperature over the Pacific Northwest, for each of the 39 days in the test period: (a) intercept, (b)–(f) member model weights, (g) and (h) variance terms c and d.

TABLE 6. Comparison of deterministic-style forecasts of surface temperature over the Pacific Northwest. The climatological, bias-corrected, EMOS, and EMOS$^+$ forecasts were trained on a sliding 40-day period.

	MAE	RMSE
AVN-MM5	2.45	3.15
GEM-MM5	2.64	3.40
ETA-MM5	2.52	3.23
NGM-MM5	2.56	3.28
NOGAPS-MM5	2.96	3.76
AVN-MM5 bias-corrected	2.31	3.00
GEM-MM5 bias-corrected	2.48	3.24
ETA-MM5 bias-corrected	2.39	3.10
NGM-MM5 bias-corrected	2.42	3.13
NOGAPS-MM5 bias-corrected	2.50	3.25
Mean of raw ensemble	2.49	3.18
Mean of bias-corrected ensemble	2.28	2.95
EMOS forecast	2.23	2.91
EMOS$^+$ forecast	2.23	2.91

RMSE of the ensemble member forecasts between 4% and 14%, and the NOGAPS-MM5 forecast showed the highest percentage improvement. The results in terms of the MAE were similar. The deterministic-style EMOS and EMOS$^+$ forecasts performed equally well, and they were more accurate than any of the other forecasts. The percentage improvement over the bias-corrected ensemble was less pronounced than for forecasts of sea level pressure.

We now turn to a discussion of probabilistic forecast skill. Table 7 shows that the bias-corrected ensemble was slightly better calibrated than the raw ensemble. However, both the raw ensemble and the bias-corrected ensemble were strikingly underdispersive, and this was reflected in the CRPS and IGN scores, which were computed on the basis of standard ensemble smoothing. When computed directly from the ensemble CDF, the CRPS scores for the raw ensemble and for the bias-corrected ensemble were 2.13 and 1.95, respectively. The EMOS and EMOS$^+$ techniques performed equally well, and they were better than any of the other forecasts, with a CRPS score that was 15% lower than for the bias-corrected ensemble, and an IGN score that was 13 points lower. The verification rank histograms and PIT histograms are shown in Figs. 10 and 11. The PIT histograms accentuate the underdispersion of the ensemble forecasts, and the histograms for the EMOS and EMOS$^+$ forecasts are close to being uniform.

TABLE 7. Comparison of predictive PDFs for surface temperature over the Pacific Northwest. The bias-corrected ensemble, the EMOS, and the EMOS$^+$ forecasts were trained on a sliding 40-day period.

	66⅔% prediction interval		Score	
	Coverage	Average width	CRPS	IGN
Raw ensemble	28.7	2.55	2.07	21.45
Bias-corrected ensemble	31.1	2.44	1.89	15.50
EMOS forecast	68.6	5.43	1.61	2.49
EMOS$^+$ forecast	68.6	5.41	1.61	2.49

d. Results for reduced ensembles

In our experiments for sea level pressure and surface temperature forecasts, the EMOS and EMOS$^+$ predictive PDFs were equally skillful. This comparison is summarized in Table 8, and the scores for the two techniques are almost indistinguishable. We interpreted the vanishing EMOS$^+$ coefficients in terms of reduced ensembles and argued that ensemble member models with consistently vanishing EMOS$^+$ weights could be removed from the ensemble, without sacrificing forecast skill.

For forecasts of sea level pressure, Fig. 5 suggests the use of a three-member ensemble, consisting of the AVN-MM5, GEM-MM5, and NOGAPS-MM5 forecasts. We applied the standard EMOS technique to the reduced three-member ensemble, and the results are shown in Table 8. The MAE, RMSE, CRPS, and IGN scores for the EMOS forecasts, the EMOS$^+$ forecasts and the EMOS forecasts using the reduced ensemble are almost indistinguishable. Similarly, Fig. 9 suggests that for forecasts of surface temperature the NGM-MM5 forecast could be removed from the ensemble. Table 8 compares the EMOS predictive PDFs based on the reduced four-member ensemble to the EMOS and EMOS$^+$ forecasts based on the full ensemble, and the comparison is favorable. Clearly, minimal differences in scores must not be overinterpreted, and ensemble members that contribute little to forecasting one variable might be useful for others. That said, the NGM-MM5 model was removed from the University of Washington ensemble on 1 August 2000, shortly after the end of our test period.

4. Discussion

It is well documented in the literature that multiple regression or superensemble techniques improve the deterministic-style forecast accuracy of ensembles systems (Krishnamurti et al. 1999, 2000; Kharin and Zwiers 2002). Regression-based forecasts correct for model biases and therefore are more accurate than the ensemble mean forecast. The novelty of our ensemble model output statistics (EMOS) approach is threefold. We apply linear regression techniques to obtain full predictive PDFs and CDFs, rather than deterministic-style forecasts, for continuous weather variables. For estimating the EMOS coefficients, we use the novel method of minimum CRPS estimation. Finally, the EMOS$^+$ implementation constrains the regression coefficients to be nonnegative, thereby allowing for an interpretation in terms of the relative usefulness of the ensemble member models, given all the others.

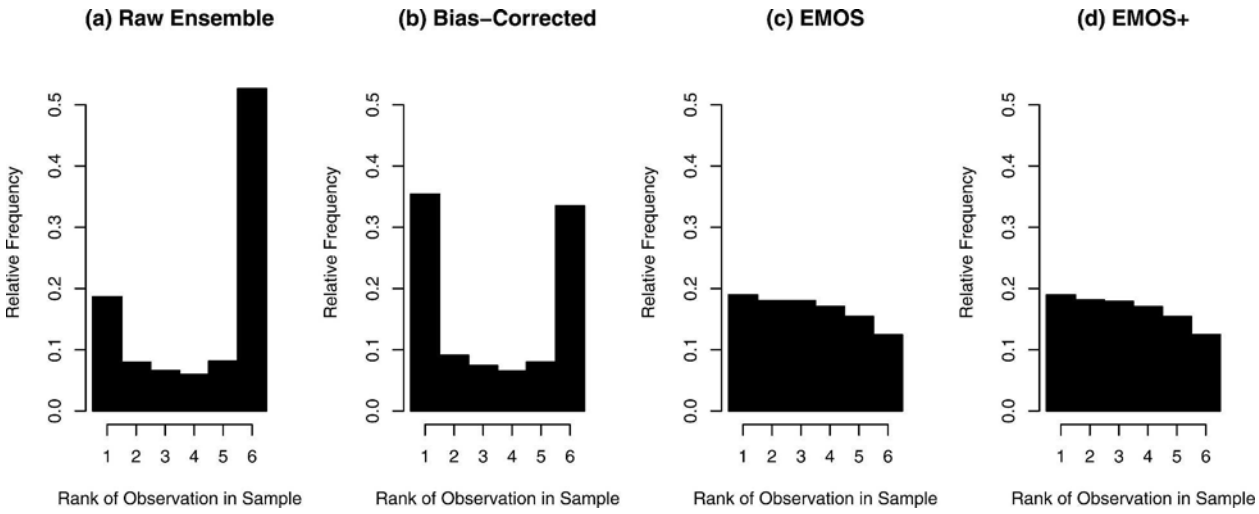

FIG. 10. Verification rank histograms for ensemble forecasts of surface temperature over the Pacific Northwest: (a) raw ensemble, (b) bias-corrected ensemble, (c) EMOS ensemble, and (d) EMOS$^+$ ensemble.

EMOS$^+$ identifies ensemble members whose relative contributions are negligible, typically as a result of collinearity, and removes them from the regression equation. The EMOS and EMOS$^+$ predictive PDFs are Gaussian, and they take account of the spread-skill relationship, in that the predictive variance is a linear function of the ensemble variance. However, both methods adapt to the absence of spread-error correlation, by estimating the variance coefficient d as negligibly small. Drawing a random sample from the Gaussian predictive CDF is a straightforward exercise, and forecast ensembles of any size can be generated. An alternative, and likely preferable, way of forming an m-member ensemble from the predictive CDF is by taking the forecast quantiles at level $i/(m + 1)$ for $i = 1, \ldots, m$, respectively.

We applied the EMOS and EMOS$^+$ techniques to sea level pressure and surface temperature forecasts over the North American Pacific Northwest in spring 2000, using the University of Washington mesoscale ensemble (Grimit and Mass 2002). The EMOS and EMOS$^+$ predictions were equally accurate, and they had lower RMSE and MAE than any of the member model forecasts, the bias-corrected member model forecasts, the ensemble mean forecast, and the ensemble mean of the bias-corrected member models. We also assessed the probabilistic forecast skill of the EMOS and EMOS$^+$ predictive PDFs. Both methods performed equally well and had substantially lower CRPS and IGN scores than the raw ensemble or bias-corrected ensemble. The predictive PDFs were much better calibrated than the raw ensemble or bias-

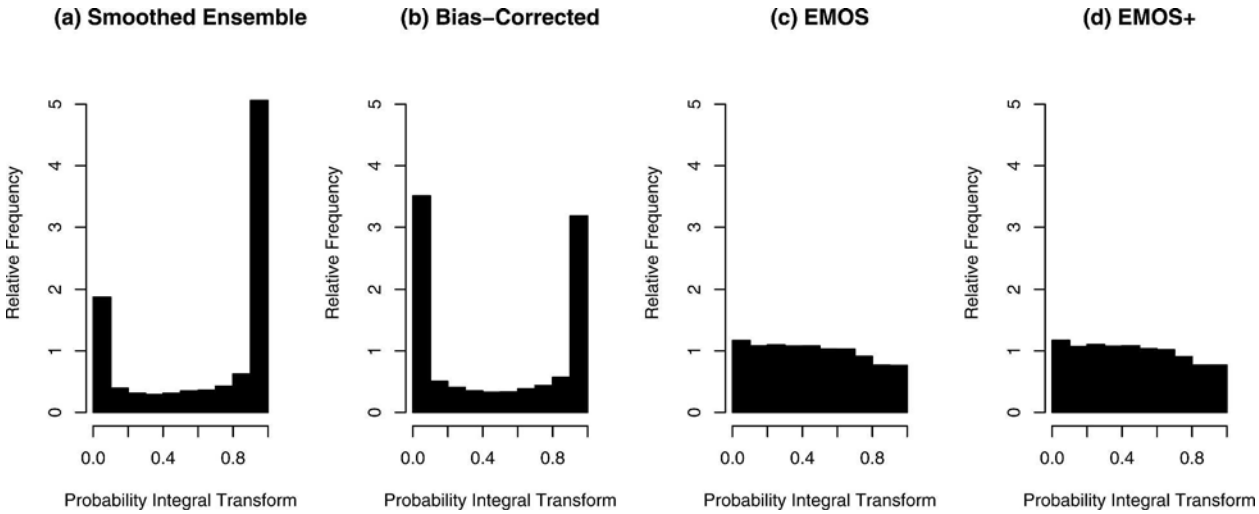

FIG. 11. PIT histograms for PDF forecasts of surface temperature over the Pacific Northwest: (a) smoothed ensemble forecast, (b) bias-corrected smoothed ensemble forecast, (c) EMOS forecast, and (d) EMOS$^+$ forecast.

TABLE 8. Comparison of EMOS forecasts, EMOS$^+$ forecasts, and EMOS forecasts based on reduced ensembles, as described in section 3d. The reduced ensemble for forecasts of sea level pressure excludes ETA-MM5 and NGM-MM5; the reduced temperature ensemble excludes NGM-MM5 only.

	Score		Score		66⅔% prediction interval	
	MAE	RMSE	CRPS	IGN	Coverage	Average width
			Sea level pressure			
EMOS	1.966	2.484	1.393	2.326	65.91	4.712
EMOS$^+$	1.953	2.487	1.389	2.326	67.61	4.747
EMOS (three-member ensemble)	1.952	2.486	1.388	2.327	67.66	4.748
			Surface temperature			
EMOS	2.231	2.907	1.606	2.487	68.58	5.427
EMOS$^+$	2.230	2.906	1.606	2.488	68.57	5.411
EMOS (four-member ensemble)	2.227	2.904	1.604	2.486	68.82	5.433

corrected ensemble and they were sharp, in that the prediction intervals were much shorter on average than prediction intervals based on climatology. Perhaps surprisingly, the EMOS and EMOS$^+$ predictive PDFs for sea level pressure were frequently sharper than the raw ensemble forecasts. With small modifications, as explained in section 2c, our methods apply to all ensemble systems, including weather and climate, synoptic-scale, poor person's, multimodel, multianalysis, perturbed observations, singular vector, and bred ensembles. EMOS and EMOS$^+$ can be applied to gridded ensemble output, thereby providing probabilistic forecasts on a grid. The resulting forecast fields can be visualized in the form of percentile maps, as in Fig. 6 of Raftery et al. (2005). In our experiments, we used observations to estimate the EMOS and EMOS$^+$ coefficients, but this could also be done using an analysis.

Bias correction results in more accurate deterministic-style forecasts, and bias correction of the individual member model forecasts reduces the ensemble spread, by pulling the individual members toward the verification mean (Eckel 2003). Verification rank histograms typically become more symmetric after bias correction, as in our Fig. 10, or in Fig. 46 of Eckel (2003). However, bias correction does not necessarily result in improved calibration, and the need for statistical postprocessing remains. We anticipate significant improvements in probabilistic forecast skill through the use of advanced bias correction schemes, followed by statistical postprocessing of the bias-corrected member model ensemble. Further research in this direction is desirable.

We close with a discussion of potential extensions as well as limitations of the EMOS method. The predictive PDFs produced by the EMOS and EMOS$^+$ techniques are Gaussian and therefore unimodal. This is unlikely to be a great disadvantage for a five-member ensemble, such as the University of Washington ensemble that we considered. However, larger ensembles sometimes suggest multimodal forecast PDFs. The ensemble smoothing approach of Wilks (2002) and the Bayesian model averaging approach of Raftery et al. (2005) address this issue.

We obtained EMOS and EMOS$^+$ forecasts of sea level pressure and surface temperature. These are variables for which the forecast error distributions are approximately Gaussian. The forecast error distributions for other variables, such as precipitation or cloud cover, are unlikely to be close to normal. Wilks (2002) proposes ways of transforming forecast ensembles to Gaussian distributions, and EMOS and EMOS$^+$ can be applied to the transformed ensemble. Another approach that remains largely unexplored uses the framework of generalized linear models (McCullagh and Nelder 1989).

Our methods provide predictive PDFs of continuous weather variables at a given location, but they do not reproduce the spatial correlation patterns of observed weather fields. Gel et al. (2004), among others, suggested a way of creating ensembles of entire weather fields, each of which honors the spatial correlation structure of verifying fields. However, this approach uses only one numerical weather prediction model rather than an ensemble of forecasts. This method could be combined with EMOS or EMOS$^+$ to yield calibrated ensembles of entire weather fields, by simulating correlated error fields and adding them to the spatially varying mean of the predictive distributions. Such an approach could also be viewed as a dressing method (Roulston and Smith 2003). A more straightforward approach to visualizing the forecast fields uses percentile maps, as suggested above. Percentile maps do not reproduce the spatial correlation structure of observed weather fields, nor do they take account of dynamical features. However, they provide concise summaries of the predictive PDFs and may facilitate the interpretation and thereby foster the acceptance and the use of probabilistic forecasts.

Acknowledgments. The authors are grateful to Mark Albright, Jeffrey L. Anderson, Anthony F. Eckel, Eric P. Grimit, James A. Hansen, Clifford F. Mass, and Jon A. Wellner for constructive comments, helpful discussions, and providing data. This research was supported by the DoD Multidisciplinary University Research Initiative (MURI) program administered by the Office of Naval Research under Grant N00014-01-10745.

REFERENCES

Anderson, J. L., 1996: A method for producing and evaluating probabilistic forecasts from ensemble model integrations. *J. Climate,* **9,** 1518–1530.

Atger, F., 2003: Spatial and interannual variability of the reliability of ensemble-based probabilistic forecasts: Consequences for calibration. *Mon. Wea. Rev.,* **131,** 1509–1523.

Birgé, L., and P. Massart, 1993: Rates of convergence for minimum contrast estimators. *Probab. Theory Rel. Fields,* **97,** 113–150.

Dawid, A. P., 1984: Statistical theory: The prequential approach. *J. Roy. Stat. Soc.,* **A147,** 278–292.

Déqué, M., J. T. Royer, and R. Stroe, 1994: Formulation of Gaussian probability forecasts based on model extended-range integrations. *Tellus,* **46A,** 52–65.

Eckel, F. A., 2003: Effective mesoscale, short-range ensemble forecasting. Ph.D. dissertation, University of Washington, 224 pp. [Available online at www.atmos.washington.edu/~ens/pubs_n_pres.html.]

——, and M. K. Walters, 1998: Calibrated probabilistic quantitative precipitation forecasts based on the MRF ensemble. *Wea. Forecasting,* **13,** 1132–1147.

Gel, Y., A. E. Raftery, and T. Gneiting, 2004: Calibrated probabilistic mesoscale weather field forecasting: The geostatistical output perturbation (GOP) method (with discussion and rejoinder). *J. Amer. Stat. Assoc.,* **99,** 575–590.

Glahn, H. R., and D. A. Lowry, 1972: The use of model output statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.,* **11,** 1203–1211.

Gneiting, T., and A. E. Raftery, 2004: Strictly proper scoring rules, prediction, and estimation. Tech. Rep. 463, Department of Statistics, University of Washington, 29 pp. [Available online at www.stat.washington.edu/tech.reports.]

——, ——, F. Balabdaoui, and A. Westveld, 2003: Verifying probabilistic forecasts: Calibration and sharpness. *Proc. Workshop on Ensemble Weather Forecasting in the Short to Medium Range,* Val-Morin, QC, Canada. [Available online at www.cdc.noaa.gov/people/tom.hamill/ef_workshop_2003_schedule.html.]

Good, I. J., 1952: Rational decisions. *J. Roy. Stat. Soc.,* **B14,** 107–114.

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

——, and ——, 2004: Forecasting mesoscale uncertainty: Short-range ensemble forecast error predictability. Preprints, *16th Conf. on Numerical Weather Prediction,* Seattle, WA, Amer. Meteor. Soc., CD-ROM, 24.3. [Available online at www.atmos.washington.edu/~ens/pubs_n_pres.html.]

Hamill, T. M., 2001: Interpretation of rank histograms for verifying ensemble forecasts. *Mon. Wea. Rev.,* **129,** 550–560.

——, and S. J. Colucci, 1997: Verification of Eta-RSM short-range ensemble forecasts. *Mon. Wea. Rev.,* **125,** 1312–1327.

——, and ——, 1998: Evaluation of Eta-RSM ensemble probabilistic precipitation forecasts. *Mon. Wea. Rev.,* **126,** 711–724.

——, J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447.

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

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.

Huber, P. J., 1964: Robust estimation of a location parameter. *Ann. Math. Stat.,* **35,** 73–101.

——, 1981: *Robust Statistics.* John Wiley, 308 pp.

Jewson, S., A. Brix, and C. Ziehmann, 2004: A new parametric model for the assessment and calibration of medium-range ensemble temperature forecasts. *Atmos. Sci. Lett.,* **5,** 96–102.

Kharin, V. V., and F. W. Zwiers, 2002: Climate predictions with multimodel ensembles. *J. Climate,* **15,** 793–799.

Krishnamurti, T. N., C. M. Kishtawal, T. E. LaRow, D. R. Bachiochi, Z. Zhang, C. E. Williford, S. Gadgil, and S. Surendan, 1999: Improved weather and seasonal climate forecasts from multimodel superensemble. *Science,* **285,** 1548–1550.

——, ——, Z. Zhang, T. E. LaRow, D. Bachiochi, E. Williford, S. Gadgil, and S. Surendan, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Mass, C. F., 2003: IFPS and the future of the National Weather Service. *Wea. Forecasting,* **18,** 75–79.

——, and Coauthors, 2003: Regional environmental prediction over the Pacific Northwest. *Bull. Amer. Meteor. Soc.,* **84,** 1353–1366.

McCullagh, P., and J. A. Nelder, 1989: *Generalized Linear Models.* 2d ed. Chapman & Hall, 511 pp.

Molteni, F., R. Buizza, T. N. Palmer, and T. Petroliagis, 1996: The ECMWF ensemble prediction system: Methodology and validation. *Quart. J. Roy. Meteor. Soc.,* **122,** 73–119.

Murphy, A. H., and R. L. Winkler, 1979: Probabilistic temperature forecasts: The case for an operational program. *Bull. Amer. Meteor. Soc.,* **60,** 12–19.

Pfanzagl, J., 1969: On the measurability and consistency of minimum contrast estimates. *Metrika,* **14,** 249–272.

Press, W. H., S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, 1992: *Numerical Recipes in FORTRAN: The Art of Scientific Computing.* 2d ed. Cambridge University Press, 963 pp.

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

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of sample size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489.

Rosenblatt, M., 1952: Remarks on a multivariate transformation. *Ann. Math. Stat.,* **23,** 470–472.

Roulston, M. S., and L. A. Smith, 2002: Evaluating probabilistic forecasts using information theory. *Mon. Wea. Rev.,* **130,** 1653–1660.

——, and ——, 2003: Combining dynamical and statistical ensembles. *Tellus,* **55A,** 16–30.

Saetra, Ø., H. Hersbach, J.-R. Bidlot, and D. S. Richardson, 2004: Effects of observation errors on the statistics for ensemble spread and reliability. *Mon. Wea. Rev.,* **132,** 1487–1501.

Scherrer, S. C., C. Appenzeller, P. Eckert, and D. Cattani, 2004: Analysis of the spread-skill relations using the ECMWF ensemble prediction system over Europe. *Wea. Forecasting,* **19,** 552–565.

Stefanova, L., and T. N. Krishnamurti, 2002: Interpretation of seasonal climate forecast using Brier score, the Florida State University superensemble, and the AMIP-I dataset. *J. Climate,* **15,** 537–544.

Stensrud, D. J., and N. Yussouf, 2003: Short-range predictions of 2-m temperature and dewpoint temperature over New England. *Mon. Wea. Rev.,* **131,** 2510–2524.

Talagrand, O., R. Vautard, and B. Strauss, 1997: Evaluation of probabilistic prediction systems. *Proc. Workshop on Predictability,* Reading, United Kingdom, European Centre for Medium-Range Weather Forecasts, 1–25.

Tibshirani, R., 1996: Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc.,* **B58,** 267–288.

Toth, Z., and E. Kalnay, 1997: Ensemble forecasting at NCEP and the breeding method. *Mon. Wea. Rev.,* **125,** 3297–3319.

——, O. Talagrand, G. Candille, and Y. Zhu, 2003: Probability and ensemble forecasts. *Forecast Verification: A Practitioner's Guide in Atmospheric Science,* I. T. Jolliffe and D. B. Stephenson, Eds., Wiley, 137–163.

Unger, D. A., 1985: A method to estimate the continuous ranked probability score. Preprints, *Ninth Conf. on Probability and Statistics in Atmospheric Sciences,* Virginia Beach, VA, Amer. Meteor. Soc., 206–213.

Van den Dool, H. M., and L. Rukhovets, 1994: On the weights for an ensemble-averaged 6–10-day forecast. *Wea. Forecasting,* **9,** 457–465.

Weigend, A. S., and S. Shi, 2000: Predicting daily probability distributions of S&P500 returns. *J. Forecasting,* **19,** 375–392.

Whitaker, J. S., and A. F. Loughe, 1998: The relationship between ensemble spread and ensemble mean skill. *Mon. Wea. Rev.,* **126,** 3292–3302.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 467 pp.

——, 2002: Smoothing forecast ensembles with fitted probability distributions. *Quart. J. Roy. Meteor. Soc.,* **128,** 2821–2836.

Wilson, L. J., W. R. Burrows, and A. Lanzinger, 1999: A strategy for verification of weather element forecasts from an ensemble prediction system. *Mon. Wea. Rev.,* **127,** 956–970.