Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging

WILLIAM KLEIBER* AND ADRIAN E. RAFTERY

Department of Statistics, University of Washington, Seattle, Washington

JEFFREY BAARS

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

TILMANN GNEITING

Institute of Applied Mathematics, University of Heidelberg, Heidelberg, Germany

CLIFFORD F. MASS

Department of Atmospheric Sciences, University of Washington, Seattle, Washington

ERIC GRIMIT

3Tier Environmental Forecast Group, Seattle, Washington

(Manuscript received 24 May 2010, in final form 16 November 2010)

ABSTRACT

The authors introduce two ways to produce locally calibrated grid-based probabilistic forecasts of temperature. Both start from the Global Bayesian model averaging (Global BMA) statistical postprocessing method, which has constant predictive bias and variance across the domain, and modify it to make it local. The first local method, geostatistical model averaging (GMA), computes the predictive bias and variance at observation stations and interpolates them using a geostatistical model. The second approach, Local BMA, estimates the parameters of BMA at a grid point from stations that are close to the grid point and similar to it in elevation and land use. The results of these two methods applied to the eight-member University of Washington Mesoscale Ensemble (UWME) are given for the 2006 calendar year. GMA was calibrated and sharper than Global BMA, with prediction intervals that were 8% narrower than Global BMA on average. Examples using sparse and dense training networks of stations are shown. The sparse network experiment illustrates the ability of GMA to draw information from the entire training network. The performance of Local BMA was not statistically different from Global BMA in the dense network experiment, and was superior to both GMA and Global BMA in areas with sufficient nearby training data.

1. Introduction

Probabilistic forecasting has experienced a recent surge of interest in the atmospheric sciences community. Early on, it was recognized that ensembles of forecasts could provide a measure of forecasting confidence for a given variable (Epstein 1969; Leith 1974). There was hope that ensembles of forecasts would produce an estimate of the predictive distribution for a specific weather quantity, and much research has been devoted to methods of generating representative ensembles (Toth and Kalnay 1993; Houtekamer and Derome 1995; Molteni et al. 1996; Stensrud et al. 1999; Hamill et al. 2000; Buizza et al. 2005). However, ensembles are often underdispersed (Hamill and Colucci 1997) and require postprocessing to properly calibrate the resulting distribution.

[*] Current affiliation: Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, Colorado.

Corresponding author address: William Kleiber, Department of Statistics, University of Washington, Box 354322, Seattle, WA 98195-4322.
E-mail: wkleiber@ucar.edu

DOI: 10.1175/2010MWR3511.1

Recently, work on postprocessing of ensembles has focused on generating calibrated probabilistic forecasts. Some approaches include nonhomogeneous Gaussian regression (Gneiting et al. 2005; Hagedorn et al. 2008), the best member method (Roulston and Smith 2003; Wang and Bishop 2005; Fortin et al. 2006), and logistic regression (Hamill et al. 2004), all of which have been recently compared by Wilks and Hamill (2007). Related approaches include kernel dressing (Bröcker and Smith 2008), moving average estimation (Johnson and Swinbank 2009), model output statistics (MOS; Glahn et al. 2009b), ensemble regression (Unger et al. 2009) and extended logistic regression (Wilks 2009). The need for postprocessing of ensemble output is also discussed in the climate literature (Kharin and Zwiers 2002; Tebaldi and Knutti 2007; Smith et al. 2009).

Postprocessing of ensembles using Bayesian model averaging (BMA), introduced by Raftery et al. (2005), has enjoyed success in forecasting weather quantities such as 2-m temperature, sea level pressure, precipitation (Sloughter et al. 2007), and wind speed (Sloughter et al. 2010), as well as hydrologic streamflow (Duan et al. 2007). BMA is a method of combining predictive densities generated by individual members of an ensemble. We focus on surface temperature as the variable of interest here, though the methods we introduce can be adapted to other quantities. If y_{st} is the temperature at site s, valid at time t, with K forecasts $f_{1st}, \ldots, f_{Kst}$, the BMA predictive density for y_{st} is

$$p(y_{st}|f_{1st}, \ldots, f_{Kst}) = \sum_{\ell=1}^{K} w_\ell g(y_{st}|f_{\ell st}), \qquad (1)$$

where $g(y_{st}|f_{\ell st})$ is a normal density with bias-corrected mean $f_{\ell st} - a_\ell$ and variance σ^2. We refer to this model for temperature as Global BMA. Global BMA is a global model in the sense that it does not adjust the statistical parameters (such as bias and predictive variance) locally.

Postprocessing of numerical weather forecasts has been carried out since the advent of MOS (Glahn and Lowry 1972). Systematic errors in numerical weather prediction models can be removed using MOS, but often these errors (which we refer to as biases; see Dee 2005) vary spatially. These biases can only be computed at observation locations. However, recent interest has focused on removing bias across the entire model grid, where there are usually no direct observations. The most common approach to gridded bias correction is to interpolate relevant information from surrounding observation stations to a given grid point. Various ways of doing this interpolation have been proposed in recent years: Yussouf and Stensrud (2006) interpolated observed biases to the model grid using a Cressman (1959) scheme. Hacker and Rife (2007) interpolated bias analyses using minimum variance estimates. Glahn et al. (2009a) described an approach to gridding MOS predictions that accounts for elevation and the distinction between water- and land-based model grid points. Mass et al. (2008) introduced an approach to gridded bias correction that is sensitive to features that affect model bias, such as elevation, land-use type, and forecast value. Their gridded bias estimation is based on an interpolation scheme, which we refer to as the Mass–Baars interpolation method. Mass–Baars interpolation is used extensively in one of our two locally adaptive probabilistic approaches, and is described fully in section 4.

Generally, the two fields of probabilistic forecasting and grid-based model corrections do not overlap. There have been some recent developments in combining numerical model output and observational data that suggest a hybrid approach that locally adjusts a statistical postprocessing model based on observations. Berrocal et al. (2009) described a statistical model with spatially varying parameters to downscale average gridcell level numerical model output for ozone concentration, which was generalized to the bivariate case in a follow-up study (Berrocal et al. 2010). A similar, but not fully Bayesian, approach has been implemented by Liu et al. (2008).

In this paper we explore postprocessing methodologies to generate locally calibrated predictive distributions based on an ensemble of forecasts. We introduce two approaches, both based on the BMA work of Raftery et al. (2005). The first can be thought of as a local generalization of Global BMA, which we call geostatistical model averaging (GMA). In particular, GMA will allow the bias correction and predictive variance parameters to vary by location. GMA belongs to the general class of spatially varying-coefficient models (Hastie and Tibshirani 1993; Gelfand et al. 2003, 2005). Our second method interpolates relevant forecast errors first according to the Mass–Baars interpolation scheme, and then estimates the Global BMA model at each model grid point. We call this approach Local BMA.

The remainder of the paper is structured as follows: section 2 introduces the GMA model for the case of a single forecast. This is extended to an ensemble of forecasts in section 3. Section 4 describes Local BMA. The following two sections are devoted to illustrating the models: aggregate results over the Pacific Northwest are considered in section 5, followed by a detailed case study at four locations in section 6. We end the paper with a discussion and possible extensions.

2. Geostatistical single forecast model

The basic model for temperature y_{st} at site s valid at time t is

$$y_{st} = f_{st} - a_s + \varepsilon_{st}, \tag{2}$$

where f_{st} is the corresponding forecast, a_s is an additive bias correction, and ε_{st} has a normal distribution with mean 0 and variance σ_s^2.

a. Estimation

Suppose we have training data at n sites $s = s_1, \ldots, s_n$ with forecasts f_{st} and validating observations y_{st}. For any given training site s, empirical estimates of bias and variance are

$$\hat{a}_s = \frac{1}{T}\sum_{t=1}^{T}(f_{st} - y_{st}) \quad \text{and} \quad \hat{\sigma}_s^2 = \frac{1}{T}\sum_{t=1}^{T}(f_{st} - y_{st} - \hat{a}_s)^2,$$

where the sum and variance are over some prespecified training period of length T.

We view the empirical estimates $\{\hat{a}_{s_i}\}_{i=1}^{n}$ as a partial realization from a stationary Gaussian random field in three dimensions, $\mathbb{R}^3$, with mean μ_a and the covariance function:

$$\begin{aligned} C_a(s_1, s_2) &= \mathrm{Cov}(a_{s_1}, a_{s_2}) \\ &= \tau_a^2 + \rho_a^2 \exp\left[-\frac{\|s_1 - s_2\|}{r_{a1}} - \frac{|h(s_1) - h(s_2)|}{r_{a2}}\right], \end{aligned} \tag{3}$$

where $\|\cdot\|$ is the Euclidean norm. The parameter τ_a^2 is the nugget effect, which corresponds to measurement error or microscale variability; ρ_a^2 is a variance parameter; r_{a1} is the range corresponding to horizontal distance; and r_{a2} is the range corresponding to vertical distance, where $h(s)$ is the elevation at location s.

We define $\upsilon_s = \log\sigma_s^2$, with empirical estimates $\hat{\upsilon}_s = \log\hat{\sigma}_s^2$. We view $\{\hat{\upsilon}_{s_i}\}_{i=1}^{n}$ as a partial realization from a stationary Gaussian random field with mean μ_υ and the covariance function:

$$\begin{aligned} C_\upsilon(s_1, s_2) &= \mathrm{Cov}(\upsilon_{s_1}, \upsilon_{s_2}) \\ &= \tau_\upsilon^2 + \rho_\upsilon^2 \exp\left[-\frac{\|s_1 - s_2\|}{r_{\upsilon 1}} - \frac{|h(s_1) - h(s_2)|}{r_{\upsilon 2}}\right]. \end{aligned} \tag{4}$$

These random field parameters are estimated by maximum likelihood using the empirical values $\{\hat{a}_{s_i}\}_{i=1}^{n}$ and $\{\hat{\upsilon}_{s_i}\}_{i=1}^{n}$ as data. There is no closed form for these estimates, so they must be found via numerical maximization; we use the limited memory quasi-Newton bound constrained optimization method of Byrd et al. (1995).

We now introduce some notation. First, denote the maximum likelihood estimates of the random field parameters by hats (e.g., $\hat{\mu}_a$). Let $\hat{C}_a(\cdot,\cdot)$ and $\hat{C}_\upsilon(\cdot,\cdot)$ be the covariance functions for the a_s and υ_s processes defined by (3) and (4), respectively, with the maximum likelihood estimates plugged in. Define the covariance matrices $\Sigma_a = \{\hat{C}_a(s_i, s_j)\}_{i,j=1}^{n}$ and $\Sigma_\upsilon = \{\hat{C}_\upsilon(s_i, s_j)\}_{i,j=1}^{n}$. For any site of interest, s_0, let $\hat{\mathbf{c}}_a = [\hat{C}_a(s_0, s_1), \ldots, \hat{C}_a(s_0, s_n)]'$ and $\hat{\mathbf{c}}_\upsilon = [\hat{C}_\upsilon(s_0, s_1), \ldots, \hat{C}_\upsilon(s_0, s_n)]'$ be the vectors of estimated covariances for the two processes between the site of interest and the station locations. Finally, let $\hat{\mathbf{a}} = (\hat{a}_{s_1}, \ldots, \hat{a}_{s_n})'$ and $\hat{\boldsymbol{\upsilon}} = (\hat{\upsilon}_{s_1}, \ldots, \hat{\upsilon}_{s_n})'$ be the vectors of empirical estimates of the bias and log variance at the observation sites.

b. Forecasting

The predictive distribution at site s_0 valid at time t is specified by (1). Unless s_0 is a training site, there are no direct estimates of a_{s_0} or $\sigma_{s_0}^2$, so we use a geostatistical method of interpolation known as kriging (Cressie 1993; Stein 1999). Kriging yields the best linear unbiased predictor under a quadratic loss function. The kriging estimates of a_{s_0} and υ_{s_0} are

$$\hat{a}_{s_0} = \hat{\mu}_a + \hat{\mathbf{c}}_a'\Sigma_a^{-1}(\hat{\mathbf{a}} - \hat{\mu}_a\mathbf{1}) \tag{5}$$

and

$$\hat{\upsilon}_{s_0} = \hat{\mu}_\upsilon + \hat{\mathbf{c}}_\upsilon'\Sigma_\upsilon^{-1}(\hat{\boldsymbol{\upsilon}} - \hat{\mu}_\upsilon\mathbf{1}), \tag{6}$$

where $\mathbf{1}$ is a vector of ones of length n. In this case, the kriging estimates correspond to the conditional expectations of a_{s_0} and υ_{s_0} given $\hat{\mathbf{a}}$ and $\hat{\boldsymbol{\upsilon}}$, respectively, under the assumption that the maximum likelihood estimates of the spatial parameters are the true underlying spatial parameters (Chilès and Delfiner 1999). The final predictive distribution for y_{s_0t} is then normal with mean $f_{s_0t} - \hat{a}_{s_0}$ and variance $\hat{\sigma}_{s_0}^2 = \exp(\hat{\upsilon}_{s_0})$.

3. Geostatistical model averaging

In the last section we considered the situation where we have just one forecast for each site and valid time. We now extend this to the situation where we have an ensemble of forecasts at each site and valid time.

Suppose that at each of n training sites $s = s_1, \ldots, s_n$, with $s \in \mathbb{R}^3$, we have K forecasts $f_{1st}, \ldots, f_{Kst}$ at site s valid at time t. The BMA approach to combining forecasts is described in Raftery et al. (2005), where the predictive density of temperature, y_{st}, given the K forecasts, is defined by (1), where $g(y_{st}|f_{\ell st})$ is a normal density with mean $f_{\ell st} - a_\ell$ and variance σ^2. Here, the additive bias corrections a_ℓ and predictive variance σ^2 are common among all sites for any time t, and hence define the Global BMA model.

In the GMA formulation, the additive bias term becomes, for forecast ℓ at site s, $a_{\ell s}$. Raftery et al. (2005) suggest the assumption of a common variance among all component densities is reasonable, so that the GMA predictive variance becomes $c\sigma_s^2$, where $\sigma_s^2 = \exp(\upsilon_s)$. The variance deflation factor c is chosen so as to produce a calibrated predictive density (Berrocal et al. 2007).

a. Estimation

Suppose we have training data consisting of K forecasts $\{f_{\ell st}\}_{\ell=1}^K$ along with observations at the n observation sites. For training site s, empirical estimates of bias and variance are

$$\hat{a}_{\ell s} = \frac{1}{T}\sum_{t=1}^{T}(f_{\ell st} - y_{st}) \quad \text{and}$$

$$\hat{\sigma}_s^2 = \frac{1}{KT}\sum_{t=1}^{T}\sum_{\ell=1}^{K}(f_{\ell st} - y_{st} - \overline{e}_s)^2,$$

where the sum is over a training period, usually of length $T = 25$ days, and the variance is over this training period along with all K forecasts, where $\overline{e}_s$ is the average of the $K \times T$ errors $f_{\ell st} - y_{st}$.

As in the single forecast case, we view the empirical estimates $\{\hat{a}_{\ell s_i}\}_{i=1}^n$ as being drawn from stationary Gaussian random fields with covariance functions of the form (3), but with forecast-specific parameters $\mu_{a\ell}$, $\tau_{a\ell}^2$, $\rho_{a\ell}^2$, $r_{a1\ell}$, $r_{a2\ell}$, for $\ell = 1, \ldots, K$. The site-specific log variances are $\upsilon_s = \log\sigma_s^2$, with estimates collapsed across ensemble members denoted by $\hat{\upsilon}_s = \log\hat{\sigma}_s^2$. The model for υ_s follows directly from the single forecast case. The random field parameters are estimated by maximum likelihood, using the empirical estimates $\{\hat{a}_{\ell s_i}\}_{i=1}^n$ and $\{\hat{\upsilon}_{s_i}\}_{i=1}^n$.

b. Forecasting

As in the single forecast case, the first step is to set up the predictive densities $\{g(y_{s_0 t}|f_{\ell s_0 t})\}_{\ell=1}^K$ for any site of interest s_0. The kriging equations (5) and (6) yield estimates $\hat{a}_{1s_0}, \ldots, \hat{a}_{Ks_0}$, and $\hat{\upsilon}_{s_0}$ that are plugged into the component densities. Thus, the final predictive density for $y_{s_0 t}$ is (1), where $g(y_{s_0 t}|f_{\ell s_0 t})$ is normal with mean $f_{\ell s_0 t} - \hat{a}_{\ell s_0}$ and variance $c\exp(\hat{\upsilon}_{s_0})$, for $\ell = 1, \ldots, K$. The BMA weights $w_1, \ldots, w_K$ and the variance deflation parameter c are estimated via the expectation-maximization (EM) algorithm (Dempster et al. 1977), which we describe in the appendix. Once a stopping criterion for the EM algorithm is reached, the estimates $w_1, \ldots, w_K$ and c are used in the predictive density (1).

4. Local Bayesian model averaging

This paper describes two ways to approach the local prediction problem: GMA first estimates forecast error characteristics such as bias at the available observation stations, and then interpolates this information spatially. Alternatively, forecast errors could be interpolated first, followed by model estimation; this is the approach behind Local Bayesian model averaging (Local BMA). Local BMA is currently used operationally in the University of Washington's probabilistic forecasting project, Probcast (Mass et al. 2009); it combines the Mass–Baars interpolation technique with the Global BMA model to produce a predictive distribution that adapts to local characteristics. Mass–Baars interpolation is sensitive to features that affect model bias, such as elevation, land-use type, and forecast value, and works as follows.

Given forecasts f_{st} and observations y_{st} at n observation stations $s = s_1, \ldots, s_n$ at time t, the goal of the Mass–Baars interpolation scheme is to interpolate past forecast errors $f_{st} - y_{st}$ at observation stations s to grid point s_0. The interpolation scheme selects relevant forecast errors using the following criteria: observation sites must have an elevation that is close to the elevation at s_0, and must fall within some prespecified radius of s_0, with observing locations that are closer to s_0 given preference over those farther away. Observation sites must have a similar land-use category as s_0, as forecast biases may have different characteristics depending on the land-use type of the forecast site. Land-use types are split into nine groups sharing similar characteristics, as defined in Mass et al. (2008). To mitigate the effects of change of meteorological regime, the forecast errors must arise from a similar forecast value; for instance, if the forecast at s_0 is 20°C, then the only errors considered come from forecasts between, say, 18° and 22°C. To account for diurnal effects and differences due to model projection time, only errors from the same forecast hour as the forecast grid in question are used. Only recent errors are considered, and errors beyond a certain magnitude are ignored as they may be indicative of a problem station. The unspecified parameters here such as the interpolation radius are obtained using an optimization routine that minimizes mean absolute error based on training data. The Mass–Baars bias-correction technique estimates bias at any given grid point s_0 by interpolating forecast errors based on the above criteria, and averaging the resulting set of interpolated errors. The estimated bias may then be removed from the forecast at s_0. Local BMA, on the other hand, uses interpolated forecast errors from observation stations to produce a probabilistic forecast.

Local BMA operates as follows: given observations and an ensemble of K forecasts at sites $s = s_1, \ldots, s_n$, the Mass–Baars interpolation scheme is used to interpolate a predefined number of forecast errors $f_{\ell st} - y_{st}$ to a model grid point s_0, where t runs through some predefined training period, usually of no longer than several weeks. Then, BMA is applied, using these errors as data, at each model grid point separately, yielding a predictive density of the form (1), but with weights $w_{1s_0}, \ldots, w_{Ks_0}$, bias corrections $a_{1s_0}, \ldots, a_{Ks_0}$, and predictive variance $\sigma^2_{s_0}$ that are all specific to the grid point s_0. Finally, to produce a predictive distribution at any site within the domain, the locally estimated Global BMA model parameters are bilinearly interpolated from each of the surrounding four grid points. The control parameters that define the interpolation radius, elevation band, and forecast value band that determine the relevant forecast errors are chosen by minimizing the domain averaged mean absolute error on a set of held out station data. The specific values of these parameters, and the algorithm used to minimize the domain-averaged mean absolute error are described in Mass et al. (2008).

5. Aggregate results

Our two methods are applied to a temperature dataset over the North American Pacific Northwest during calendar years 2005–06. We use 48-h forecasts initialized at 0000 UTC from the 8-member University of Washington Mesoscale Ensemble (UWME), described in Eckel and Mass (2005). First we examine aggregate results over the entire forecast domain, followed by a focused study of results at four stations.

Since Local BMA chooses stations based on characteristics other than just station proximity, GMA and Local BMA usually choose different sets of observations. Depending on the density of the observation network, we might expect one of the two methods to be superior. For instance, if the observation network is very dense, GMA will focus on the nearest stations for local adjustments with no concern for features such as land type. In contrast, the nearby stations that Local BMA draws from may be more similar in terms of forecast errors, and hence may account for important information overlooked by GMA. On the other hand, if the observation network is sparse, Local BMA may not have enough "similar" stations nearby to make any local corrections (we always default to the Global BMA predictive density in these cases), while GMA is able to draw from all available sites in the network. Below we look at both scenarios, starting with the sparse network experiment.

a. Sparse network

For the sparse network, we restrict attention to observation stations on land, as exploratory analysis suggested no spatial correlation of forecast biases or variances over the Pacific Ocean, where there would be little or no gain in using a geostatistical approach; at these locations Global BMA will suffice for probabilistic predictions. That is, if one were interested in producing probabilistic forecasts at a station in the Pacific Ocean, one could use Global BMA (using, say, only training stations that are also ocean based), without the added complication of the locally varying model parameters. Land-based observation stations show strong spatial correlations: see the empirical estimates of the biases a_s and the log variances v_s for the GFS ensemble member using a 25-day training period leading up to 7 July 2005 in Fig. 1.

The 294 observation stations in this sparse experiment were chosen as if they represented a sparse, but reliable network, having an observation on at least 90% of all available days. Hence, each station has a nearly complete set of observations across the 2005–06 period, with very few short periods of missing data. Initially, we randomly divide the 294 observation stations into 194 for model fitting and 100 for validation. Figure 2 shows the locations of these two sets of stations. Empirical estimates of the bias and log variance processes a_s and v_s require a choice of training period; for both GMA and Global BMA, we adopt the 25-day training period that has been recommended in previous studies (Raftery et al. 2005).

GMA relies on a Gaussian process representation of the bias a_s and log variance v_s parameters. Exploratory analysis, such as empirical variograms, suggests that the functional form of the covariance functions (3) and (4) is justified, and also that the parameters defining the Gaussian structure of a_s and v_s are constant across seasons. Thus, we estimate these random field parameters only once, using 11 independent realizations generated by disjoint 25-day training periods from 20 January to 31 December 2005, and hold these estimates constant across the validation year of 2006. This duration is longer than 11 × 25 days since forecasts are missing on some days due to machine failure and disruptions in communications. These maximum likelihood estimates are presented in Table 1. The second-order parameters are almost indistinguishable between ensemble members, but the mean bias varies between forecasts.

With these estimates in hand, we perform validation on the entire year 2006. All models use a sliding training window of the previous T available days where $T = 25$

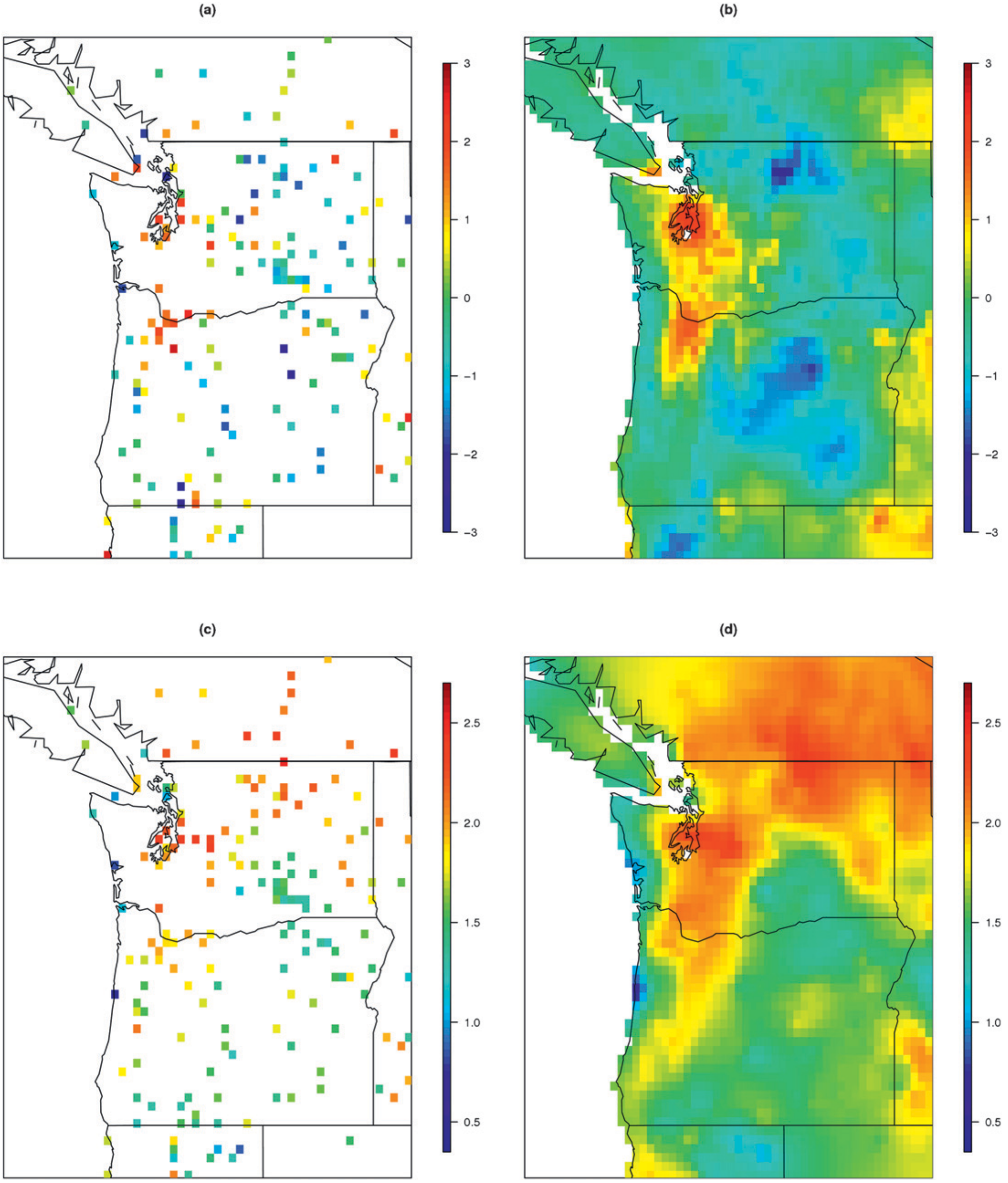

FIG. 1. Empirical estimates of (a) the bias process a_s (in °C), and (c) the log variance process v_s (in log Celsius^2) with kriged estimates of (b) bias and (d) log variance on the 12-km forecast grid, from the GFS ensemble member, using a 25-day training period leading up to 7 Jul 2005.

for GMA and Global BMA, and $T = 59$ for Local BMA. Other lengths of training periods were considered for GMA, but 25 days was found to produce the best results in terms of domain-averaged mean absolute error and continuous ranked probability score, similar to the experience of Raftery et al. (2005). The length of training period for Local BMA was chosen as that which minimized the domain aggregated mean absolute error, and we refer to Mass et al. (2008) for details. The network is sparse enough that occasionally Local BMA does not have enough similar stations near a grid point to interpolate forecast errors (there must be 8 nearby stations with 11 similar forecasts each for Local BMA to be available at a grid point). In these situations, we will substitute the Global BMA predictive density, thereby always guaranteeing a probabilistic prediction.

To assess the quality of the predictive distributions, we adopt the principle that probabilistic forecasts aim

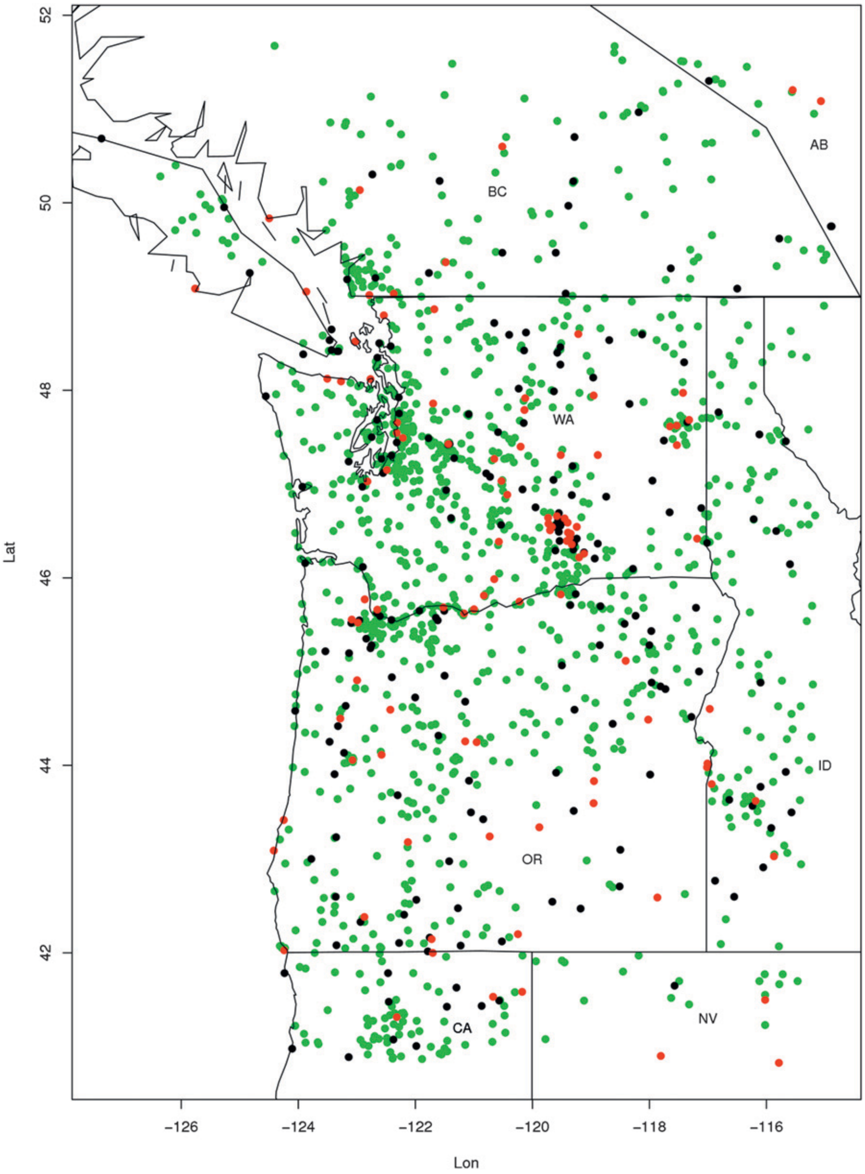

FIG. 2. Observation station locations: fitting stations in the sparse network are shown by black dots, fitting stations in the dense network are shown by both black and green dots, and validation stations are shown by red dots.

to maximize sharpness subject to calibration (Gneiting et al. 2007). Calibration requires statistical consistency between the predictive distributions and validating observations, and may be assessed via the probability integral transform (PIT) histogram (Diebold et al. 1998; Raftery et al. 2005; Gneiting et al. 2007). If F is the predictive cumulative distribution function for the observed quantity y, then the PIT histogram is a plot of $F(y)$ over many instances. A perfectly calibrated distribution will result in a uniform PIT histogram, while an overdispersed predictive distribution will put more mass in the center, and finally underdispersion is indicated by a U-shaped histogram. PIT histograms are a continuous analog of the rank histogram, which we use to describe the ensemble's calibration (Hamill 2001). Figure 3 shows the PIT histograms for the Global BMA model, Local BMA, and GMA, with a rank histogram for the raw ensemble.

It is immediately seen that the raw ensemble is underdispersed, a common feature of many types of forecasts (Hamill and Colucci 1997). The Global BMA model and GMA show much better calibration, while Local BMA reduces the underdispersion, but does not remove it completely.

TABLE 1. Maximum likelihood estimates for the additive bias processes $a_{\ell s}$ (in °C) and the log variance process υ_s (in log Celsius^2), for each member of the UWME. Distance is in km, and elevation is in m. The Global Forecast System (GFS) is from the National Centers for Environmental Prediction (NCEP), the Global Environmental Multiscale Model from the Canadian Meteorological Centre (CMCG), Eta is the limited-area mesoscale model from NCEP, the Global Analysis and Prediction (GASP) model is from the Australian Bureau of Meteorology, the Global Spectral Model is from the Japan Meteorological Agency (JMA), the Navy Operational Global Atmospheric Prediction System (NGPS) is from the Fleet Numerical Meteorological and Oceanographic Center, the Global Forecast System is from the Taiwan Central Weather Bureau (TCWB), and the Unified Model is from the Met Office.

	Bias processes, $a_{\ell s}$				
Forecast	μ_a	τ_a^2	ρ_a^2	r_{a1}	r_{a2}
GFS	−0.21	0.48	3.36	307	2159
CMCG	0.14	0.46	3.40	287	2090
ETA	0.10	0.52	3.64	356	2429
GASP	−0.43	0.43	3.57	304	2139
JMA	−0.47	0.46	3.35	300	1992
NGPS	−0.47	0.44	3.67	321	2240
TCWB	−0.69	0.44	3.76	333	2193
UKMO	−0.16	0.45	3.37	301	2109
	Log variance process, υ_s				
	μ_υ	τ_υ^2	ρ_υ^2	$r_{\upsilon 1}$	$r_{\upsilon 2}$
	1.78	0.0076	0.23	136	2800

TABLE 2. MAE and mean CRPS (both in °C) for the raw ensemble, Global BMA, Local BMA, and GMA over the calendar year 2006.

	Sparse network		Dense network	
	MAE	CRPS	MAE	CRPS
Raw ensemble	1.958	1.603	1.958	1.603
Global BMA	1.865	1.350	1.875	1.356
Local BMA	1.889	1.377	1.883	1.375
GMA	1.834	1.326	1.849	1.333

Sharpness refers to the concentration of the predictive distribution; the continuous ranked probability score (CRPS) assesses both sharpness and calibration, and is defined by

$$\mathrm{CRPS}(F, x) = \int_{-\infty}^{+\infty} [F(y) - \mathbb{I}(y \geq x)]^2 \, dy, \qquad (7)$$

where F is the predictive distribution and x is the observed temperature (Matheson and Winkler 1976; Hersbach 2000; Grimit et al. 2006; Gneiting and Raftery 2007). The CRPS and mean absolute error (MAE) between the validating observations and median of the predictive distribution are displayed in Table 2. Global BMA improves the raw ensemble's MAE by 4.8%, Local BMA improves it by 3.5%, while GMA improves it by 6.3%. Similarly, the raw ensemble's CRPS is improved by 15.8% using Global BMA, 14.1% by Local BMA, and 17.3% by GMA. Indeed, GMA improves the aggregate mean CRPS and MAE over Global BMA; the standard error of the difference in CRPS between Global BMA and GMA is 0.003°C and for MAE the standard error is 0.005°C, indicating that the improvement in both scores is statistically significant. Similarly, the standard error for difference between Global BMA and Local BMA is 0.003° and 0.005°C for CRPS and MAE, respectively, also indicating that the differences shown in Table 2 are significantly different, in particular that Local BMA is performing worse than Global BMA at an aggregate level.

We also calculate the CRPS value for each model at each validation station separately, with Fig. 4a summarizing the results. Each validation station is color coded corresponding to the model with the best local CRPS value. GMA has the best CRPS value at 47 stations, while Global BMA has the lowest value at 42, Local BMA at 10, and the raw ensemble is best at 1 station.

The aggregate scores of CRPS and MAE show improvement using all models over the raw ensemble, and

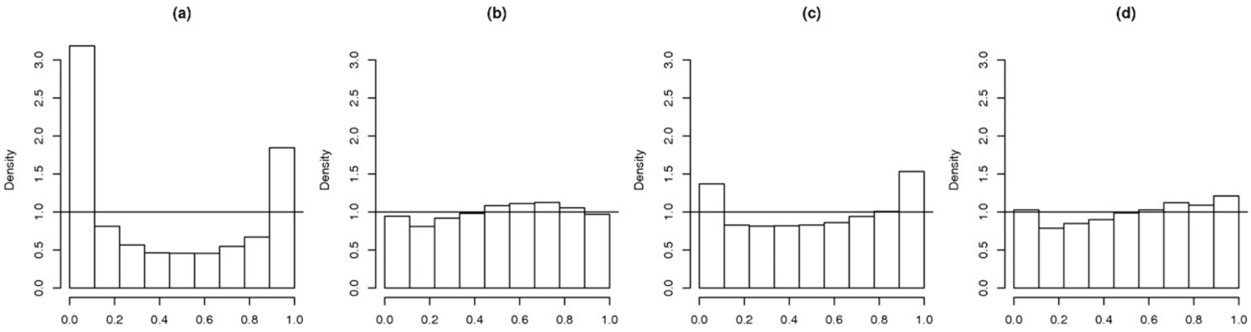

FIG. 3. Sparse network experiment: (a) rank histogram for the raw ensemble, and PIT histograms for (b) Global BMA, (c) Local BMA, and (d) GMA.

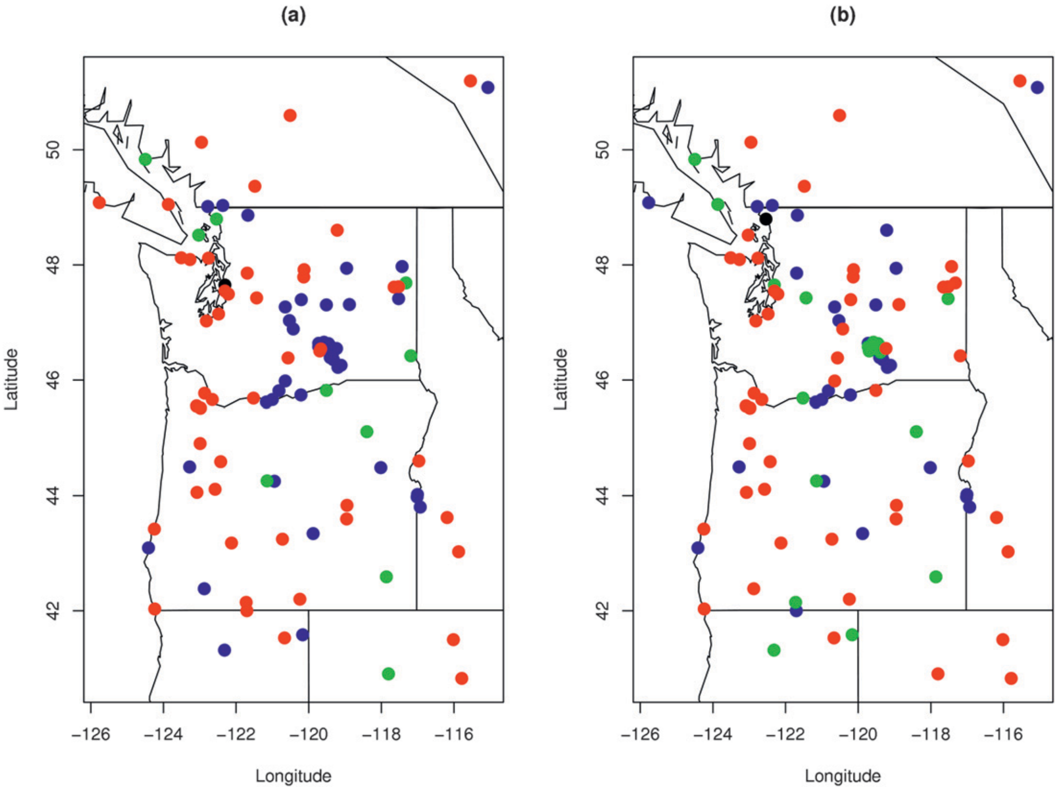

FIG. 4. Station-based CRPS with color corresponding to model with the lowest (best) value. Color codes are black for the raw ensemble, blue for Global BMA, green for Local BMA, and red for GMA. Scores are shown for (a) the sparse training network and (b) the dense training network.

suggest further improvement over Global BMA using a locally adaptive model. Indeed, GMA improves MAE at 54% of stations over Global BMA and improves the local CRPS at 55% of validation stations. Local BMA improves MAE and CRPS over Global BMA at 30% and 26% of validation stations, respectively. To assess local calibration and sharpness of the postprocessing models, Fig. 5 shows box plots for the validation-station-based coverage and average width for the nominal 80%, 90%, and 95% prediction intervals. Locally, we see that Local BMA is underdispersed at most stations while GMA and Global BMA are closer to nominal coverage at most validation stations. The box plots of average widths in the second row illustrate the effect of a constant predictive variance parameter as in the Global BMA model, all average prediction intervals are nearly of equal width at every station. GMA and Local BMA allow the prediction interval widths to vary substantially between locations, and both produce much narrower prediction intervals at most validation stations than Global BMA.

While CRPS takes account of both sharpness and calibration, the former may be checked directly by examining the width of the predictive intervals. Table 3 provides average width and coverage of the 80%, 90%, and 95% prediction intervals for Global BMA, Local BMA, and GMA. As the PIT histogram suggests, Global BMA produces globally calibrated prediction intervals; for instance the 80% interval covers the verifying value in 80.6% of cases. GMA yields only slightly underdispersed but significantly sharper prediction intervals, consistently narrowing each of the 80%, 90%, and 95% intervals by approximately 8% relative to Global BMA. Local BMA displays the underdispersion seen in its PIT histogram with the average interval coverages generally being 8%–10% lower than nominal, but narrows the predictive intervals by approximately 20% over Global BMA on average.

The greater accuracy of GMA than Global BMA displayed in the MAE is a result of a locally varying bias correction, while the better calibration and sharper prediction intervals are also influenced by the locally varying predictive variance. The ability to adjust parameters locally should result in better-calibrated distributions at each given site, which is not necessarily guaranteed using Global BMA. We examined the discrepancy criteria at each validation station, that is, the deviation of the PIT histogram from uniformity, and we

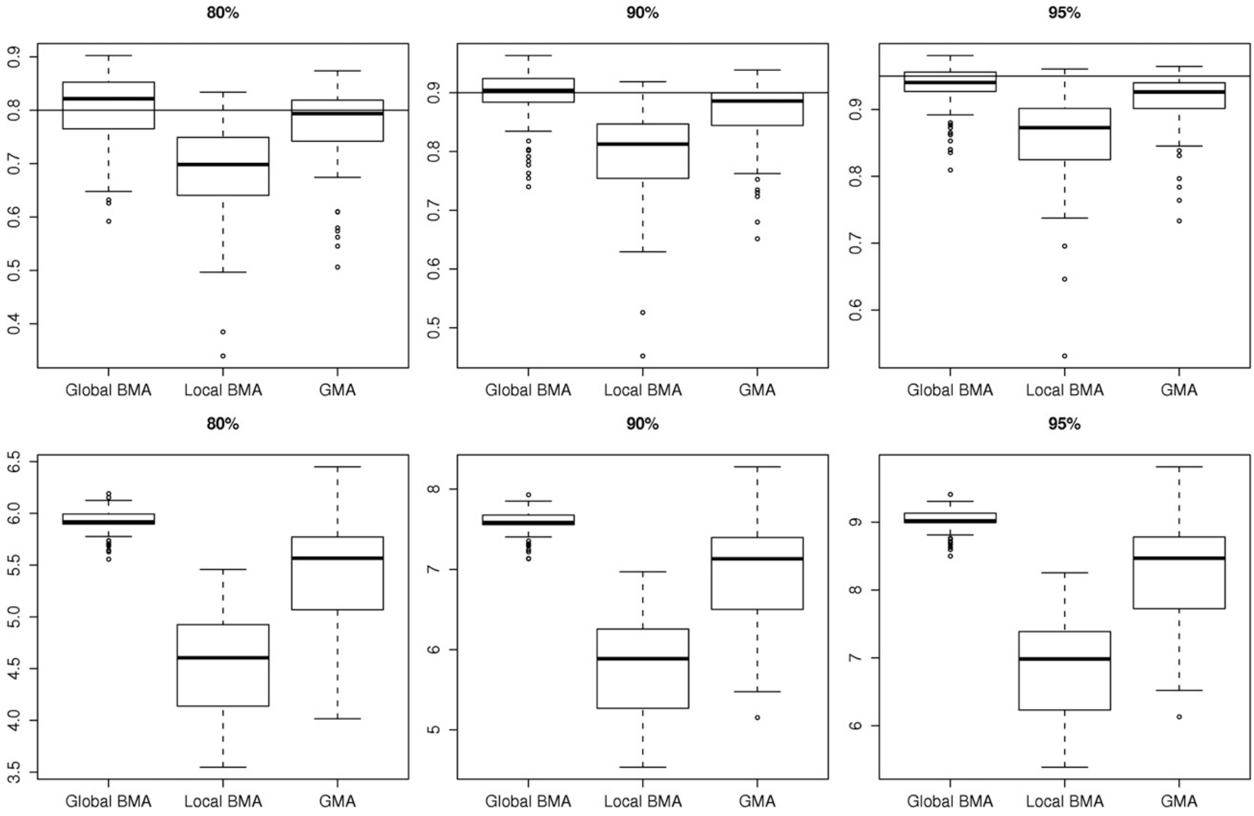

FIG. 5. Station-based prediction interval coverage and average width for Global BMA, Local BMA, and GMA for the nominal 80%, 90%, and 95% prediction intervals. The first row contains coverage while the second row shows predictive interval width. Coverages and width are calculated at each validation station separately and summarized by box plots. The box shows the interquartile range, while the whiskers extend to no more than 1.5 times the interquartile range.

found that GMA improved calibration over Global BMA in 66% of cases.

Local BMA suffers from the lack of available information using the sparse training network, and must interpolate forecast errors from great distances because of the sparsity of fitting stations. In fact, Local BMA was available at only 85 of the 100 validation sites; we substituted in Global BMA's predictive density at the remaining 15 locations where Local BMA was not available.

b. *Dense network*

For the dense network, we allow any station available during the 2005–06 time period to be included as a fitting station. This yields 1457 fitting stations on land, with an additional 263 water-based stations; we use the same 100 validation stations as for the sparse network, which are held out of the training set.

Table 2 displays the MAE and CRPS scores for the dense network experiment. The order of performance of the three methods is the same as for the sparse method, with GMA best, followed by Global BMA and then Local BMA. The standard errors for differences between Global BMA and GMA or Local BMA are the same as in the sparse experiment (i.e., 0.003°C for CRPS and 0.005°C for MAE). In this case, the differences in CRPS between all three methods are statistically

TABLE 3. Average prediction interval coverage and width (in °C) for 2006.

	Sparse network						Dense network					
	Coverage (%)			Width (°C)			Coverage (%)			Width (°C)		
	80%	90%	95%	80%	90%	95%	80%	90%	95%	80%	90%	95%
Global BMA	80.6	89.4	93.5	5.92	7.59	9.03	81.9	90.2	94.1	6.17	7.91	9.41
Local BMA	70.0	80.5	86.7	4.77	6.08	7.20	68.6	79.4	85.7	4.64	5.92	7.00
GMA	77.2	86.6	91.4	5.45	6.98	8.30	77.4	86.8	91.8	5.50	7.06	8.39

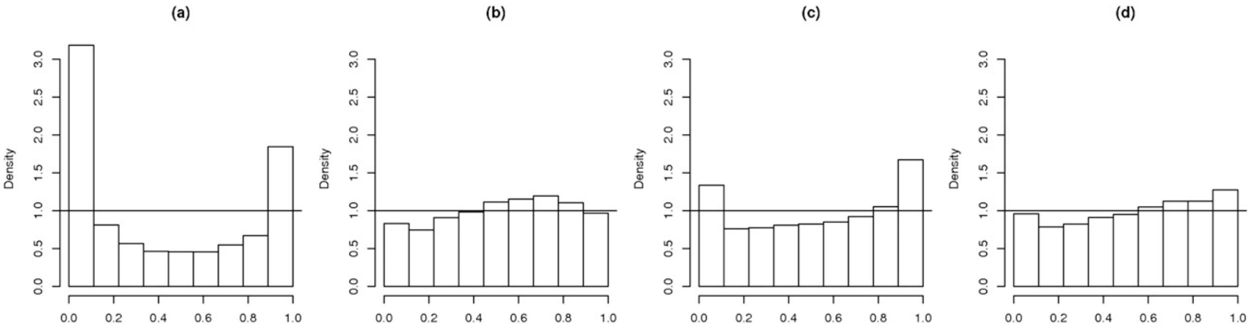

FIG. 6. Dense network experiment: (a) Rank histogram for the raw ensemble, and PIT histograms for (b) Global BMA, (c) Local BMA, and (d) GMA.

significant, but unlike the sparse experiment, the MAE is not significantly different between Local BMA and Global BMA. Hence, the skill of Local BMA has improved over the sparse experiment, largely because of the higher density of training stations. Whereas Local BMA was available at 85 of the 100 validation stations in the sparse experiment, it is available at 99 in the dense network case.

We examine local CRPS values for the dense training network in Fig. 4b, where Global BMA is best at 31 stations, and the locally adaptive methods Local BMA and GMA are superior at 19 and 49 stations, respectively. Local calibration can be assessed similar to Fig. 5 for the sparse training network, and the results are similar to those of Fig. 5 (not shown). While Global BMA shows aggregate calibration, GMA displays local calibration and substantially shrinks the predictive interval widths. For this network, GMA improves the local CRPS value at 57% of stations over Global BMA, while Local BMA improves local CRPS at 45% of stations over Global BMA.

The PIT histograms for this training network are displayed in Fig. 6. Local BMA still displays some underdispersion, reflected in Table 3, with lower coverage percentages and narrower prediction intervals. Global BMA is now slightly overdispersed. The PIT histogram for GMA does not indicate either overdispersion or underdispersion.

Figure 7 shows the predictive temperature fields for Global BMA, GMA, and Local BMA valid on 22 May 2006. Each point is the median of the predictive distribution at the corresponding grid location. Global BMA and GMA are available everywhere, but Local BMA is not available in areas of sparse observations. Local BMA and GMA both decrease the median forecasts compared to Global BMA in the Puget Sound region east of the Olympic Peninsula in Washington, and Local BMA shows greater adjustment. The observations in the Puget Sound region are generally cooler than Global BMA's median forecast on 22 May 2006, while GMA and Local BMA identify and adjust for the local bias.

6. Case study

We now present four stations in detail: Peony Creek, Washington; Roberts Field, Oregon; Franklin Falls, Washington; and Friday Harbor Airport, Washington. The following results are under the conditions of the sparse training network.

a. Peony Creek

The first case study station is at Peony Creek in northeast Washington State; see Fig. 8. The black time series in Fig. 9a consists of the forecast errors of the GFS ensemble member forecast at Peony Creek in 2006. The red line is the bias correction for the GFS member equal to the kriged value of a_s from GMA, the blue line represents Global BMA's bias correction, while the green line is Local BMA's bias correction. The bias correction of Global BMA is constant across all sites. Figure 9b illustrates GMA's ability to adapt to bias locally, predicting the local bias at Peony Creek significantly more accurately than Global BMA.

Similar behavior is seen in the predictive standard deviation, shown in Fig. 9c. Global BMA's predictive standard deviation was nearly constant across 2006. In contrast, GMA's predictive standard deviation was able to identify and adapt to periods of changing uncertainty, as indicated by the vertical dashed line. Local BMA is not available every day that GMA and Global BMA are; there is a slight gap at day 54 in the bias correction and predictive standard deviation. On these days there were not enough nearby stations that met the requirements of the Mass–Baars interpolation scheme (see section 4).

The stations on which Local BMA's interpolation is based change from day to day, while GMA uses every available station. To get a sense of which stations were chosen on a particular day, see Fig. 8. On 22 May 2006,

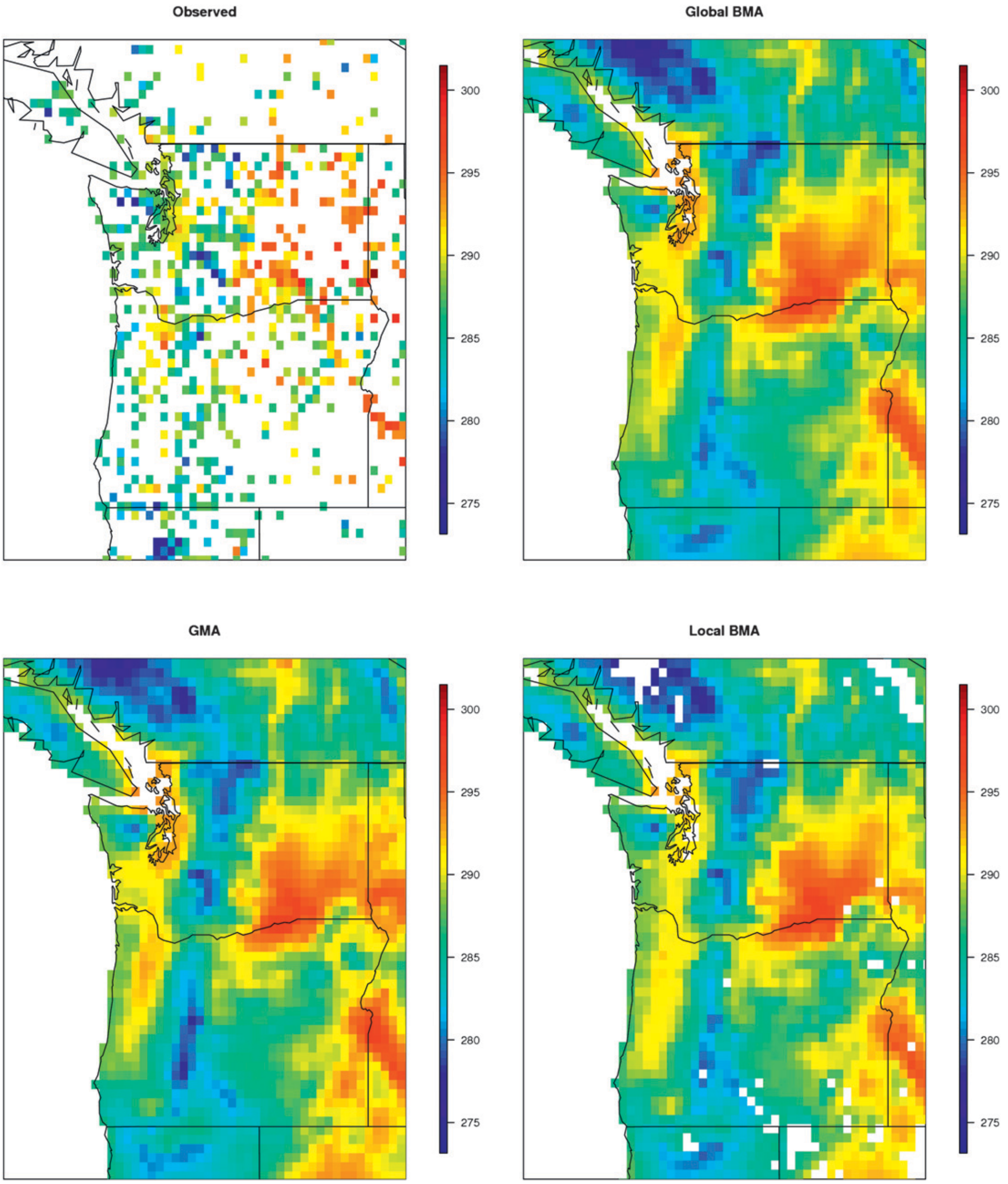

FIG. 7. Median of predictive density on 22 May 2006 for Global BMA, GMA, and Local BMA using the dense network for training, with observed values.

eight stations were available to interpolate to the nearest grid point adjacent to Peony Creek and Roberts Field for Local BMA; these are represented by green points. The eight nearest stations used by GMA are colored red, while the stations shared by both are in purple. At Peony Creek, the biases from the nearby stations used by GMA reflect the empirical bias more accurately than those stations chosen by Local BMA, as seen in Fig. 9b. This is partially due to the fact that two of the nearest stations selected by GMA are of the same land type as that at Peony Creek, namely forest, while the land type of the stations chosen by Local BMA (which are of the same land type as the nearest grid point to Peony Creek but not as Peony Creek itself) are grassland.

Calibration may again be assessed by looking at the PIT histograms generated by predictive distributions at Peony Creek for the year 2006; these are displayed in Fig. 10. As expected, the raw ensemble is underdispersed, while the PIT histogram for Global BMA

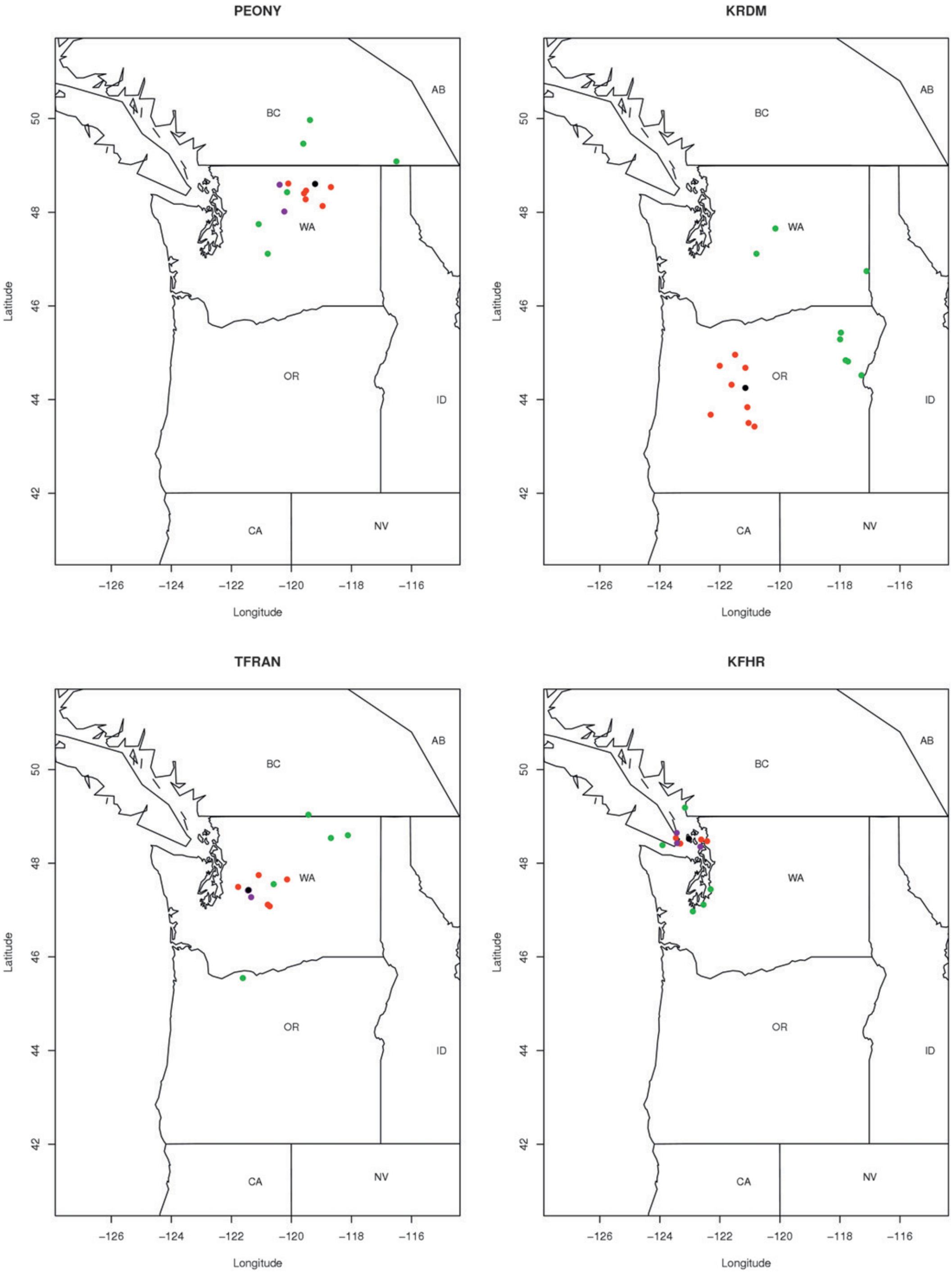

FIG. 8. The eight nearest training stations chosen by GMA in red and Local BMA in green on 22 May 2006, with shared stations in purple. The case study stations are in black with corresponding codes PEONY for Peony Creek, KRDM for Roberts Field, TFRAN for Franklin Falls, and KFHR for Friday Harbor airport.

puts more mass at higher values, suggesting a tendency to underpredict at Peony Creek. Local BMA improves the calibration of Global BMA, but also has the tendency to underpredict. The PIT histogram for GMA shows much better calibration, which is a result of locally accurate bias correction and predictive variance.

Predictive densities for 11 September 2006 and 28 December 2006 are shown in Fig. 11. The ensemble

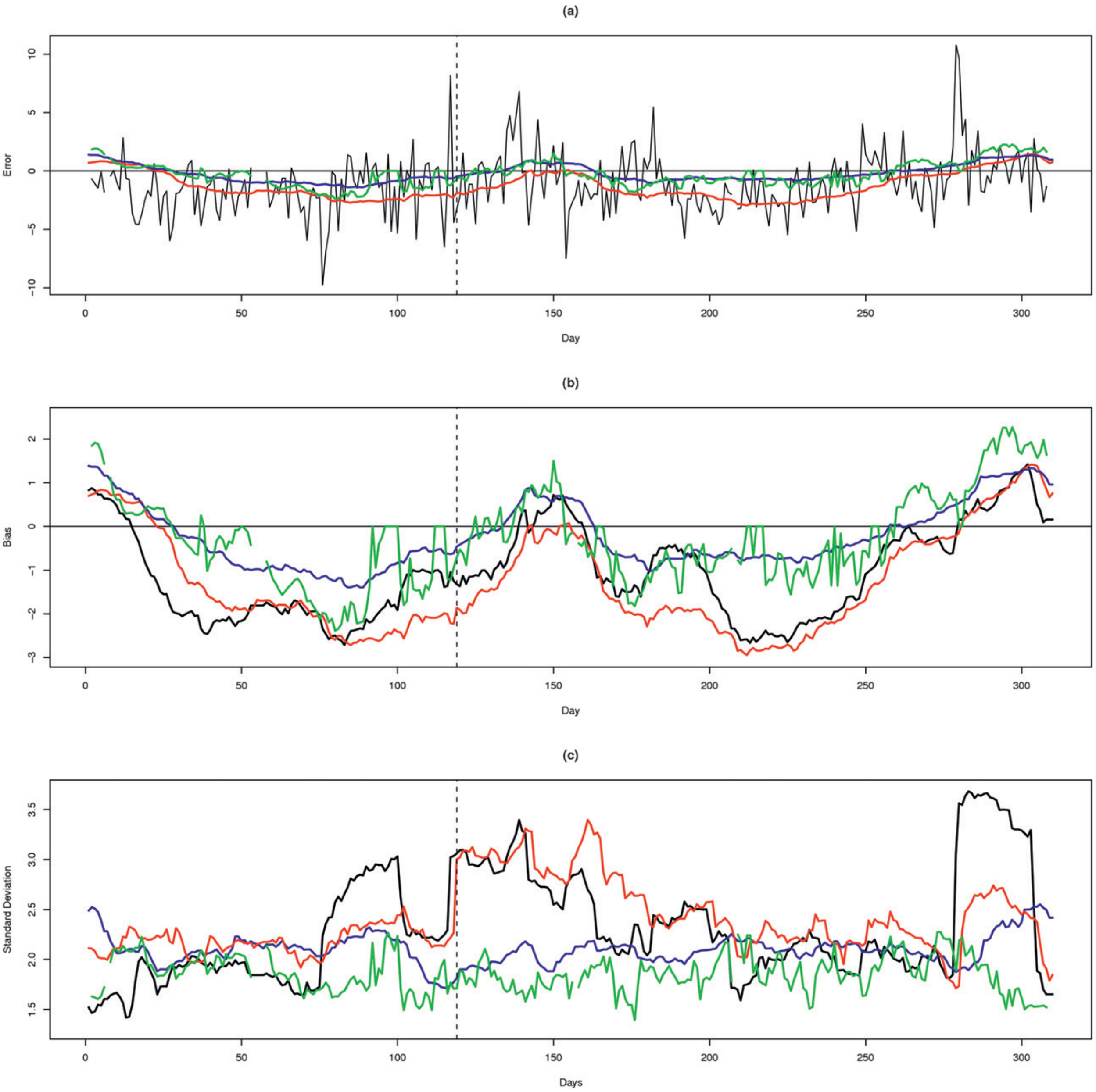

FIG. 9. (a) Time series of GFS member forecast errors (forecast minus observation) for 2006 at Peony Creek in black, with bias corrections from Global BMA (blue), Local BMA (green), and GMA (red). (b) Empirical bias of GFS member at Peony Creek in black with bias corrections from Global BMA (blue), Local BMA (green), and GMA (red). (c) Empirical standard deviation of GFS member forecast errors at Peony Creek in black with predictive standard deviations from Global BMA (blue), Local BMA (green), and GMA (red). The vertical dashed line marks the beginning of a period of greater predictive uncertainty, as seen in the empirical standard deviation.

spread does not capture the realizing value in either case. On 11 September 2006 we see the local bias correction of GMA shifting the predictive distribution toward higher temperatures, centering exactly around the realizing value. In both cases, GMA's 80% predictive interval was much narrower than that of Global BMA. This is most easily seen on 28 December where GMA's interval was completely contained within that of Global BMA, while both intervals captured the realizing value.

Table 4 shows the MAE and CRPS for the raw ensemble, Global BMA, GMA, and Local BMA at Peony Creek for the 2006 calendar year. Global BMA performed

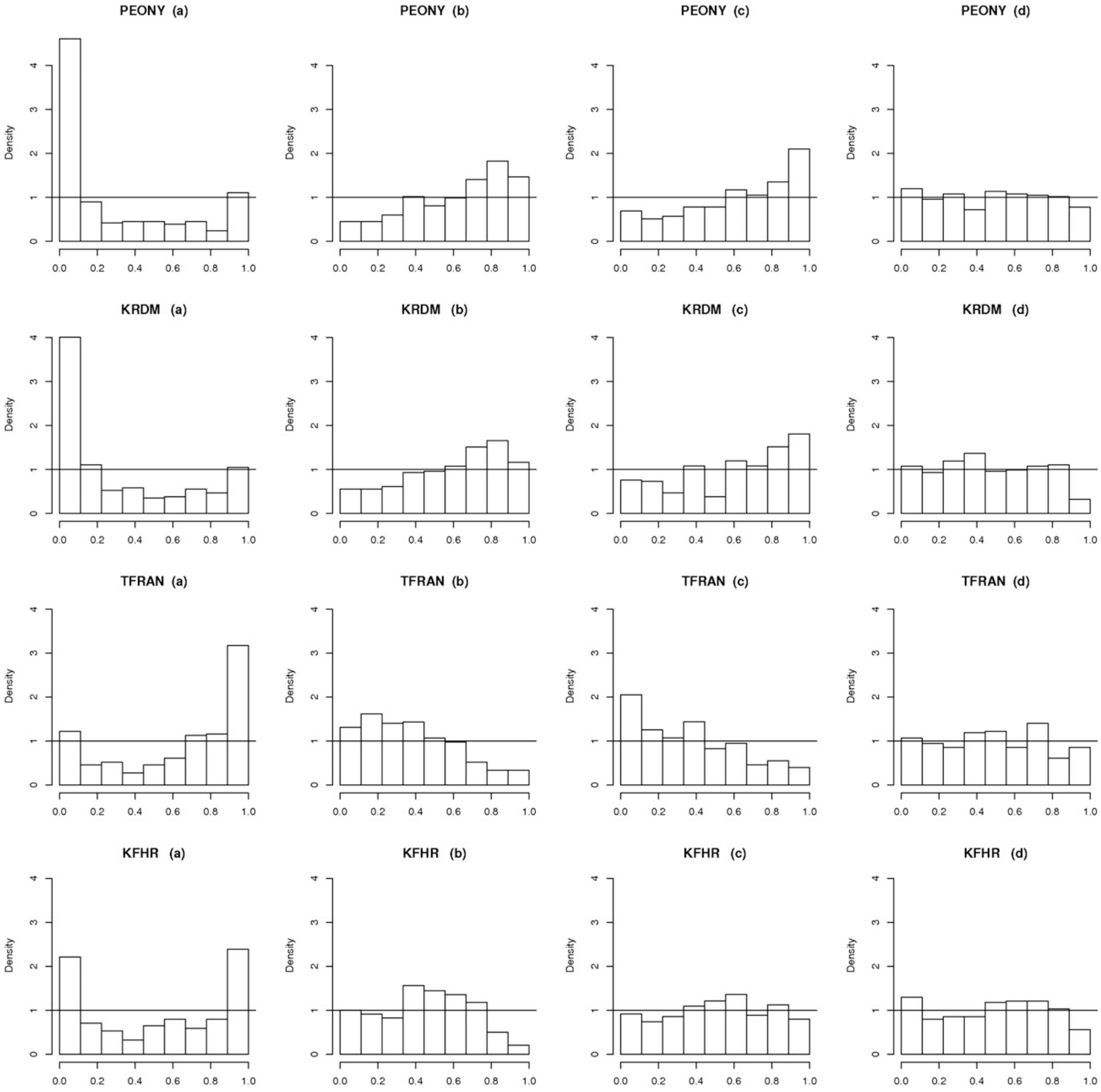

FIG. 10. (a) Rank histogram for the raw ensemble, and PIT histograms for (b) Global BMA, (c) Local BMA, and (d) GMA. Each row is a case study station with codes PEONY for Peony Creek, KRDM for Roberts Field, TFRAN for Franklin Falls, and KFHR for Friday Harbor airport.

better than the raw ensemble, while GMA showed further improvement. Local BMA did not improve over the global model; this is due to a lack of available training stations nearby whose forecast errors follow the same behavior as the grid points surrounding Peony Creek. GMA was calibrated and at the same time provided sharper prediction intervals, narrowing the average interval width of Global BMA by 7.6%, as seen in Table 5. This table also indicates that Local BMA was underdispersed at these three intervals.

b. Roberts Field

Our second case study station is at Roberts Field for which Fig. 12 shows empirical errors, biases, and predictive standard deviations for the 2006 calendar year. The situation here is the opposite of that at Peony Creek. GMA pulled information mainly from the stations shown in Fig. 8, which are geographically closer to Roberts Field, but the information interpolated by Local BMA from the further stations represent the

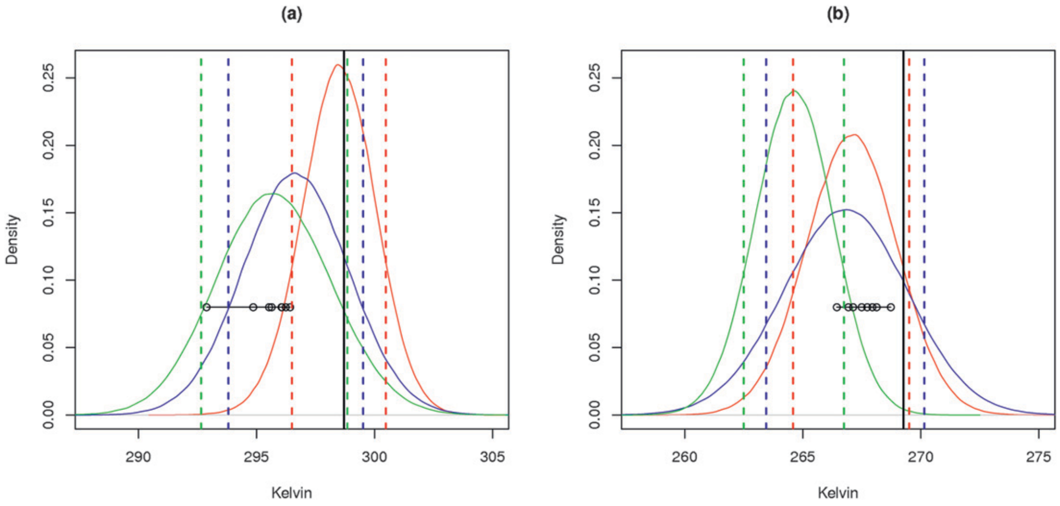

FIG. 11. Predictive densities for (a) 11 Sep 2006 and (b) 28 Dec 2006 at Peony Creek. The blue density is Global BMA, the red density is GMA, the green density is Local BMA, the black vertical line is the verifying observation, while the dashed vertical lines correspond to the 80% prediction intervals. The dots show the eight ensemble member forecasts, with a horizontal line to illustrate ensemble spread.

behavior of the GFS member forecast errors at Roberts Field more accurately. This can be seen in Fig. 12b, where the bias estimation by GMA is consistently too low during most of 2006, while Local BMA's correction tends to follow the empirical trend more closely. The interpolated standard deviation of GMA agrees with the empirical variability of the GFS member, similarly to the situation at Peony Creek.

The PIT histograms in Fig. 10 show that the raw forecast tended to overpredict, Global BMA usually underpredicted, and GMA and Local BMA were better calibrated. Local BMA was slightly underdispersed, as seen in Table 5, but it gave significantly sharper prediction intervals at Roberts Field than Global BMA and GMA, by approximately 22% on average. This is reflected in the superior CRPS score for Local BMA, see Table 4.

Local BMA performed well at Roberts Field because there were appropriate training stations nearby for it to select. Indeed, all 8 stations chosen by Local BMA were within 280 m of the elevation at Roberts Field, and these stations were characterized as grassland, the same land use as Roberts Field. GMA chose stations of up to 507 m different in elevation, and only 3 of the nearest locations were grassland, while the other 5 were either cropland or

TABLE 4. MAE and mean CRPS (both in °C) for the raw ensemble, Global BMA, GMA, and Local BMA over the calendar year 2006 at four stations with codes PEONY for Peony Creek, KRDM for Roberts Field, TFRAN for Franklin Falls, and KFHR for Friday Harbor airport.

	Model	MAE	CRPS
PEONY	Raw ensemble	2.007	1.703
	Global BMA	1.877	1.339
	Local BMA	1.950	1.400
	GMA	1.760	1.283
KRDM	Raw ensemble	1.948	1.550
	Global BMA	1.843	1.303
	Local BMA	1.804	1.274
	GMA	1.785	1.289
TFRAN	Raw ensemble	1.716	1.400
	Global BMA	1.750	1.262
	Local BMA	1.732	1.243
	GMA	1.552	1.135
KFHR	Raw ensemble	1.404	1.132
	Global BMA	1.400	1.054
	Local BMA	1.373	0.992
	GMA	1.375	1.009

TABLE 5. Average prediction interval coverage and width at four stations with codes PEONY for Peony Creek, KRDM for Roberts Field, TFRAN for Franklin Falls, and KFHR for Friday Harbor airport.

		Coverage (%)			Width (°C)		
		80%	90%	95%	80%	90%	95%
PEONY	Global BMA	82.4	89.7	94.0	5.81	7.46	8.88
	Local BMA	71.7	85.3	91.0	5.32	6.81	8.08
	GMA	81.4	89.0	92.4	5.37	6.89	8.20
KRDM	Global BMA	84.2	91.9	95.2	6.04	7.72	9.17
	Local BMA	74.4	85.8	90.6	4.74	6.02	7.12
	GMA	86.1	90.6	92.9	6.03	7.72	9.15
TFRAN	Global BMA	82.0	89.8	94.2	5.92	7.59	9.03
	Local BMA	73.8	85.7	91.5	5.02	6.40	7.58
	GMA	80.0	87.5	92.5	5.14	6.60	7.85
KFHR	Global BMA	88.5	93.1	96.4	5.71	7.32	8.72
	Local BMA	81.9	91.1	96.1	4.74	6.06	7.20
	GMA	81.6	87.2	93.1	4.31	5.52	6.57

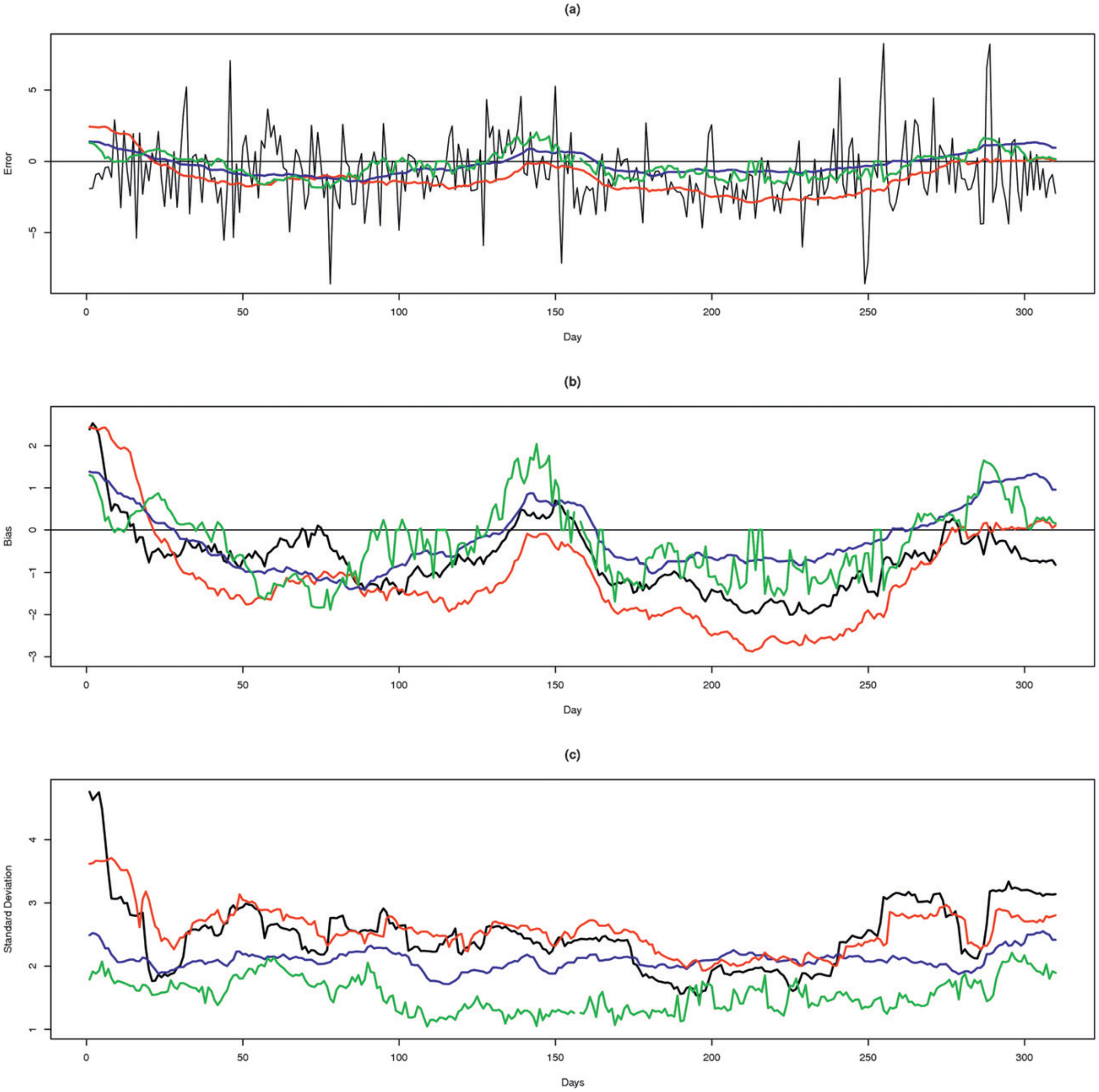

FIG. 12. (a) Time series of GFS member forecast errors (forecast minus observation) for 2006 at Roberts Field in black, with bias corrections from Global BMA (blue), Local BMA (green), and GMA (red). (b) Empirical bias of GFS member at Roberts Field in black with bias corrections from Global BMA (blue), Local BMA (green), and GMA (red). (c) Empirical standard deviation of GFS member forecast errors at Roberts Field in black with predictive standard deviations from Global BMA (blue), Local BMA (green), and GMA (red).

forest, leading to a poorer estimate of the forecast bias at Roberts Field.

However, Local BMA's approach to station selection can also lead to difficulty in obtaining predictions. Of the 100 validation stations used in section 5, only 85 had sufficient training stations nearby to give a Local BMA predictive distribution. This behavior in areas of a sparse observation network is overcome by GMA, which is available everywhere, while still locally adjusting forecasts.

c. Franklin Falls and Friday Harbor Airport

The final two case study stations serve to illustrate the applicability of GMA and Local BMA to regions of complex terrain. Franklin Falls is only a few hundred meters away from Snoqualmie Pass in the Cascade Mountain range, and sits at an elevation of 829 m. Indeed the sensitivity of both methods to elevation is important in the mountains, and Table 4 displays the importance of local adjustments in different types of terrain. Both methods produce significantly sharper prediction intervals that are calibrated for GMA and only slightly undercalibrated for Local BMA (see Table 5).

Friday Harbor Airport is located on the San Juan Island chain in the Puget Sound. The airport itself is only approximately 200 m from the ocean, and experiences vastly different climate than the three other case study stations. The local methods improve the point estimate of the median of the predictive distribution over Global BMA, but all are seen to have comparable CRPS values (see Table 4). Notice that Global BMA is slightly overdispersed at Friday Harbor Airport, while GMA and Local BMA display good calibration, and narrow the predictive interval widths (at all levels) relative to Global BMA by 17% on average for Local BMA and by 25% on average for GMA (seen in Table 5). Both local methods choose coastal stations, as shown in Fig. 8, for improving local bias and predictive variance, while Global BMA equally weights all coastal, mountain, grassland, cropland, forest, and other land-type-based stations for parameter estimation at Friday Harbor Airport.

7. Discussion

We have presented two different ways to produce locally calibrated probabilistic grid-based forecasts, using station observations. Local calibration refers to statistical calibration at an individual location, while global calibration refers to calibration over all locations. Both models are based on the Bayesian model averaging method of Raftery et al. (2005). Geostatistical model averaging first estimates the statistical parameters of the BMA model at each station and then interpolates the parameter estimates using a geostatistical model. Local BMA estimates the bias at a grid point as an average of the observed forecast errors at stations that are close to the grid point and have similar elevation and land use.

The two methods have advantages and disadvantages. GMA yields local adjustments everywhere, putting the most weight on information from the nearest stations, irrespective of land use or other characteristics. Local BMA uses only stations with similar land type, elevation, and forecast value, and thus provides a more physical interpolation; however, it has the disadvantage of not always being available everywhere. This problem is related to station density—for a dense observation network, Local BMA performs well at locations with sufficiently many nearby stations, while GMA is better adapted to a sparse network. However, occasionally the similarity of bias in a certain land type reflects the true bias at a model grid point more so than the nearest stations; in these situations Local BMA will estimate the local parameters more accurately than GMA.

There is no guarantee that the predictive distribution from a global model at any single location will be calibrated, whereas locally adaptive procedures can produce locally calibrated forecasts. GMA was locally calibrated and sharp, and Local BMA was significantly sharper than Global BMA and GMA on average, but was underdispersed.

Our example of 48-h temperature forecasts illustrates a strength of GMA relative to the Global BMA model. The predictive densities generated by both methods lead to calibrated probabilistic forecasts. However, GMA yields sharper predictive densities than Global BMA, due to the locally varying predictive variance. At the second case study station, Roberts Field, Oregon, we see Global BMA and GMA producing essentially equivalent sharpness, where Local BMA yields much sharper densities, narrowing the predictive intervals as much as 22%. Both methods perform well and improve over Global BMA in areas of complex terrain, as seen with the final two case study stations: one in the Cascade Mountains and the other on an island in the Puget Sound.

The example section focused on 48-h temperature forecasts, but we also evaluated the models' 36- and 42-h forecasts, finding similar results to those described above. In particular, GMA and Local BMA substantially reduced the predictive interval width over Global BMA, but GMA displayed slight underdispersion and Local BMA had more pronounced underdispersion than at the 48-h horizon. For 36- and 42-h forecasts, GMA reduced the domain aggregated MAE and CRPS over both Global BMA and the raw ensemble. Local BMA reduced the domain-aggregated scores over Global BMA, except for the CRPS in the 36-h experiment.

We have specified GMA in terms of a specific geostatistical model with an exponential covariance function. However, any other geostatistical model could be used, and other specifications could provide better performance. Also, GMA and Local BMA both have strengths, and it might be profitable to combine them, for example by restricting the estimation of GMA parameters for a grid point to stations that are similar to the grid point in location, elevation, and land use.

Local BMA relies on the Mass–Baars interpolation algorithm, whose interpolation parameters are estimated by minimizing the domain-aggregated mean absolute error. The parameters used in our experiments are described in Mass et al. (2008). However, these parameters are likely to differ depending on the region of interest, so reoptimization is likely required in a different setting. It also may be profitable to consider allowing these parameters to vary by location. For example, in areas of dense observations the forecaster may want stricter rules on what errors are interpolated, while in areas of sparse network coverage, there may be benefit to relaxing the forecast error selection criteria.

Gridded bias correction is built into both GMA and Local BMA. Gel (2007) investigated two other approaches to gridded bias correction that are similar to our methods. The first, local observation-based (LOB) bias removal, defines neighborhoods based on the spatial structure of historical biases at observing locations. However, rather than using all available bias information, Gel (2007) defined a neighborhood based on the spatial parameters. Her other method is a nonlinear regression method that uses classification and regression trees (CART) and alternating conditional expectations (ACE). Gel (2007) pointed out that her CART–ACE method works well in areas of sparse data, but it requires a long training period and is not fully automated. Both LOB and CART–ACE only result in deterministic gridded bias estimates, rather than a calibrated probabilistic prediction.

Acknowledgments. This research was sponsored by the Department of Defense MURI Program (N00014-01-1-0745), the National Science Foundation under Joint Ensemble Forecast System (JEFS) Subawards S06-47225 and S10-75538 with the University Corporation for Atmospheric Research (UCAR), as well as National Weather Service CSTAR program (Grant 2073958) and Grants ATM-0724721 and DMS-0706745.

APPENDIX

EM Algorithm

We describe the EM algorithm of section 3. The EM algorithm is an iterative process that finds a maximum of the log likelihood function:

$$\ell(w_1,\ldots,w_K,c)=\sum_{s,t}\log\left[\sum_{\ell=1}^{K}w_\ell g(y_{st}|f_{\ell st})\right]. \qquad (A1)$$

One cannot guarantee that the algorithm converges to a global maximum, but using different initial values can assist in avoiding local maxima.

We introduce latent variables $z_{\ell st}$, which can be thought of as being 1 if the ℓth forecast is best for site s and time t, and 0 otherwise. In the jth E (expectation) step of the EM algorithm, we start with current estimates $\{w_\ell^{(j)}\}_{\ell=1}^{K}$ and $c^{(j)}$, and, based on these, calculate estimates

$$\hat{z}_{kst}^{(j)}=\frac{w_k^{(j)}g^{(j)}(y_{st}|f_{kst})}{\displaystyle\sum_{\ell=1}^{K}w_\ell^{(j)}g^{(j)}(y_{st}|f_{\ell st})} \qquad (A2)$$

for every ensemble member, where $g^{(j)}(y_{st}|f_{\ell st})$ is a normal density with mean $f_{\ell st}-\hat{a}_{\ell s}$ and variance $c^{(j)}\exp(\hat{v}_s)$. In the M (maximization) step, we update the weights via

$$w_k^{(j+1)}=\frac{1}{N}\sum_{s,t}\hat{z}_{kst}^{(j)} \qquad (A3)$$

and c via

$$c^{(j+1)}=\frac{1}{N}\sum_{s,t}\frac{1}{\exp(\hat{v}_s)}\sum_{\ell=1}^{K}\hat{z}_{kst}^{(j)}[y_{st}-(f_{\ell st}-\hat{a}_{\ell s})]^2, \qquad (A4)$$

where N is the number of station and time pairs, and $\hat{a}_{\ell s}$ and $\exp(\hat{v}_s)$ depend on time t, though we have not included the subscript t here for simplicity. The algorithm is iterated between the E and M steps until some stopping criterion is reached.

REFERENCES

Berrocal, V. J., A. E. Raftery, and T. Gneiting, 2007: Combining spatial statistical and ensemble information in probabilistic weather forecasts. *Mon. Wea. Rev.,* **135,** 1386–1402.

——, A. E. Gelfand, and D. M. Holland, 2009: A spatio-temporal downscaler for output from numerical models. *J. Agric. Biol. Environ. Stat.,* **15,** 176–197.

——, ——, and ——, 2010: A bivariate-space time downscaler under space and time misalignment. *Annu. Appl. Stat.,* **4,** 1942–1975.

Bröcker, J., and L. Smith, 2008: From ensemble forecasts to predictive distribution functions. *Tellus,* **60A,** 663–678.

Buizza, R., P. L. Houtekamer, Z. Toth, G. Pellerin, M. Wei, and Y. Zhu, 2005: A comparison of the ECMWF, MSC, and NCEP global ensemble prediction systems. *Mon. Wea. Rev.,* **133,** 1076–1097.

Byrd, R. H., P. Lu, J. Nocedal, and C. Zhu, 1995: A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Comput.,* **16,** 1190–1208.

Chilès, J. P., and P. Delfiner, 1999: *Geostatistics: Modeling Spatial Uncertainty*. Wiley, 695 pp.

Cressie, N. A. C., 1993: *Statistics for Spatial Data*. rev. ed. Wiley, 900 pp.

Cressman, G. P., 1959: An operational objective analysis system. *Mon. Wea. Rev.,* **87,** 367–374.

Dee, D. P., 2005: Bias and data assimilation. *Quart. J. Roy. Meteor. Soc.,* **131,** 3323–3344.

Dempster, A. P., N. M. Laird, and D. B. Rubin, 1977: Maximum likelihood from incomplete data via EM algorithm. *J. Roy. Stat. Soc.,* **39B,** 1–38.

Diebold, F. X., T. A. Gunther, and A. S. Tay, 1998: Evaluating density forecasts with applications to financial risk management. *Int. Econ. Rev.,* **39,** 863–883.

Duan, Q., N. K. Ajami, X. Gao, and S. Sorooshian, 2007: Multimodel ensemble hydrologic prediction using Bayesian model averaging. *Adv. Water Resour.,* **30,** 1371–1386.

Eckel, F. A., and C. F. Mass, 2005: Aspects of effective mesoscale short-range ensemble forecasting. *Wea. Forecasting,* **20,** 328–350.

Epstein, E. S., 1969: Stochastic dynamic prediction. *Tellus,* **21,** 739–759.

Fortin, V., A. C. Favre, and M. Said, 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,** 1349–1370.

Gel, Y. R., 2007: Comparative analysis of the local observation-based (LOB) method and the nonparametric regression-based method for gridded bias correction in mesoscale weather forecasting. *Wea. Forecasting,* **22,** 1243–1256.

Gelfand, A. E., H. J. Kim, C. F. Sirmans, and S. Banerjee, 2003: Spatial modeling with spatially varying coefficient processes. *J. Amer. Stat. Assoc.,* **98,** 387–396.

——, S. Banerjee, and D. Gamerman, 2005: Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics,* **16,** 465–479.

Glahn, B., K. Gilbert, R. Cosgrove, D. P. Ruth, and K. Sheets, 2009a: The gridding of MOS. *Wea. Forecasting,* **24,** 520–529.

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

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, 2007: Strictly proper scoring rules, prediction, and estimation. *J. Amer. Stat. Assoc.,* **102,** 359–378.

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

——, F. Balabdaoui, and A. E. Raftery, 2007: Probabilistic forecasts, calibration and sharpness. *J. Roy. Stat. Soc.,* **69B,** 243–268.

Grimit, E. P., T. Gneiting, V. J. Berrocal, and N. A. Johnson, 2006: The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quart. J. Roy. Meteor. Soc.,* **132,** 2925–2942.

Hacker, J. P., and D. L. Rife, 2007: A practical approach to sequential estimation of systematic error on near-surface mesoscale grids. *Wea. Forecasting,* **22,** 1257–1273.

Hagedorn, R., T. M. Hamill, and J. S. Whitaker, 2008: Probabilistic forecast calibration using ECMWF and GFS ensemble reforecasts. Part I: Two-meter temperatures. *Mon. Wea. Rev.,* **136,** 2608–2619.

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.

——, C. Snyder, and R. E. Morss, 2000: A comparison of probabilistic forecasts from bred, singular-vector, and perturbed observation ensembles. *Mon. Wea. Rev.,* **128,** 1835–1851.

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

Hastie, T., and R. Tibshirani, 1993: Varying-coefficient models. *J. Roy. Stat. Soc.,* **55B,** 757–796.

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

Houtekamer, P. L., and J. Derome, 1995: Methods for ensemble prediction. *Mon. Wea. Rev.,* **123,** 2181–2196.

Johnson, C., and R. Swinbank, 2009: Medium-range multimodel ensemble combination and calibration. *Quart. J. Roy. Meteor. Soc.,* **135,** 777–794.

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

Leith, C. E., 1974: Theoretical skill of Monte Carlo forecasts. *Mon. Wea. Rev.,* **102,** 409–418.

Liu, Z., N. Le, and J. V. Zidek, 2008: Combining measurements and physical model outputs for the spatial prediction of hourly ozone space–time fields. Department of Statistics, University of British Columbia, Tech. Rep. 239, 23 pp. [Available online at http://www.stat.ubc.ca/Research/TechReports/techreports/239.pdf.]

Mass, C. F., J. Baars, G. Wedam, E. Grimit, and R. Steed, 2008: Removal of systematic model bias on a model grid. *Wea. Forecasting,* **23,** 438–459.

——, and Coauthors, 2009: PROBCAST: A Web-based portal to mesoscale probabilistic forecasts. *Bull. Amer. Meteor. Soc.,* **90,** 1009–1014.

Matheson, J. E., and R. L. Winkler, 1976: Scoring rules for continuous probability distributions. *Manage. Sci.,* **22,** 1087–1096.

Molteni, R., 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.

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.

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

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

——, T. Gneiting, and A. E. Raftery, 2010: Probabilistic wind speed forecasting using ensembles and Bayesian model averaging. *J. Amer. Stat. Assoc.,* **105,** 25–35.

Smith, R. L., C. Tebaldi, D. Nychka, and L. O. Mearns, 2009: Bayesian modeling of uncertainty in ensembles of climate models. *J. Amer. Stat. Assoc.,* **104,** 97–116.

Stein, M. L., 1999: *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, 247 pp.

Stensrud, D. J., H. E. Brooks, J. Du, M. S. Tracton, and E. Rogers, 1999: Using ensembles for short-range forecasting. *Mon. Wea. Rev.,* **127,** 433–446.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Soc. London,* **365,** 2053–2075.

Toth, Z., and E. Kalnay, 1993: Ensemble forecasting at NMC: The generation of perturbations. *Bull. Amer. Meteor. Soc.,* **74,** 2317–2330.

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

Wang, X., and C. H. Bishop, 2005: Improvement of ensemble reliability with a new dressing kernel. *Quart. J. Roy. Meteor. Soc.,* **131,** 965–986.

Wilks, D. S., 2009: Extending logistic regression to provide full-probability-distribution MOS forecasts. *Meteor. Appl.,* **16,** 361–368.

——, and T. M. Hamill, 2007: Comparison of ensemble-MOS methods using GFS reforecasts. *Mon. Wea. Rev.,* **135,** 2379–2390.

Yussouf, N., and D. J. Stensrud, 2006: Prediction of near-surface variables at independent locations from a bias-corrected ensemble forecasting system. *Mon. Wea. Rev.,* **134,** 3415–3424.