Geostatistical Model Averaging for Locally Calibrated Probabilistic Quantitative Precipitation Forecasting

William KLEIBER, Adrian E. RAFTERY, and Tilmann GNEITING

Accurate weather benefit many key societal functions and activities, including agriculture, transportation, recreation, and basic human and infrastructural safety. Over the past two decades, ensembles of numerical weather prediction models have been developed, in which multiple estimates of the current state of the atmosphere are used to generate probabilistic forecasts for future weather events. However, ensemble systems are uncalibrated and biased, and thus need to be statistically postprocessed. Bayesian model averaging (BMA) is a preferred way of doing this. Particularly for quantitative precipitation, biases and calibration errors depend critically on local terrain features. We introduce a geostatistical approach to modeling locally varying BMA parameters, as opposed to the extant method that holds parameters constant across the forecast domain. Degeneracies caused by enduring dry periods are overcome by Bayesian regularization and Laplace approximations. The new approach, called geostatistical model averaging (GMA), was applied to 48-hour-ahead forecasts of daily precipitation accumulation over the North American Pacific Northwest, using the eight-member University of Washington Mesoscale Ensemble. GMA had better aggregate and local calibration than the extant technique, and was sharper on average.

KEY WORDS: Bayesian model averaging; Calibration; Ensemble prediction system; Gaussian process; Laplace approximation; Numerical weather prediction; Probabilistic forecast; Regularization.

1. INTRODUCTION

Scientists have been forecasting the weather for well over a century, and following the advent of numerical weather prediction models, the forecasts have become increasingly accurate. However, even with state-of-the-art numerical models, there are still significant uncertainties in the forecasts. Imperfect representations of atmospheric physics, incomplete initial and boundary conditions, and imperfect numerical schemes all lead to point forecasts that are uncertain. Understanding and quantifying the uncertainty in a forecast is crucial. The natural alternative to a point forecast then is a probabilistic forecast that takes the form of a predictive distribution over future weather quantities and events (Gneiting 2008).

Probabilistic weather forecasts benefit many realms of society. Dutton (2002, p. 1306) estimated that upward of $3 trillion in annual private industry activities in the United States are subject to weather-related risk. In this article, we focus on quantitative precipitation, where probabilistic forecasting allows for optimal decision making in a wealth of applications (Krzysztofowicz 2001; Palmer 2002; Zhu et al. 2002). For example, extreme precipitation can force the transportation industry to cancel flights or reroute ships, and cause authorities to salt roads or clear snow. In mountainous regions, sudden heavy localized precipitation can lead to flash floods. Locally accurate forecasts are also an important tool in agricultural planning, such as to avoid unnecessary irrigation when anticipating precipitation events (Stern and Coe 1982; Katz and Murphy 1997).

In meteorology, the preferred way of producing a probabilistic forecast is to run an ensemble of numerical weather prediction models, in which multiple estimates of the current state of the atmosphere and/or multiple physics options are used to generate an estimate of the probability distribution of future weather events (Palmer 2002; Gneiting and Raftery 2005). However, operational ensemble prediction systems have biases, and typically they lack spread, and thus must be statistically postprocessed to generate calibrated predictive distributions (Hamill and Colucci 1997). Raftery et al. (2005) introduced Bayesian model averaging (BMA) as a way of doing this for temperature and pressure, with more recent extensions to quantitative precipitation (Sloughter et al. 2007), wind direction (Bao et al. 2010), and wind speed (Sloughter, Gneiting, and Raftery 2010).

These postprocessing methods are globally calibrated, that is, when verification data from many locations are aggregated, the predictive distributions and the observations are statistically compatible. There is no guarantee that a postprocessed predictive distribution will be calibrated at any single location, and Atger (2003) and Hamill and Juras (2006) warned that the aggregation of validation data across the forecast domain can lead to overestimates of the predictive performance.

Precipitation is highly affected by local terrain features such as elevation and leeward or windward siting, and while numerical models attempt to account for these dependencies, significant systematic bias remains. One way to guarantee local calibration is to fit the statistical model at an individual location, or at many different locations separately (Thorarinsdottir and Gneiting 2010). However, this approach does not explicitly generate predictive models between observation stations, and thus does not apply on forecast grids, as is commonly required in practice. In fact, there has been a recent call to generalize the BMA work of Raftery et al. (2005) to "account for geographical variations" (Iversen et al. 2011, p. 513). In this light, Kleiber

William Kleiber is Post-Graduate Scientist, Institute for Mathematics Applied to Geosciences, National Center for Atmospheric Research, Boulder, CO 80307-3000 (E-mail: *wkleiber@uw.edu*). Adrian E. Raftery is Professor of Statistics and Sociology, Department of Statistics, University of Washington. Tilmann Gneiting is Professor of Mathematical Statistics, Institut für Angewandte Mathematik, Universität Heidelberg, Germany. We are indebted to Jeff Baars, Chris Fraley, Cliff Mass, and McLean Sloughter for sharing their insights and providing data. This research was sponsored by the National Science Foundation under Joint Ensemble Forecasting System (JEFS) subaward S06-47225 with the University Corporation for Atmospheric Research (UCAR), as well as grants ATM-0724721 and DMS-0706745. Tilmann Gneiting furthermore acknowledges support by the Alfried Krupp von Bohlen und Halbach Foundation.

Journal of the American Statistical Association
December 2011, Vol. 106, No. 496, Applications and Case Studies
DOI: 10.1198/jasa.2011.ap10433

et al. (2011) introduced geostatistical model averaging (GMA) to produce locally calibrated probabilistic forecasts of surface temperature on forecast grids. Their methodology applies to any meteorological variable for which Gaussian predictive distributions are plausible, such as surface temperature or pressure. Precipitation, however, is incompatible with any Gaussian assumption, as it can take on nonnegative values only, and there is a positive probability of exactly zero precipitation occurring.

In this article we introduce a GMA approach that generates locally calibrated and sharp predictive distributions for quantitative precipitation from an ensemble of numerical weather forecasts. The method builds on those of Sloughter et al. (2007), who used BMA to globally calibrate precipitation forecasts, and combines BMA with the ideas of Kleiber et al. (2011). We apply the GMA method to the eight-member University of Washington Mesoscale Ensemble (UWME; Grimit and Mass 2002; Eckel and Mass 2005), where it captures the local behavior of daily precipitation accumulation over the North American Pacific Northwest. The region has many terrain features that make weather forecasting in general especially difficult (Mass 2008). In close proximity to the Pacific Ocean, weather systems encounter major mountain ranges with an elevation change from sea level to well over 4000 meters. The Pacific Northwest thus features extreme precipitation climates, with the rain forests of the Olympic Peninsula receiving up to 5000 mm of precipitation annually, while only a few dozen kilometers away the city of Sequim, Washington averages only about 400 mm. The Mount Baker area in the Cascade Mountains east of the town of Bellingham, Washington owns the world record for the highest seasonal total snowfall. Throughout the region, both rain and snow contribute to precipitation accumulations, depending on altitude and season, among other factors.

Operational probabilistic weather forecasting requires probabilistic forecasts on a grid for forecast horizons ranging from as little as three hours to up to several days ahead. The grid currently used in UWME is at a spacing of 12 km, and in the Pacific Northwest amounts to nearly 10,000 grid points. The computational load required to generate 10,000 forecast distributions is substantial, and as forecasts must be generated in real time, fully Bayesian approaches to estimation that integrate out parameter uncertainty are not feasible. We describe a two-step estimation procedure that is fast to implement by fixing estimated parameters, and whose forecasts are effective.

The article is structured as follows. Section 2 reviews the gamma-point mass mixture model of Sloughter et al. (2007) and the geostatistical approach of Kleiber et al. (2011) in the simplified case in which only a single numerical precipitation prediction is available. In Section 3 we complete the GMA methodology so that it applies to the general case of an ensemble prediction system. Section 4 reviews methods for assessing predictive performance. Section 5 presents aggregate results for probabilistic forecasts of daily precipitation accumulation over the Pacific Northwest. In addition, an individual location is studied in detail. In Section 6, we discuss extensions and possible directions for future research.

2. SINGLE FORECAST MODELS

In this section we combine the gamma-point mass mixture model of Sloughter et al. (2007) and the geostatistical approach of Kleiber et al. (2011) for the simplified case in which there is only a single numerical quantitative precipitation forecast available. In this context, we review the standard global model of Sloughter et al. (2007), and introduce our geostatistical approach for postprocessing the numerical prediction. All subsequent models specify predictive densities for the variable y_s, the cube root of daily precipitation accumulation at site s, conditionally on a numerical forecast, f_s, for the (untransformed) daily precipitation accumulation at this site. In modeling the cube root of precipitation accumulation we follow Sloughter et al. (2007), who used this transformation to linearize the relationship between numerical forecasts and observations. Other authors have considered more extreme transformations, such as the fourth root (Hamill, Whittaker, and Wei 2004). Specifically, we let

$$p(y_s|f_s) = \mathrm{P}(y_s = 0|f_s)\mathbb{1}_{[y_s=0]} + \mathrm{P}(y_s > 0|f_s)g(y_s|f_s)\mathbb{1}_{[y_s>0]}, \quad (1)$$

where the first term is a point mass at zero and $g(y_s|f_s)$ is a gamma density. The probability of precipitation, $\mathrm{P}(y_s > 0|f_s)$, and the mean, μ_s, and variance, σ_s^2, of the gamma density $g(y_s|f_s)$ depend on the numerical forecast f_s as described below.

2.1 Single Forecast Global Model

The global model of Sloughter et al. (2007) specifies the probability of precipitation via the logistic regression equation

$$\operatorname{logit} \mathrm{P}(y_s > 0|f_s) = a_0 + a_1\left(f_s^{1/3} - \overline{f^{1/3}}\right) + a_2\mathbb{1}_{[f_s=0]}, \quad (2)$$

where the parameters a_0, a_1, and a_2 do not vary by location, and $\overline{f^{1/3}}$ is a global average over the forecasts in the training period. Equation (2) is a slight variation on that introduced by Sloughter et al. (2007) which does not include $\overline{f^{1/3}}$; we include this offset to aid direct comparisons to the geostatistical model, discussed below. Conditionally on there being precipitation, we model the cube root of the precipitation accumulation, y_s, by a gamma distribution with mean

$$\mu_s = b_0 + b_1 f_s^{1/3} \quad (3)$$

and variance

$$\sigma_s^2 = c_0 + c_1 f_s, \quad (4)$$

where the parameters b_0, b_1, c_0, and c_1 do not depend on location. The conversion from the predictive distribution for the cube root to the predictive distribution for the original, nontransformed precipitation accumulation, which is the quantity typically required in practice, is straightforward.

For estimation, we follow Sloughter et al. (2007). The logistic regression parameters in Equation (2) are estimated by maximum likelihood, using domain wide training data from a sliding window training period, consisting of the most recent T available days. Sloughter et al. (2007) recommended a 30-day training period as a length that experimentally minimized both domain averaged mean absolute error and continuous ranked probability score. Typically over the training period both positive and vanishing daily precipitation accumulations are observed, especially since training data are used from multiple locations. The raw forecasts themselves include most of the temporally varying dynamics of the atmosphere, but often the forecast bias evolves over time. Using a sliding training window allows the statistical model to update to the temporally varying bias, where a longer window will stabilize parameter estimates, but a shorter period allows the statistical model to adjust better to shifts in weather regimes.

The gamma mean and variance parameters in Equations (3) and (4) are then estimated using only cases with positive precipitation accumulation. Specifically, b_0 and b_1 are estimated by ordinary least squares regression, while c_0 and c_1 are estimated by maximum likelihood, viewing the observation-forecast pairs as independent. As the maximum likelihood estimates (MLEs) are not available in closed form, we numerically optimize the log-likelihood function via the limited memory quasi-Newton bound constrained optimization method of Byrd et al. (1995), as implemented in R (Ihaka and Gentleman 1996). Implementing a full Bayesian hierarchical model is not an appealing option in real time, nor likely to result in improved predictive performance (Vrugt, Diks, and Clark 2008).

Turning now to forecasting, for any location s of interest, the predictive distribution of the cube root of daily precipitation accumulation, y_s, conditional on the numerical forecast, f_s, is given by Equation (1) with the global parameter estimates plugged in.

2.2 Single Forecast Geostatistical Model

We now introduce the single forecast geostatistical model that allows the parameters to vary by location. Under the geostatistical model, the logistic regression equation for the probability of precipitation becomes

$$\operatorname{logit} \mathrm{P}(y_s > 0|f_s) = a_{0s} + a_{1s}\left(f_s^{1/3} - \overline{f_s^{1/3}}\right) + a_{2s}\mathbb{1}_{[f_s=0]}, \qquad (5)$$

where $\overline{f_s^{1/3}}$ is now a local average over the forecasts in the training period, using only forecasts at site s. The motivation for including the $\overline{f_s^{1/3}}$ term is to reduce correlation between a_{0s} and a_{1s}, which facilitates our subsequent Bayesian regularization approach. The mean and variance of the gamma component in the predictive density (1) are now modeled as

$$\mu_s = b_{0s} + b_{1s}f_s^{1/3} \qquad (6)$$

and

$$\sigma_s^2 = c_{0s} + c_{1s}f_s, \qquad (7)$$

where all parameters are location dependent. Note that the single forecast global model is a special case of our geostatistical model, where $a_{is} = a_i$ for $i = 0, 1, 2$ and $b_{is} = b_i$ with $c_{is} = c_i$ for $i = 0, 1$, where, at location s, we re-center the logistic regression by $a_{1s}(\overline{f_s^{1/3}} - \overline{f^{1/3}})$.

We model all parameters $a_{0s}, a_{1s}, a_{2s}, b_{0s}, b_{1s}, c_{0s}$, and c_{1s} as independent spatial Gaussian processes. Each process has a constant mean μ_{pn}, where $p = a, b, c$ and $n = 0, 1$ and possibly 2, and a covariance function of the form

$$\begin{aligned} &\mathrm{C}_{pn}(s_1, s_2) \\ &\quad = \mathrm{Cov}(p_{ns_1}, p_{ns_2}) \\ &\quad = \tau_{pn}^2\mathbb{1}_{[s_1=s_2]} + \rho_{pn}^2 \exp\left(-\frac{\|s_1 - s_2\|}{r_{pn1}} - \frac{|h(s_1) - h(s_2)|}{r_{pn2}}\right), \end{aligned} \qquad (8)$$

where $h(s)$ is the elevation at location s. Here, τ_{pn}^2 is the nugget effect in geostatistical terminology, ρ_{pn}^2 is a variance parameter, r_{pn1} is a range parameter that corresponds to horizontal distance, and r_{pn2} is a range parameter that corresponds to differences in elevation. This form of second-order structure is designed to reflect the fact that precipitation strongly depends on elevation as well (Basist, Bell, and Meentemeyer 1994; Daly, Neilson, and Phillips 1994).

Estimation proceeds in two steps, as the probabilistic forecasts must be produced in real time operationally, and implementing a full Bayesian hierarchical model is not tenable, nor likely to result in improved predictive performance (Vrugt, Diks, and Clark 2008). First, we gather point estimates of all forecasting parameters $a_{0s}, a_{1s}, a_{2s}, b_{0s}, b_{1s}, c_{0s}$, and c_{1s} at training stations. The hyperparameters governing the spatial structure of each process are then estimated by maximum likelihood, conditional on the point estimates. In our experience, the spatial structures do not evolve substantially across time, and hence the time-expensive likelihood estimation may be done once on training data, and held fixed during the forecasting period. Point estimates of the forecasting parameters are updated at each time step, outlined next.

In estimating the parameters b_{0s}, b_{1s}, c_{0s}, and c_{1s} of the model for the amount of precipitation, we follow the scheme described in Section 2.1, except that the training data are not aggregated across the domain, and each parameter is estimated at every training station separately, using the most recent available T days of positive precipitation accumulation as training data. As in the global model, our sliding training window of length T allows the statistical parameters to update with temporally varying local biases and adjust for sudden shifts in weather regimes.

To estimate the parameters a_{0s}, a_{1s}, and a_{2s} that determine the probability of precipitation, we proceed as follows. For the global model, data are aggregated across all locations, and in our experience there are sufficiently many observations of zero as well as nonzero precipitation to avoid degeneracies in estimating the global logistic regression model. Moving to the geostatistical model introduces this difficulty, especially during the dry season when many weeks may pass without any rain at a given station. Maximum likelihood estimation on a 30-day training period, say, at a location with exclusively zero observations can result in poorly behaved, degenerate estimates. In particular, after a string of days with no rain at a given location, the maximum likelihood estimates are $\pm\infty$, as the maximizing value of probability is either 0 or 1, depending on the context. This problem is well known and has been addressed by several authors, who generally suggest Bayesian methods of regularizing parameter estimates (Clogg et al. 1991; Fraley and Raftery 2007; Berrocal et al. 2010). Ideally we would like a regularization procedure to agree with traditional logistic regression estimates in well-behaved cases, and to shrink toward reasonable values in poorly behaved cases. An ad hoc method is to increase the length of the training period, so as to capture training days with zero as well as nonzero precipitation accumulation. Our experiences with this approach have been mixed (Kleiber 2010).

Here we take another approach that preserves the training period, but replaces maximum likelihood by a Bayesian technique that imposes independent normal priors on a_{0s}, a_{1s}, and a_{2s}, using the posterior mean as the plug-in estimate. The use of independent priors can be relaxed, but in our experience works well for predictive purposes. Hyperparameter choices depend on the situation at hand and will be discussed later on in Section 5.

Genkin, Lewis, and Madigan (2007) noted that maximum likelihood for logistic regression models with a large number of predictors can fail, and recommended using a Laplace prior to induce a sparse structure on the regression coefficients. Alternatively, Gelman et al. (2008) developed a Cauchy prior which uses minimal knowledge, and can be applied universally. We use informative normal priors due to the availability of a historical data record, and have no need to induce sparsity on our three coefficients. We avoid the Cauchy prior here as we have a wealth of historical data which can inform the prior parameters and shrink posterior means toward stable global estimates.

As we do not require full posterior distributions for the logistic regression parameters, we avoid sampling algorithms, and instead use Laplace approximations to find the posterior means (Tierney and Kadane 1986; Wong 2001), as described in the Appendix. As an example, Figure 1 compares maximum likelihood and Laplace-approximated posterior mean estimates of a_{0s}, a_{1s}, and a_{2s} at Vancouver International Airport, British Columbia across 2008, based on training data from a sliding window 30-day training period. The Bayesian approach regularizes the MLEs, while showing good agreement when the latter is well behaved. Our use of independent normal priors is important in the Laplace approximation, which depends on the Hessian of the log-likelihood function. The Hessian, and subsequently our posterior mean approximations, would change if a multivariate prior were used instead.

In the second stage of estimation, the spatial parameters that define the Gaussian process structure of a_{0s}, a_{1s}, a_{2s}, b_{0s}, b_{1s}, c_{0s}, and c_{1s} are estimated by maximum likelihood conditional on the above point estimates. To give an example, suppose that at sites $s = s_1, \ldots, s_N$, we have posterior mean estimates $\hat{a}_{0s}$ based on a historical data record. We view $\{\hat{a}_{0s} : s = s_1, \ldots, s_N\}$ as a partial realization of a Gaussian random field, whose log-likelihood is then maximized with respect to the covariance parameters μ_{a0}, τ^2_{a0}, ρ^2_{a0}, r_{a01}, and r_{a02}. We discuss the details of our implementation in Section 5.

Turning now to forecasting with the geostatistical model, we first consider the training stations $s = s_1, \ldots, s_N$, where we apply Equation (1) with the station-specific estimates for the probability of precipitation and gamma density parameters plugged in.

If interest lies in prediction at a site s, where no training data are available, we use the geostatistical method of inter-

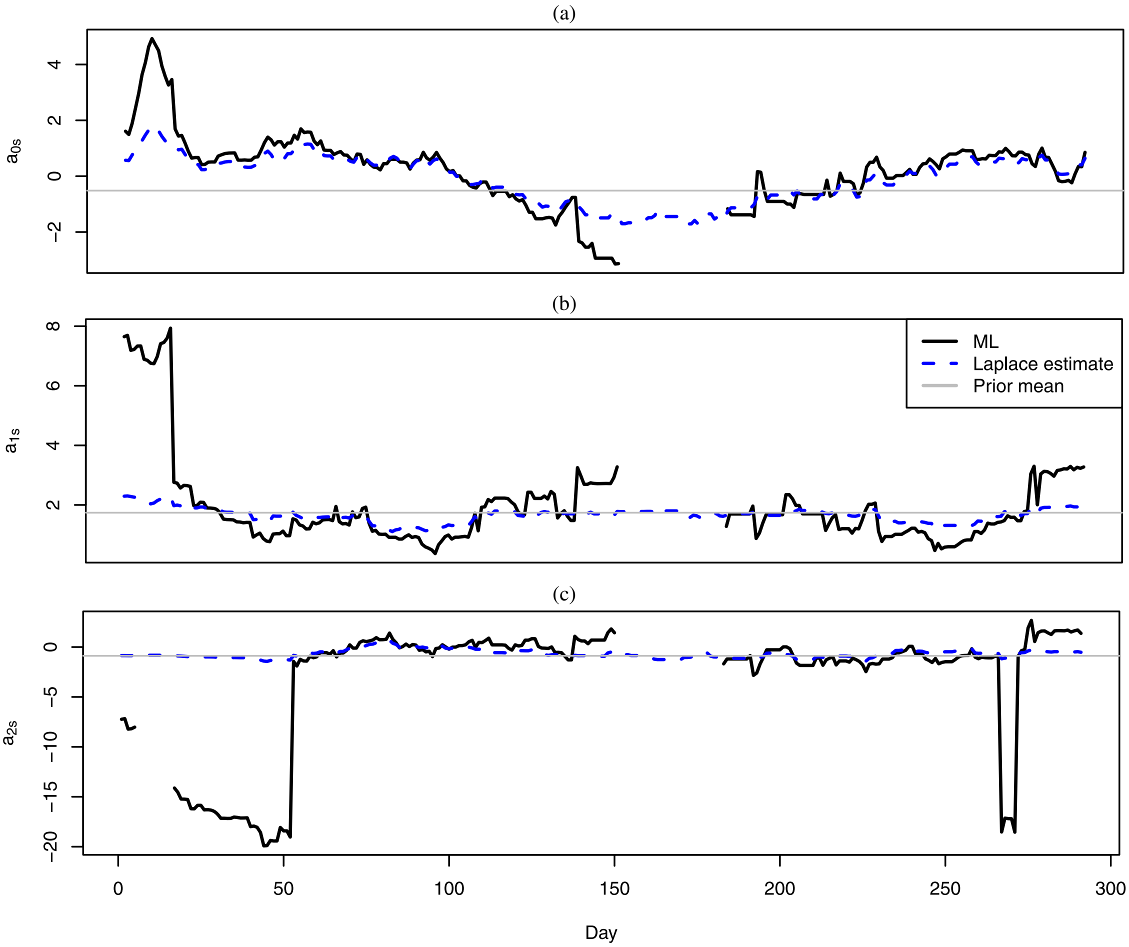

Figure 1. Maximum likelihood estimates (MLEs) and Laplace-approximated posterior mean estimates for the probability of precipitation parameters (a) a_{0s}, (b) a_{1s}, and (c) a_{2s} at the training station at Vancouver International Airport, British Columbia across 2008, using a 30-day sliding window training period. When the MLE is missing, it is degenerate and estimated as ±∞. The online version of this figure is in color.

polation known as kriging (Cressie 1993; Stein 1999) based on the covariance structure (8) and plug-in MLEs for its parameters, $\mu_{pn}, \tau_{pn}^2, \rho_{pn}^2, r_{pn1}$, and r_{pn2} where p refers to a, b, c and n to $0, 1$ and possibly 2. Kriging results in interpolated estimates $\hat{a}_{0s}, \hat{a}_{1s}, \hat{a}_{2s}, \hat{b}_{0s}, \hat{b}_{1s}, \hat{c}_{0s}$, and $\hat{c}_{1s}$ at the desired location s, which we plug into the predictive density in Equation (1). Here, we condition on the kriging estimates and other parameters; in our experience and that of other authors (Vrugt, Diks, and Clark 2008) this works well, and a fully Bayesian treatment does not result in improved predictive performance. Specifically, we also explored the fully Bayesian posterior predictive distribution with our Gaussian process priors, but the results were comparable to our conditional point estimate approach, and the computational load is significantly higher.

Kriging is an exact interpolator; that is, if s is a training location, then the interpolated value agrees with the site-specific value. The kriging predictor is a linear function of the observed values, and puts most weight on nearby observations at similar altitudes. Hence, kriging provides a unified way to do parameter estimation at and between training stations.

We close this section by noting that we have not imposed any sign constraints on b_{0s}, b_{1s}, c_{0s}, and c_{1s}. It is possible that these take on very small negative values, and in fact in our case study below, they sometimes do. However, this happens rarely (in less than half a percent of cases in our example), and then the member forecasts are such that the gamma means and variances remain positive. A possible alternative would be to model $\log b_{is}$ and $\log c_{is}$, for example, as Gaussian processes, thereby precluding the possibility of invalid parameter estimates in the gamma components.

3. ENSEMBLE MODELS

Thus far in this article, we have considered the generation of predictive distributions for future precipitation accumulation based on a single numerical weather forecast. However, it is common practice to run ensembles of numerical weather prediction models, in which multiple estimates of the current state of the atmosphere and/or multiple physics options are used to generate an estimate of the probability distribution of future weather events (Palmer 2002; Gneiting and Raftery 2005).

Thus, we extend our model to take advantage of the information in multiple ensemble members, and turn to the situation where we have an ensemble of $K > 1$ numerical forecasts of precipitation accumulation. We follow Raftery et al. (2005) and Sloughter et al. (2007) in postprocessing the ensemble forecasts using Bayesian model averaging (BMA). The geostatistical model now becomes geostatistical model averaging (GMA), and the global model becomes Global BMA.

Let y_{st} denote the cube root of the precipitation accumulation at site s and time t, and consider the ensemble member forecasts $f_{1st}, \ldots, f_{Kst}$ for the (non-transformed) precipitation accumulation. Following Sloughter et al. (2007), each ensemble member is associated with a density $p(y_{st}|f_{kst})$ of the form in Equation (1) with probability of precipitation parameters $a_{k0s}, a_{k1s}, a_{k2s}$ and gamma mean parameters b_{k0s} and b_{k1s}, where $k = 1, \ldots, K$. These parameters are estimated in the same way as in the previous section, except that they are now member-specific. The associated covariance parameters in Equation (8) also become member-specific. Furthermore, both Global BMA and GMA use gamma variance parameters c_{0s} and c_{1s} that do not depend on the ensemble member, and whose estimation we discuss below. Of course, for Global BMA, the parameters do not vary with location.

The BMA predictive density then is

$$p(y_{st}|f_{1st}, \ldots, f_{Kst}) = \sum_{k=1}^{K} w_k p(y_{st}|f_{kst}), \tag{9}$$

where the member weights $w_1, \ldots, w_K$ are probabilities, and thus are nonnegative and sum to 1. This approach to combining forecast densities accommodates both Global BMA and GMA, and remains valid when some or all of the ensemble members are exchangeable, with straightforward adaptations (Fraley, Raftery, and Gneiting 2010). We now discuss the estimation of the gamma variance parameters c_{0s} and c_{1s}, which do not depend on the ensemble member, and the BMA weights $w_1, \ldots, w_K$. Note that the weights for both models are global parameters, and do not vary by location, in line with the experience of Brentnall, Crowder, and Hand (2011, p. 1158), who concluded that "performance is worse when the weights in the predictions are allowed to vary across the stations."

In GMA, c_{0s} and c_{1s} are estimated via maximum likelihood by numerically maximizing the joint log-likelihood which is a sum of the logarithms of gamma densities over all K ensemble members and time points in the training period with estimates $\hat{b}_{k0s}$ and $\hat{b}_{k1s}$ plugged in. Conditionally on these estimates, we fit the member weights $w_1, \ldots, w_K$ by maximum likelihood using a version of the expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin 1977). This version introduces latent variables $z_{1st}, \ldots, z_{Kst}$ for each site and time, where z_{kst} can be interpreted informally as the probability of member k being the most skillful for this site and time. The notion of a most skillful member is frequently invoked by operational weather forecasters, who tend to select a "best member" based on recent performance (Joslyn and Jones 2008). The EM algorithm is iterative, and the jth E step takes current estimates $\{w_k^{(j)}\}_{k=1}^{K}$, and, based on these, computes

$$\hat{z}_{kst}^{(j)} = \frac{w_k^{(j)} p^{(j)}(y_{st}|f_{kst})}{\sum_{\ell=1}^{K} w_\ell^{(j)} p^{(j)}(y_{st}|f_{\ell st})} \tag{10}$$

for every ensemble member, where $p^{(j)}(y_s|f_{\ell st})$ is the density of Equation (1) with the GMA estimates plugged in. In the M step, the weights are updated via

$$w_k^{(j+1)} = \frac{1}{N} \sum_{s,t} \hat{z}_{kst}^{(j)}, \tag{11}$$

where N is the number of station and time pairs. The E and M steps are then iterated until convergence is reached.

In Global BMA, the gamma variance parameters and the weights are estimated simultaneously. Specifically, we update $c_0^{(j)}$ and $c_1^{(j)}$ by using a variation of EM called ECME (Expectation/Conditional Maximization Either; Liu and Rubin 1994). The algorithm consists of E (expectation) and CM (conditional maximization) steps. The jth E step is analogous to Equation (10), but with current estimates $c_0^{(j)}$ and $c_1^{(j)}$ used in $p^{(j)}(y_s|f_{\ell st})$. The CM step is split into two steps, with the initial (CM-1) step being identical to GMA's M step, Equation (11).

Conditionally on these updated $w_k^{(j+1)}$ estimates, the CM-2 step maximizes the mixture likelihood (9) numerically as a function of $c_0^{(j+1)}$ and $c_1^{(j+1)}$.

4. ASSESSING PREDICTIVE PERFORMANCE

The basic goal in probabilistic forecasting is to maximize the sharpness of the predictive distributions subject to calibration (Gneiting, Balabdaoui, and Raftery 2007).

Probability integral transform (PIT) histograms are useful tools for visually examining calibration. The method relies on the fact that if the random variable Y has a continuous cumulative distribution function F, then $U = F(Y)$ is uniformly distributed. In practice, one evaluates the predictive cumulative distribution function at the realizing observation to compute the PIT, aggregates over forecast cases, and plots a histogram. In our case, the predictive distributions have a point mass at the origin, and if there is no precipitation, we randomize, that is, the PIT is a random value uniformly between zero and the probability of no precipitation (Sloughter et al. 2007; Czado, Gneiting, and Held 2009). Uniformity of the PIT histogram corresponds to the probabilistic calibration criterion of Gneiting, Balabdaoui, and Raftery (2007). We quantify the deviation of the PIT histogram from uniformity by averaging the absolute difference of histogram values from unity; we call this the PIT discrepancy criterion. The choice of the number of bins for the PIT histogram is not straightforward, but there is a related diagram for ensemble forecasts called the verification rank histogram (Hamill and Colucci 1997). If an ensemble of size K is calibrated, then the rank of the observation within the combined set of the ensemble members and the observation has a discrete uniform distribution on the integers from 1 to $K + 1$. As we will be comparing to an ensemble forecast, we use $K + 1$ bins in our PIT histograms as well.

Sharpness can be examined by looking at the width of prediction intervals. For unimodal densities, central intervals are often appropriate, but for highly skewed positive variables such as precipitation and wind speed, lower intervals are preferable. As well as looking at average interval widths, we consider a proper scoring rule called the quantile score (QS). The special case for the $(1 - \alpha) \times 100\%$ quantile arises from equation (40) of the article by Gneiting and Raftery (2007) with $s(x) = \frac{x}{\alpha}$ and $h(x) = -\frac{x}{\alpha}$, and is defined by

$$\mathrm{QS}_{1-\alpha}(u, y) = u + \frac{1}{\alpha}(y - u)\mathbb{1}_{[y>u]},$$

where u is the $(1 - \alpha) \times 100\%$ quantile of the predictive density and y is the realizing observation. Note that we take the negative of the score here to make it negatively oriented, that is, the smaller the better. This score is especially attractive for lower intervals with a cutoff at zero, such as are encountered in precipitation or wind forecasting. The mean score then equals the average prediction interval width when the predictive intervals always capture the realizing observation. If the observation falls outside of the interval, the score adds a penalty term proportional to the distance between the observation and the upper boundary of the predictive interval.

To assess calibration and sharpness jointly, we consider the continuous ranked probability score (CRPS), which is defined by

$$\mathrm{CRPS}(F, y) = \int_{-\infty}^{+\infty} (F(x) - \mathbb{1}(x \geq y))^2 \, dx,$$

where F is the predictive cumulative distribution function (CDF) and y is the observed precipitation accumulation. The CRPS may equivalently be expressed as

$$\mathrm{CRPS}(F, y) = \mathbb{E}_F|Y - y| - \tfrac{1}{2}\mathbb{E}_F|Y - Y'|,$$

where Y and Y' are independent random variables with CDF F. The initial definition illustrates that the CRPS accounts for both sharpness and calibration simultaneously, while the second formulation shows that it has the same unit as the outcome, y. A natural summary to examine is the mean absolute error (MAE), that is, the average difference between the median of the predictive density and the realizing observation.

The above scores, apart from the MAE, involve the entire predictive distribution and so we include the popular Brier score (Brier 1950; Gneiting and Raftery 2007) to assess the probability of precipitation forecasts. For a series of binary observations o_i corresponding to rain/no rain and probability forecasts p_i for $i = 1, \ldots, n$, the Brier score is defined by

$$\text{Brier score} = \frac{1}{n}\sum_{i=1}^{n}(o_i - p_i)^2.$$

The Brier score is also negatively oriented with smaller values indicating better performance. Finally, we visually assess calibration via the reliability diagram, which plots the conditional frequency of precipitation occurrence against the forecast probability. A well-calibrated probability forecast shows a reliability diagram tightly along the identity line.

5. PRECIPITATION FORECASTS FOR THE PACIFIC NORTHWEST

We now apply GMA to forecasts of 24-hour aggregated precipitation accumulation over the North American Pacific Northwest during the calendar years 2007–2008. The Pacific Northwest has terrain features that challenge weather forecasters, including the coastline, major mountain ranges, and the Puget Sound region, all of which have important effects on local precipitation patterns (Mass 2008). The predictions we use are 48-hour-ahead forecasts with the eight-member University of Washington Mesoscale Ensemble, described by Eckel and Mass (2005), initialized at 0000 UTC, which is 4:00 pm local time, except when Daylight Saving Time is in effect, when it is 5:00 pm local time.

Only observation stations on land are considered. Generally, precipitation is highly affected by terrain characteristics such as elevation, while over large bodies of water such as the Pacific Ocean the distribution of precipitation is much more homogeneous, and there is little need for a locally adaptive predictive model. There are 279 stations, which we randomly divide into 179 stations used for model fitting and 100 hold-out stations for validation. The stations are displayed in Figure 2.

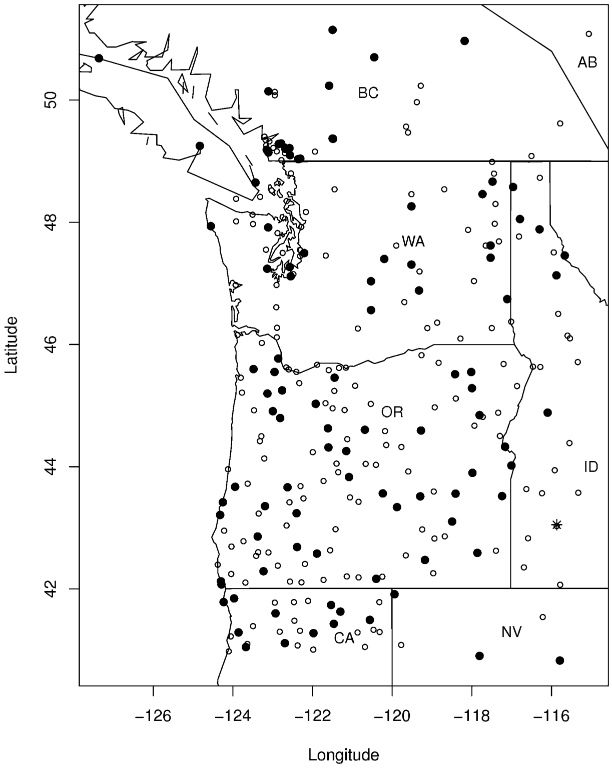

Figure 2. Pacific Northwest domain with fitting stations in small unfilled circles and validation stations in black dots. The location of Mountain Home Air Force Base, Idaho is shown by an asterisk.

5.1 Parameter Estimation

The first step in setting up the predictive distributions is to gather point estimates of the predictive model parameters. We use a sliding training window of 30 days, so that, on any particular day of interest, the previous 30 days are used to yield estimates $\hat{a}_{k0s}, \hat{a}_{k1s}, \hat{a}_{k2s}, \hat{b}_{k0s}, \hat{b}_{k1s}$, which depend on the ensemble member k, and estimates $\hat{c}_{0s}$ and $\hat{c}_{1s}$. The gamma component parameters b_{kis} and c_{is} are estimated as described in Section 2.2. Estimates of $a_{k0s}, a_{k1s}, a_{k2s}$ for GMA are set to the posterior mean as approximated by the Laplace method; this temporarily requires prior distributions. The prior distributions are independent normals, with hyperparameters estimated in the following way. Using only data from 2007, we find daily MLEs of a_{k0}, a_{k1}, and a_{k2} for the Global BMA model using a sliding window 30-day training period. The prior mean for the process $a_{k\ell s}$ is set to the median of the Global estimates for $a_{k\ell}$, while the prior standard deviation is set to twice the (asymptotically corrected) median absolute deviation of the Global estimates for $a_{k\ell}$. This variance inflation factor was also used by Berrocal et al. (2010). Note that we use robust estimates, median and median absolute deviation, to guard against the possibility that the MLEs of $a_{k\ell}$ are not well behaved. The weights are then estimated for each day of the validation year 2008 using the EM algorithm with all data at available training locations. This completes the first stage of estimation.

To interpolate the predictive parameters to locations other than training stations, GMA requires estimates of the spatial parameters $\mu_{kp}, \tau^2_{kp}, \rho^2_{kp}, r_{kp1}$, and r_{kp2} of the Gaussian processes. These are estimated once using data from the fitting stations in 2007, conditional on our point estimates of predictive parameters obtained in the first stage. Specifically, we use five realizations of $\hat{a}_{k0s}, \hat{a}_{k1s}, \hat{a}_{k2s}, \hat{b}_{k0s}, \hat{b}_{k1s}, \hat{c}_{0s}$, and $\hat{c}_{1s}$ at time points that are 60 days apart. This subset of days is used to provide approximately independent realizations of the spatial processes, giving a buffer period of two months between realizations. Conditioning on these realizations, we estimate the spatial covariance parameters by maximum likelihood. The spatial parameters are then held constant throughout the year 2008. Validation is always performed at the 100 hold-out stations in the year 2008, where the sliding training window for the statistical parameters requires the last 30 days of 2007.

5.2 Aggregate Results

We now present validation results with special attention to local predictive performance. Table 1 contains the aggregate scores for Global BMA and GMA averaged across the calendar year 2008 as well as the Pacific Northwest domain. We pause only to mention that the very use of an ensemble (compared to a single forecast) indeed led to improved scores, especially for the lower 90% prediction interval. GMA reduced the MAE of Global BMA by 7.2%, the mean CRPS by 5.5%, and the mean Brier score by 8.1%.

Table 1. Aggregate scores for Global BMA and GMA: mean absolute error (MAE), mean continuous ranked probability score (CRPS), mean Brier score, and mean quantile score, coverage percent, and average width for 50% and 90% lower prediction intervals. Units are millimeters for all scores except the Brier score, which is unitless, and coverage, which is in percent. Scores are listed for using only the GFS member forecast as well as the full eight-member ensemble

		Global BMA		GMA	
		Single forecast	Ensemble	Single forecast	Ensemble
	MAE	1.53	1.53	1.46	1.42
	CRPS	1.11	1.10	1.08	1.04
	Brier score	0.135	0.136	0.133	0.125
50%	Quantile score	3.36	3.36	3.29	3.26
	Width	0.77	0.62	0.90	0.82
	Coverage	48.9	48.3	48.6	48.0
90%	Quantile score	7.47	7.16	7.48	7.05
	Width	4.63	4.54	4.50	4.52
	Coverage	89.5	91.1	88.9	91.1

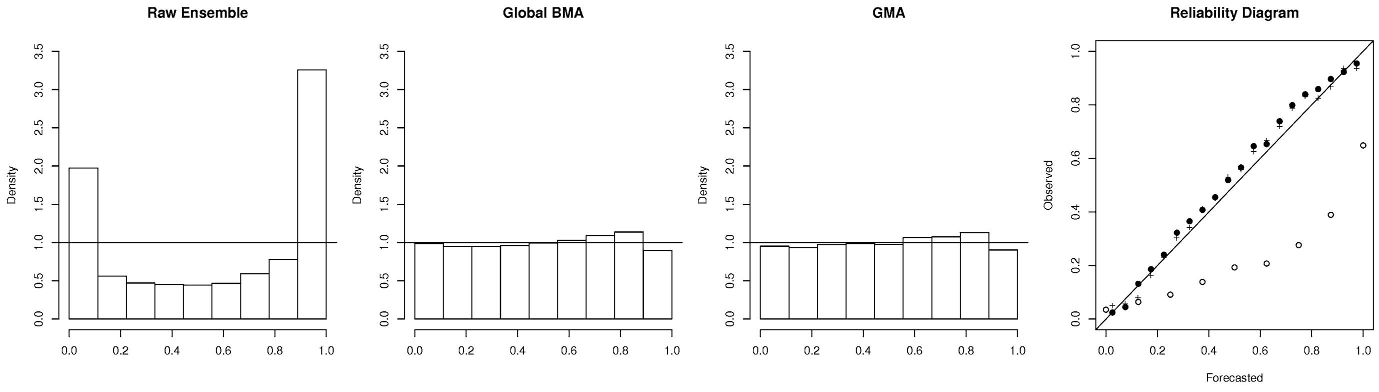

Figure 3. Aggregate results: Verification rank histogram for the raw ensemble, PIT histograms for Global BMA and GMA, and reliability diagram with the raw ensemble represented as empty circles, GMA as filled circles, and Global BMA as '+' symbols.

On the aggregate level, GMA showed slightly better quantile scores for both the nominal 50% and 90% prediction intervals, while both methods displayed good calibration at these levels. The improved scores indicate increased sharpness subject to calibration using the geostatistical model, as in the case of GMA for temperature (Kleiber et al. 2011). Notice that the average predictive interval widths are slightly worse for GMA than Global BMA at the 50% level. This is unsurprising, as GMA produces locally calibrated prediction intervals, and thus must sometimes increase the predictive interval width to achieve calibration. As we will see, GMA shows better local calibration, which follows our goal of maximizing sharpness subject to local calibration. Aggregate calibration can be visually assessed using the PIT histogram. Figure 3 shows the PIT histogram for both Global BMA and GMA, as well as reliability diagrams. Both methods showed good reliability on the aggregate level with a slight tendency to underforecast at central probabilities.

Table 2 provides a comparison of Global BMA and GMA in terms of the predictive performance at individual sites. For example, at 72% of our validation stations, GMA showed an improvement in the PIT discrepancy criterion over Global BMA, thereby demonstrating better local probabilistic calibration. Similar rates of improvement apply to other performance criteria, comprising calibration, sharpness, and proper scoring rules. To determine if these local effects could be explained by chance alone, we performed a paired *t*-test for each score, and we indeed find that GMA significantly improves all scores over Global BMA at the 1% level with the exception of predictive interval widths and the quantile score at the 90% level. As mentioned earlier, GMA adjusts the local interval widths in order to calibrate the predictive distributions locally. GMA significantly improves local calibration over Global BMA, which sometimes involves extending predictive interval widths, but this aligns with our goal to maximize sharpness subject to local calibration.

Table 2. Comparison of GMA and Global BMA in terms of performance measures at individual stations. Each entry shows the percent of validation stations that had a better result under GMA than under Global BMA. Improvement is defined by a lower mean absolute error (MAE), mean continuous ranked probability score (CRPS), mean PIT discrepancy, mean Brier score, or mean quantile score, or closer to nominal coverage for the 50% and 90% lower prediction intervals

	Performance criterion	Percent improved
	MAE	60
	CRPS	65
	Brier score	77
	PIT discrepancy	72
50%	Quantile score	60
	Width	54
	Coverage	76
90%	Quantile score	68
	Width	64
	Coverage	63

Finally, Figure 4 shows probability of precipitation forecast fields valid January 8, 2008 along with the verifying precipitation pattern for this day. GMA correctly identified the band of precipitation along the Pacific Coast and in the Puget Sound region, producing significantly higher probabilities than Global BMA. GMA was also able to identify the regions in northern Washington and eastern Oregon where there was little chance of precipitation on January 8. Since Global BMA's statistical parameters are not location dependent, the probability fields tend to be less sensitive to local terrain features, while the probability fields produced by GMA are able to adapt to the complex terrain of the Pacific Northwest.

5.3 Forecasts at Mountain Home Air Force Base, Idaho

We consider forecasts at Mountain Home Air Force Base, Idaho, whose location is shown in Figure 2. This is a hold-out station, and all parameter estimates for GMA are interpolated to this site, while Global BMA uses the same parameter values across the hold-out stations. The GMA method was locally calibrated at Mountain Home, as shown by the PIT histogram in Figure 5. The Global BMA method displayed a skewed PIT histogram, indicating slight overdispersion and a tendency to overpredict. Table 3 shows performance measures at Mountain Home, with the MAE, mean CRPS, and mean Brier score being much better for GMA. At the nominal 50% level, Global BMA was overdispersed by approximately 13.2%, while GMA was much closer to the nominal level, thus yielding calibrated and sharper intervals. At the 90% level, Global BMA also showed

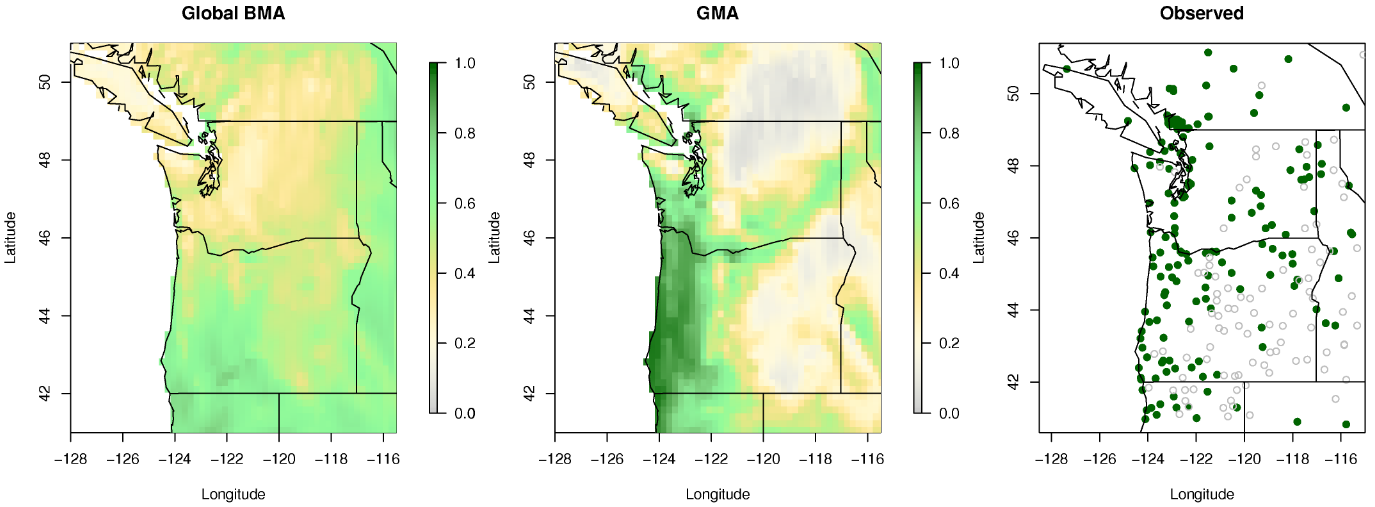

Figure 4. Forty-eight-hour-ahead probability of precipitation forecast fields valid January 8, 2008 for Global BMA and GMA. Observations are filled points if it rained on January 8, and empty points if there was no precipitation.

overdispersion, with a coverage of 97.2%, while GMA was at 91.9%.

Global BMA and GMA separate the estimation of the occurrence-of- and amount-of-precipitation models, and combine them in the final step for prediction. The gamma component for the amount of precipitation [Equations (3), (4), (6), and (7)] is weighted by the probability of precipitation for any given day, and hence is highly affected by the latter. Table 3 points to an important characteristic of GMA, namely, its ability to accurately model the probability of precipitation locally. Indeed, at Mountain Home GMA reduced the mean Brier score over Global BMA by 14.1%. The reason for the improvement is illustrated in Figure 6, which displays local estimates of a_{0s}, a_{1s}, and a_{2s}, as well as the (spatially constant) Global BMA estimates and the kriged values from GMA for the GFS member forecast. Global BMA consistently overestimated the local value of a_{0s} at Mountain Home throughout most of 2008, thereby over-weighting the amount-of-precipitation component, while GMA accurately followed the local estimates, which put a more appropriate weight on the gamma component. This effect is readily seen in Figure 7(a), which displays the Global BMA and GMA predictive densities for January 19, 2008 along with lower 90% intervals and the verifying observation. Global BMA put the probability of precipitation at 66.7% for this day, and hence assigned substantial mass to the gamma component, resulting in inflated higher quantiles. GMA, on the other hand, produced a low 31.2% probability of precipitation, even though all eight ensemble member forecasts were positive.

A second pair of predictive distributions that is valid for November 11, 2008 is displayed in Figure 7(b). Here, Global BMA and GMA agreed closely at probability of precipitation values of 97.6% and 96.0%, respectively. The effect of the local adjustment to the gamma parameters b_{0s}, b_{1s}, c_{0s}, and c_{1s} is readily seen, as GMA narrowed the lower 90% prediction interval by over 30 mm, when compared to Global BMA, while still capturing the verifying value. These two days illustrate the sharpness of the GMA forecasts, in addition to their being locally calibrated.

6. DISCUSSION

We have introduced a geostatistical model averaging (GMA) approach to generating locally calibrated probabilistic forecasts of precipitation accumulation from an ensemble of numerical weather predictions. The method builds on the Global BMA

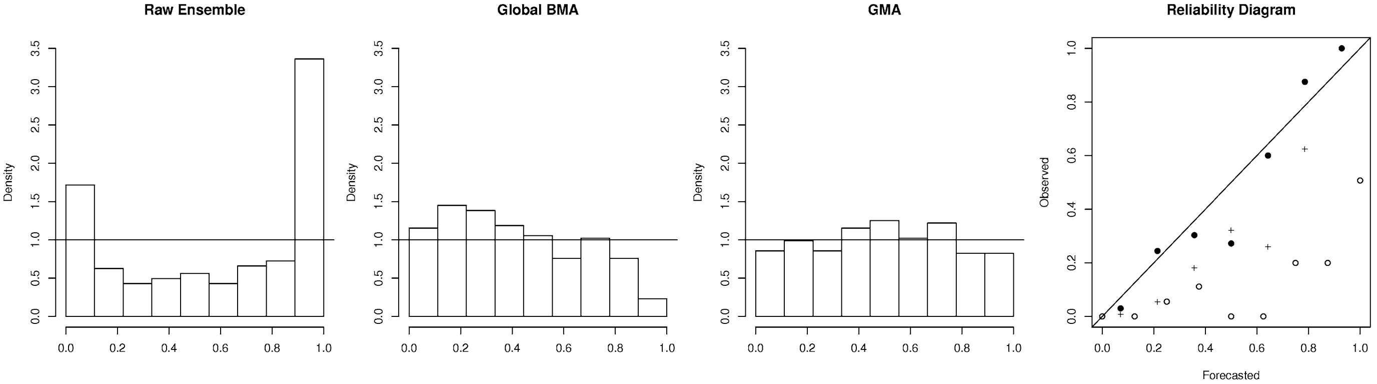

Figure 5. Results at Mountain Home Air Force Base, Idaho: verification rank histogram for the raw ensemble, PIT histograms for Global BMA and GMA, and the reliability diagram with the raw ensemble as empty circles, GMA as filled circles, and Global BMA as '+' symbols.

Table 3. Results at Mountain Home Air Force Base, Idaho: mean absolute error (MAE), mean continuous ranked probability score (CRPS), mean Brier score, and mean quantile score, coverage percent, and average width for the 50% and 90% lower prediction intervals. Units are millimeters for all scores except the Brier score, which is unitless, and coverage, which is in percent

		Global BMA	GMA
	MAE	0.45	0.27
	CRPS	0.38	0.24
	Brier score	0.11	0.09
50%	Quantile score	0.72	0.56
	Width	0.39	0.07
	Coverage	63.2	52.9
90%	Quantile score	3.39	2.11
	Width	2.99	0.97
	Coverage	97.2	91.9

model of Sloughter et al. (2007) but allows all parameters to vary by location, except for the BMA weights, for which the predictive performance is known to benefit from spatial constancy (Brentnall, Crowder, and Hand 2011). When applied to the eight-member University of Washington Mesoscale Ensemble, GMA had better aggregate and local calibration than Global BMA, and was significantly sharper on average.

The GMA approach relies on modeling each parameter as a Gaussian process with constant mean and a covariance function that decays exponentially across horizontal distance and elevation. This basic idea accommodates many adjustments, including more complex first- and second-order structures. Kleiber (2010) examined a number of extensions to this standard GMA formulation including nonconstant means and multivariate covariance functions that allow for dependencies between parameters, but found that the added complexity did not substantially improve the predictive performance. Nonetheless, these extensions may prove useful in other settings with different types of ensemble systems or other geographic domains. Berrocal, Gelfand, and Holland (2010a, 2010b) reported on related developments in the context of air quality; all these approaches are examples of general varying coefficient models (Hastie and Tibshirani 1993).

Our estimation scheme for the model parameters is split into two steps to increase efficiency and allow implementation in

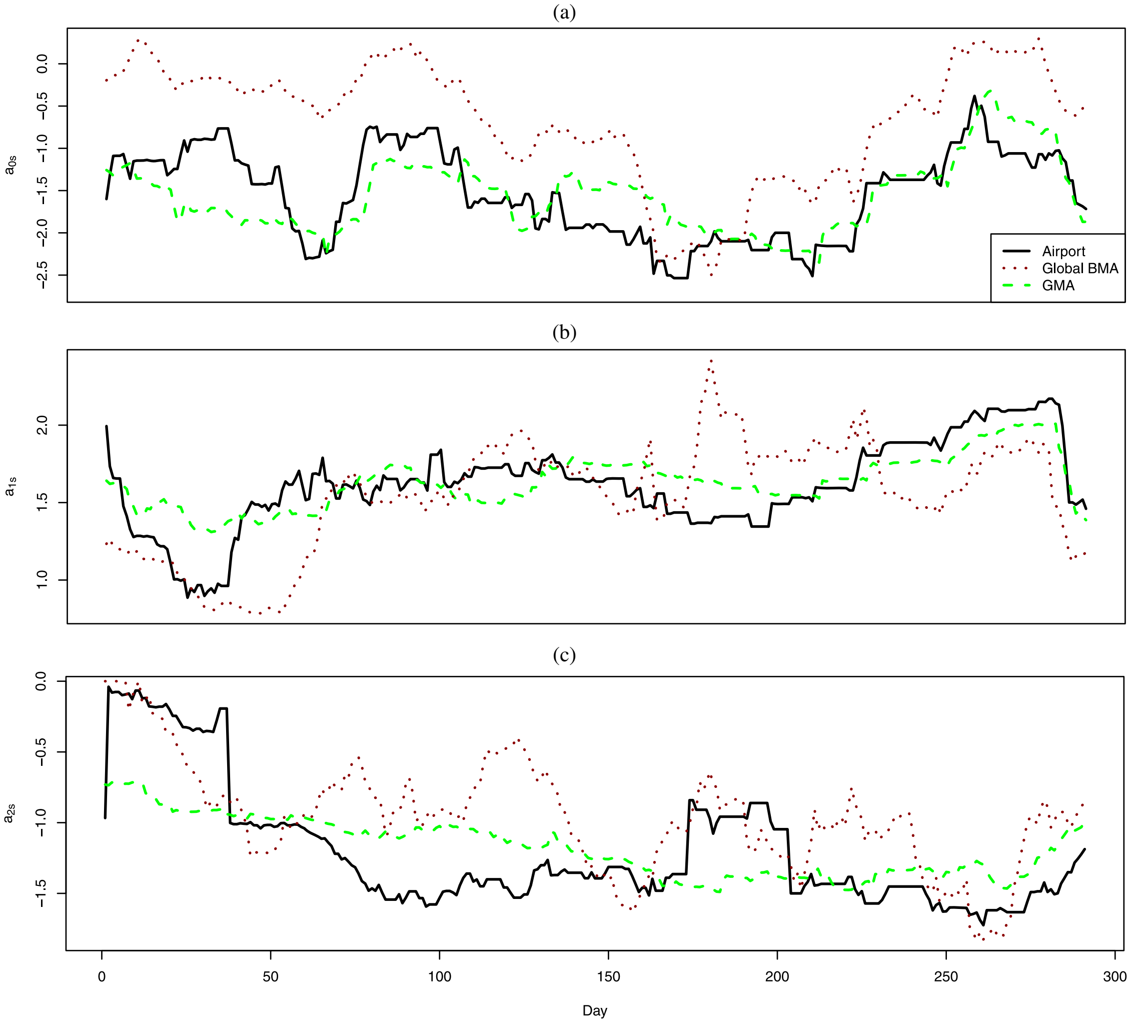

Figure 6. Local estimates for the probability of precipitation parameters (a) a_{0s}, (b) a_{1s}, and (c) a_{2s} at Mountain Home Air Force Base, Idaho (black), kriged values from GMA, and the globally constant parameters of Global BMA for the GFS member forecast over the validation year of 2008. The online version of this figure is in color.

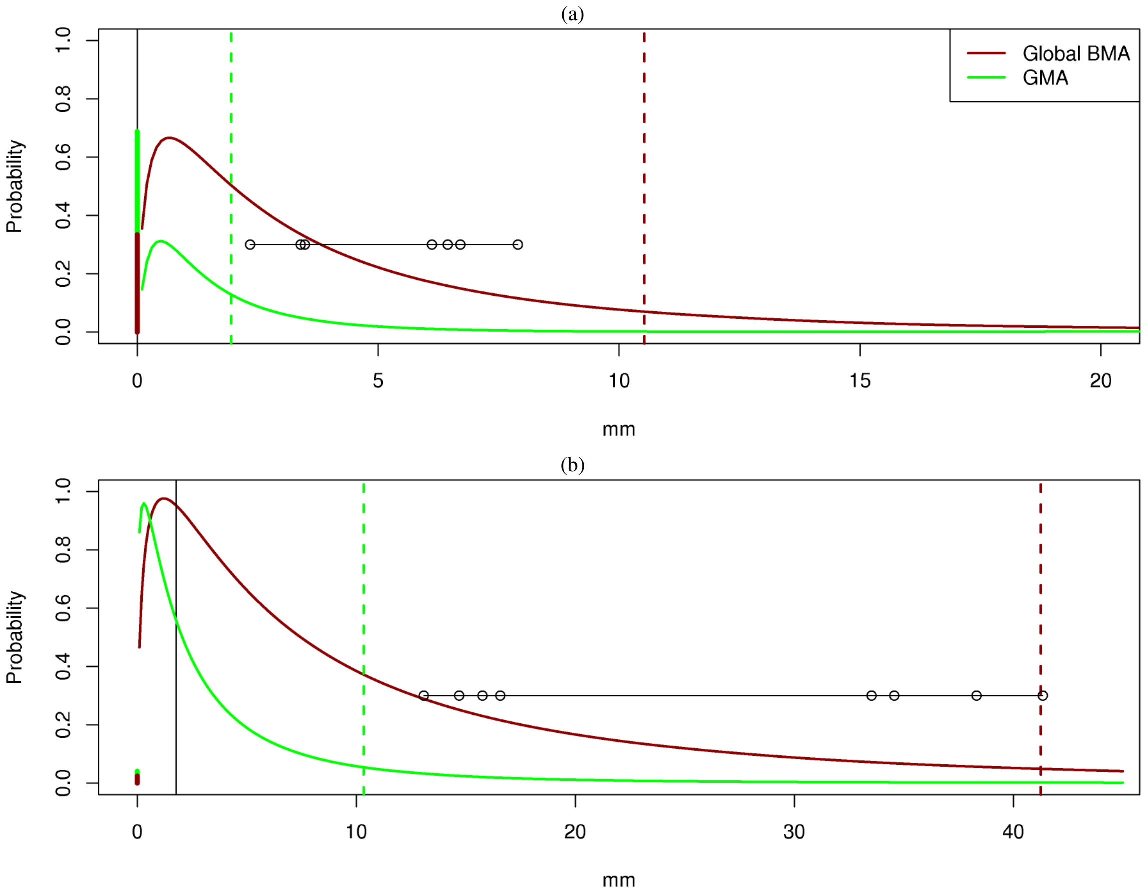

Figure 7. Forty-eight-hour-ahead predictive densities for daily precipitation accumulation at Mountain Home using GMA and Global BMA valid (a) January 19, 2008 and (b) November 11, 2008, respectively. The eight-member ensemble is represented by dots, and the realizing observation by the black vertical line. Dashed lines indicate the lower 90% prediction interval. The thick solid line segments at zero show the probability of no precipitation, with the lesser probability in the foreground.

real time, which is a critical concern here. The locally varying model parameters are estimated at each point in time using a variety of techniques, ranging from least squares estimates to posterior means. We also considered a more fully Bayesian approach that used posterior predictive distributions, rather than conditioning on point estimates. The results were comparable to those of our conditional approach, much in line with the experiences of Vrugt, Diks, and Clark (2008). In addition to failing to improve the predictive performance, a major drawback of using a fully Bayesian approach is the extra effort required to sample from posterior and conditional distributions, which is unlikely to be undertaken, or even to be feasible, in real time. Our conditional approach produces locally sharp and calibrated forecasts, and is faster than using full posterior predictive distributions. Di Narzo and Cocchi (2010) also carried out a fully Bayesian analysis similar to that of Vrugt, Diks, and Clark (2008), but they did not show any performance comparisons with the simpler method of Raftery et al. (2005).

While we consider precipitation here, GMA was originally developed for the technically less demanding case of surface temperature (Kleiber et al. 2011). Versions of GMA that apply to wind speed and wind vectors, which are also highly affected by local terrain, are highly desirable, especially with the recent surge of interest in wind energy. For wind speed, this could be achieved by combining the work of Sloughter, Gneiting, and Raftery (2010) with the ideas in our article.

Our approach, and that of Sloughter et al. (2007), is to model the occurrence of precipitation using logistic regression and, conditional on precipitation occurrence, the precipitation amount using a gamma density. Other authors have considered alternative approaches to postprocessing ensemble forecasts of quantitative precipitation, including nonparametric methods (Brown and Seo 2010) and logistic regression at a number of thresholds (Hamill and Whitaker 2006; Hamill, Hagedorn, and Whitaker 2008). Recently, Wilks (2009) proposed an appealing development that avoids inconsistencies between logistic regression equations at distinct thresholds and results in full predictive distributions. Schmeits and Kok (2010) compared his approach to BMA, with the methods performing similarly. The basic idea of GMA, namely locally varying model parameters, can be implemented in the logistic regression framework as well.

While GMA yields locally calibrated probabilistic forecasts at any given individual location and prediction horizon, it does not address the problem of joint calibration across meteorological variables, locations, and/or prediction horizons. A major challenge for the future, with further benefits in key applications such as flood management or air traffic control, is to combine our postprocessing approach with that of Berrocal, Raftery, and Gneiting (2008), which generates spatially correlated potential realizations of entire precipitation fields, while retaining the predictive density of Equation (1) at any individual location.

APPENDIX: LAPLACE METHOD

In Section 2.2 we discussed a Bayesian regularization approach that uses Laplace approximations to find the posterior means of the GMA logistic regression parameters. The Laplace method was introduced to statistics by Tierney and Kadane (1986) and rests on the approximation

$$\int \exp(g(\mathbf{a}))\, d\mathbf{a} \approx (2\pi)^{p/2} \det(-\mathbf{H}(\mathbf{m}))^{-1/2} \exp(g(\mathbf{m})),$$

where the integral is over $\mathbb{R}^p$, the function g has mode $\mathbf{m}$, and $\mathbf{H}(\mathbf{m})$ is the Hessian matrix of g evaluated at $\mathbf{m}$. Here we provide the details of our implementation.

The posterior means we seek can be written as

$$\int a_{is} q(\mathbf{a}|\mathbf{y})\, d\mathbf{a} = \left(\int a_{is} q(\mathbf{y}|\mathbf{a}) p(\mathbf{a})\, d\mathbf{a}\right) \times \left(\int q(\mathbf{y}|\mathbf{a}) p(\mathbf{a})\, d\mathbf{a}\right)^{-1} \qquad \text{(A.1)}$$

for $i = 0, 1, 2$, where $\mathbf{a} = (a_{0s}, a_{1s}, a_{2s})'$ is the parameter vector at the location s, $\mathbf{y} = (y_{1,s}, \ldots, y_{T,s})'$ is the vector of the binary observations of precipitation occurrence in the training set, q denotes a generic density, and p denotes the prior density. An application of the Laplace method to the terms on the right side requires the Hessian of the logarithm of the integrands. In particular, we need to find the Hessian of the log-likelihood, $\log q(\mathbf{y}|\mathbf{a})$, and the Hessian of the log prior, $\log p(\mathbf{a})$.

The likelihood function $q(\mathbf{y}|\mathbf{a})$ is a product of Bernoulli terms with success probability $p_t = \exp(\mathbf{a}'\mathbf{x}_t)/(1 + \exp(\mathbf{a}'\mathbf{x}_t))$ where $\mathbf{x}_t = (x_{0t}, x_{1t}, x_{2t})' = (1, f_{ts}^{1/3} - \overline{f_s^{1/3}}, \mathbb{1}_{[f_{ts}=0]})'$ for $t = 1, \ldots, T$. The Hessian of the log-likelihood function then has entry

$$(\mathbf{H}(\mathbf{a}))_{i+1,j+1} = -\sum_{t=1}^{T} x_{it} x_{jt} p_t (1 - p_t)$$

for $i, j = 0, 1, 2$. Note, in particular, that the Hessian matrix does not depend on the observation vector $\mathbf{y}$; it depends on the parameter vector $\mathbf{a}$ via the success probabilities $p_1, \ldots, p_T$. The prior density $p(\mathbf{a})$ factors as a product of independent normal densities with means μ_i and variances σ_i^2 for $i = 0, 1, 2$. The Hessian of the log prior thus is a diagonal matrix with entries $-1/\sigma_i^2$ for $i = 0, 1, 2$.

The parameters a_{is} may take on negative values, whence we cannot directly apply the Laplace method to the first term on the right side of Equation (A.1). Instead, we follow Tierney, Kass, and Kadane (1989) and find Laplace approximations for the integrals $\int (a_{is} + M) q(\mathbf{y}|\mathbf{a}) p(\mathbf{a})\, d\mathbf{a}$ and $\int M q(\mathbf{y}|\mathbf{a}) p(\mathbf{a})\, d\mathbf{a}$, where $M > 0$ is large. The difference of the two approximations provides the desired estimate.

The Laplace method is both fast and accurate, with the approximation error for posterior moments being of order $\mathcal{O}(T^{-2})$ (Tierney, Kass, and Kadane 1989). The most demanding task computationally is to find the mode of the logarithm of the integrand, but this is not an issue for the small training sets we use.

[Received July 2010. Revised June 2011.]

REFERENCES

Atger, F. (2003), "Spatial and Interannual Variability of the Reliability of Ensemble-Based Probabilistic Forecasts: Consequences for Calibration," *Monthly Weather Review*, 131, 1509–1523. [1291]

Bao, L., Gneiting, T., Grimit, E. P., Guttorp, P., and Raftery, A. E. (2010), "Bias Correction and Bayesian Model Averaging for Ensemble Forecasts of Surface Wind Direction," *Monthly Weather Review*, 138, 1811–1821. [1291]

Basist, A., Bell, G. D., and Meentemeyer, V. (1994), "Statistical Relationships Between Topography and Precipitation Patterns," *Journal of Climate*, 7, 1305–1315. [1293]

Berrocal, V. J., Gelfand, A. E., and Holland, D. M. (2010a), "A Bivariate-Space Time Downscaler Under Space and Time Misalignment," *The Annals of Applied Statistics*, 4, 1942–1975. [1300]

——— (2010b), "A Spatio-Temporal Downscaler for Output From Numerical Models," *Journal of Agricultural, Biological, and Environmental Statistics*, 15, 176–197. [1300]

Berrocal, V. J., Raftery, A. E., and Gneiting, T. (2008), "Probabilistic Quantitative Precipitation Field Forecasting Using a Two-Stage Spatial Model," *The Annals of Applied Statistics*, 2, 1170–1193. [1301,1302]

Berrocal, V. J., Raftery, A. E., Gneiting, T., and Steed, R. C. (2010), "Probabilistic Weather Forecasting for Winter Road Maintenance," *Journal of the American Statistical Association*, 105, 522–537. [1293,1297]

Brentnall, A. R., Crowder, M. J., and Hand, D. J. (2011), "Approximate Repeated-Measures Shrinkage," *Computational Statistics & Data Analysis*, 55, 1150–1159. [1295,1300]

Brier, G. W. (1950), "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3. [1296]

Brown, J. D., and Seo, D.-J. (2010), "A Nonparametric Post-Processor for Bias-Correction of Hydrometeorological and Hydrologic Ensemble Forecasts," *Journal of Hydrometeorology*, 11, 642–655. [1301]

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C. (1995), "A Limited Memory Algorithm for Bound Constrained Optimization," *SIAM Journal on Scientific Computing*, 16, 489–509. [1293]

Clogg, C. C., Rubin, D. B., Schenker, N., Schultz, B., and Weidman, L. (1991), "Multiple Imputation of Industry and Occupation Codes in Census Public-Use Samples Using Bayesian Logistic Regression," *Journal of the American Statistical Association*, 86, 68–78. [1293]

Cressie, N. A. C. (1993), *Statistics for Spatial Data* (revised ed.), New York: Wiley. [1295]

Czado, C., Gneiting, T., and Held, L. (2009), "Predictive Model Assessment for Count Data," *Biometrics*, 65, 1254–1261. [1296]

Daly, C., Neilson, R. P., and Phillips, D. L. (1994), "A Statistical-Topographic Model for Mapping Climatological Precipitation Over Mountainous Terrain," *Journal of Applied Meteorology*, 33, 140–158. [1293]

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum Likelihood From Incomplete Data via the EM Algorithm," *Journal of the Royal Statistical Society, Ser. B*, 39, 1–38. [1295]

Di Narzo, A. F., and Cocchi, D. (2010), "A Bayesian Hierarchical Approach to Ensemble Weather Forecasting," *Journal of the Royal Statistical Society, Ser. C*, 59, 405–422. [1301]

Dutton, J. A. (2002), "Opportunities and Priorities in a New Era for Weather and Climate Services," *Bulletin of the American Meteorological Society*, 83, 1303–1311. [1291]

Eckel, F. A., and Mass, C. F. (2005), "Aspects of Effective Mesoscale, Short-Range Ensemble Forecasting," *Weather and Forecasting*, 20, 328–350. [1292,1296]

Fraley, C., and Raftery, A. E. (2007), "Bayesian Regularization for Normal Mixture Estimation and Model-Based Clustering," *Journal of Classification*, 24, 155–181. [1293]

Fraley, C., Raftery, A. E., and Gneiting, T. (2010), "Calibrating Multi-Model Forecast Ensembles With Exchangeable and Missing Members Using Bayesian Model Averaging," *Monthly Weather Review*, 138, 190–202. [1295]

Gelman, A., Jakulin, A., Pittau, M. G., and Su, Y. S. (2008), "A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models," *The Annals of Applied Statistics*, 2, 1360–1383. [1294]

Genkin, A., Lewis, D. D., and Madigan, D. (2007), "Large-Scale Bayesian Logistic Regression for Text Categorization," *Technometrics*, 49, 291–304. [1294]

Gneiting, T. (2008), "Editorial: Probabilistic Forecasting," *Journal of the Royal Statistical Society, Ser. A*, 171, 319–321. [1291]

Gneiting, T., and Raftery, A. E. (2005), "Weather Forecasting With Ensemble Methods," *Science*, 310, 248–249. [1291,1295]

——— (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation," *Journal of the American Statistical Association*, 102, 359–378. [1296]

Gneiting, T., Balabdaoui, F., and Raftery, A. E. (2007), "Probabilistic Forecasts, Calibration and Sharpness," *Journal of the Royal Statistical Society, Ser. B*, 69, 243–268. [1296]

Grimit, E. P., and Mass, C. F. (2002), "Initial Results of a Mesoscale Short-Range Ensemble Forecasting System Over the Pacific Northwest," *Weather and Forecasting*, 17, 192–205. [1292]

Hamill, T. M., and Colucci, S. J. (1997), "Verification of Eta–RSM Short-Range Ensemble Forecasts," *Monthly Weather Review*, 125, 1312–1327. [1291,1296]

Hamill, T. M., and Juras, J. (2006), "Measuring Forecast Skill: Is It Real Skill or Is It the Varying Climatology?" *Quarterly Journal of the Royal Meteorological Society*, 132, 2905–2923. [1291]

Hamill, T. M., and Whitaker, J. S. (2006), "Probabilistic Quantitative Precipitation Forecasts Based on Reforecast Analogs: Theory and Application," *Monthly Weather Review*, 134, 3209–3229. [1301]

Hamill, T. M., Hagedorn, R., and Whitaker, J. S. (2008), "Probabilistic Forecast Calibration Using ECMWF and GFS Ensemble Reforecasts. Part II: Precipitation," *Monthly Weather Review*, 136, 2620–2632. [1301]

Hamill, T. M., Whittaker, J. S., and Wei, X. (2004), "Ensemble Reforecasting: Improving Medium-Range Forecast Skill Using Retrospective Forecasts," *Monthly Weather Review*, 132, 1434–1447. [1292]

Hastie, T., and Tibshirani, R. (1993), "Varying-Coefficient Models," *Journal of the Royal Statistical Society, Ser. B*, 55, 757–796. [1300]

Ihaka, R., and Gentleman, R. (1996), "R: A Language for Data Analysis and Graphics," *Journal of Computational and Graphical Statistics*, 5, 299–314. [1293]

Iversen, T. A. D., Santos, C., Sattler, K., Bremnes, J. O., Feddersen, H., and Frogner, I. (2011), "Evaluation of 'GLAMEPS'—A Proposed Multimodel EPS for Short Range Forecasting," *Tellus*, 63A, 513–530. [1291]

Joslyn, S., and Jones, D. W. (2008), "Strategies in Naturalistic Decision-Making: A Cognitive Task Analysis of Naval Weather Forecasting," in *Naturalistic Decision Making and Macrocognition*, eds. J. M. Schraagen, L. Militello, T. Ormerod, and R. Lipshitz, Aldershot, U.K.: Ashgate Publishing, pp. 183–202. [1295]

Katz, R. W., and Murphy, A. H. (eds.) (1997), *Economic Value of Weather and Climate Forecasts*, New York: Cambridge University Press. [1291]

Kleiber, W. (2010), "Multivariate Geostatistics and Geostatistical Model Averaging," Ph.D. thesis, Dept. of Statistics, University of Washington, Seattle, WA. [1293,1300]

Kleiber, W., Raftery, A. E., Baars, J., Gneiting, T., Mass, C. F., and Grimit, E. (2011), "Locally Calibrated Probabilistic Temperature Forecasting Using Geostatistical Model Averaging and Local Bayesian Model Averaging," *Monthly Weather Review*, 139, 2630–2649. [1291,1292,1298,1301]

Krzysztofowicz, R. (2001), "The Case for Probabilistic Forecasting in Hydrology," *Journal of Hydrology*, 249, 2–9. [1291]

Liu, C., and Rubin, D. B. (1994), "The ECME Algorithm: A Simple Extension of EM and ECM With Faster Monotone Convergence," *Biometrika*, 81, 633–648. [1295]

Mass, C. F. (2008), *The Weather of the Pacific Northwest*, Seattle, WA: University of Washington Press. [1292,1296]

Palmer, T. N. (2002), "The Economic Value of Ensemble Forecasts as a Tool for Risk Assessment: From Days to Decades," *Quarterly Journal of the Royal Meteorological Society*, 128, 747–774. [1291,1295]

Raftery, A. E., Gneiting, T., Balabdaoui, F., and Polakowski, M. (2005), "Using Bayesian Model Averaging to Calibrate Forecast Ensembles," *Monthly Weather Review*, 133, 1155–1174. [1291,1295,1301]

Schmeits, M. J., and Kok, K. J. (2010), "A Comparison Between Raw Ensemble Output, (Modified) Bayesian Model Averaging, and Extended Logistic Regression Using ECMWF Ensemble Precipitation Reforecasts," *Monthly Weather Review*, 138, 4199–4211. [1301]

Sloughter, J. M., Gneiting, T., and Raftery, A. E. (2010), "Probabilistic Wind Speed Forecasting Using Ensembles and Bayesian Model Averaging," *Journal of the American Statistical Association*, 105, 25–35. [1291,1301]

Sloughter, J. M. L., Raftery, A. E., Gneiting, T., and Fraley, C. (2007), "Probabilistic Quantitative Precipitation Forecasting Using Bayesian Model Averaging," *Monthly Weather Review*, 135, 3209–3220. [1291,1292,1295,1296,1300,1301]

Stein, M. L. (1999), *Interpolation of Spatial Data: Some Theory for Kriging*, New York: Springer-Verlag. [1295]

Stern, R. D., and Coe, R. (1982), "The Use of Rainfall Models in Agricultural Planning," *Agricultural Meteorology*, 26, 35–50. [1291]

Thorarinsdottir, T. L., and Gneiting, T. (2010), "Probabilistic Forecasts of Wind Speed: Ensemble Model Output Statistics by Using Heteroscedastic Censored Regression," *Journal of the Royal Statistical Society, Ser. A*, 173, 371–388. [1291]

Tierney, L., and Kadane, J. B. (1986), "Accurate Approximations for Posterior Moments and Marginal Densities," *Journal of the American Statistical Association*, 81, 82–86. [1294,1302]

Tierney, L., Kass, R. E., and Kadane, J. B. (1989), "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions," *Journal of the American Statistical Association*, 84, 710–716. [1302]

Vrugt, J. A., Diks, C. G. H., and Clark, M. P. (2008), "Ensemble Bayesian Model Averaging Using Markov Chain Monte Carlo Sampling," *Environmental Fluid Dynamics*, 8, 579–595. [1293,1295,1301]

Wilks, D. S. (2009), "Extending Logistic Regression to Provide Full-Probability-Distribution MOS Forecasts," *Meteorological Applications*, 16, 361–368. [1301]

Wong, R. (2001), *Asymptotic Approximations of Integrals*, New York: Academic Press. [1294]

Zhu, Y., Toth, Z., Wobus, R., Richardson, D., and Mylne, K. (2002), "The Economic Value of Ensemble-Based Weather Forecasts," *Bulletin of the American Meteorological Society*, 83, 73–84. [1291]