Categorical Climate Forecasts through Regularization and Optimal Combination of Multiple GCM Ensembles*

BALAJI RAJAGOPALAN

Department of Civil, Environmental, and Architectural Engineering, University of Colorado, Boulder, Colorado, and International Research Institute for Climate Prediction, Palisades, New York

UPMANU LALL

Department of Earth and Environmental Engineering and Columbia Earth Institute, Columbia University, New York, New York, and Department of Civil and Environmental Engineering and Utah Water Research Laboratory, Utah State University, Logan, Utah, and International Research Institute for Climate Prediction, Palisades, New York

STEPHEN E. ZEBIAK

International Research Institute for Climate Prediction, Palisades, New York

(Manuscript received 17 September 2000, in final form 14 December 2001)

ABSTRACT

A Bayesian methodology is used to assess the information content of categorical, probabilistic forecasts of specific variables derived from a general circulation model (GCM) forecast ensemble, and to combine a ''prior'' forecast (climatological probabilities of each category) with a categorical probabilistic forecast derived from a GCM ensemble to develop posterior, or ''regularized'' categorical probabilities. The combination algorithm assigns a weight to a particular model forecast and to climatology. The ratio of the sample likelihood of the model based on the posterior categorical probabilities, to that based on climatological probabilities, computed over the period of record of historical forecasts, provides a measure of the skill or information content of a candidate model. The weight given to a GCM forecast serves as a secondary indicator of its information content. Model weights are determined by maximizing the likelihood ratio. Results using the so-called ranked probability skill score as an objective function are also obtained, and are found to be very similar to the likelihood-based results.

The procedure is extended to the optimal combination of forecasts from multiple GCMs. An application of the method is presented for global, seasonal precipitation and temperature forecasts in two different seasons, based on 41 yr of observational and model simulation data. The multimodel combination skill is significantly better than climatology skill in only a few regions of the globe, but is generally an improvement over individual models, and over a simple average of forecasts from different models. Limitations and possible improvements of the methodology are discussed.

1. Introduction

The recent demonstration of skill in forecasting ENSO and its climatic teleconnections (Cane et al. 1986; Barnston et al. 1999a,b; Mason et al. 1999) has led to considerable interest in the use of climate forecasting models. The models in question are general circulation models that numerically integrate the governing equations of motion using prescribed boundary and initial conditions. Different models generally embody the same governing equations but vary in their parameterization of subgrid-scale processes such as clouds or convection and in their treatment of boundary conditions and data assimilation. These models generate forecasts of a large number $O(10^7)$ of state variables on a spatial grid over the planet. The model outputs may be used in turn as boundary conditions by land-surface hydrologic, ecologic, and other models to provide forecasts of ''local'' state variables relevant to these systems. Uncertainties in the boundary and initial conditions, as well as biases in process parameterizations propagate through this process. Further, nonlinear interactions introduce chaotic behavior and loss of predictability with increasing forecast lead times. To deal with this limitation, forecasts are commonly provided as ensembles (typically 5–10

* International Research Institute Contribution Number IRI-PP/02/01.

Corresponding author address: Dr. Stephen E. Zebiak, International Research Institute for Climate Prediction, 61 Rt. 9W, Monell Building, Palisades, NY 10964-8000.
E-mail: steve@iri.columbia.edu

members), generated through small perturbations of the initial and boundary conditions. Often forecast ensembles can differ significantly across models, and from the observed climate state. Consequently, there is interest in developing methods for combining such ensembles in a way that makes the resulting ensemble more representative of the observed data.

A priori, we expect model estimates of the evolution of surface state variables to be biased relative to observations. This is due in part to limitations in process and boundary condition representation, and in part to the disparity in the nature of model outputs and surface observations. The model space and time discretization does not correspond to the sampling domain for surface observations. Thus, regardless of whether one interprets the model variables as point values or grid averages, even under perfect physics, we cannot expect model outputs to correspond exactly to fields interpolated from surface observations, particularly where there is considerable subgrid heterogeneity. These biases may be manifest in the marginal (i.e., long term) distribution of the state variable, in the conditional (e.g., forecast of next season) distribution, or in both.

In this paper a framework is developed under which a categorical, probabilistic forecast can be derived from a GCM ensemble, allowing for an empirical correction of the bias (GCM vs observations) in the marginal distribution of the state variable at the location of interest. A Bayesian approach for the evaluation of forecast skill and regularization of categorical probability forecasts is developed and applied. It is presumed that a reasonably long (e.g., 30–40 yr) sequence of past model forecasts and corresponding historical data is available. This series is used for skill assessment and model regularization. The procedure allows one to assess the relative probability with which one model (e.g., a raw GCM forecast) is likely to generate the correct category of outcome when compared with another model (e.g., climatology, another GCM, or a statistical forecast). We can also develop probabilistic categorical forecasts that reflect the model skill—automatically tending to climatology where skill is low, and to raw model forecast where skill is high. A by-product of the procedure is an estimate of the uncertainty distribution of the category probabilities for a given forecast. These may be useful for Monte Carlo analysis of decisions that use the forecast as input. The framework presented is extended to the simultaneous use and combination of multiple forecast models. Results for precipitation and temperature forecasts over the entire globe are presented.

2. Background

Two classes of methods have been used for multimodel ensemble forecast combination. A simple way to combine the ensembles from different models is to pool ensembles for all models into a single ensemble (e.g., Palmer et al. 2000). Another approach uses best linear unbiased estimates (Fraedrich and Smith 1989; Pavan and Doblas-Reyes 2000) or linear regression (Krishnamurti et al. 1999, 2000). A method akin to linear regression has been used operationally for some time at the National Centers for Environmental Prediction (NCEP) to combine individual forecasts of seasonal climate (A. Barnston 2001, personal communication). Several papers under the Prediction of Climate Variations on Seasonal Timescales (PROVOST) project used multimodel ensembles (using four different European climate models) for forecasts of several variables across the globe and reported improvements over models using a single ensemble [we refer the readers to the DSP/PROVOST issue, vol. 126(567) of the *Quarterly Journal of Royal Meteorological Society*]. The increased effective sample size of the ensemble by pooling information from different models presumably offers a reduction in the variance of the statistics (e.g., mean) of the forecast. However, given the likely differing linear and nonlinear biases in each model, it is not clear that such a strategy will always be effective. Root-mean-square error or the ranked probability skill score (Epstein 1969; Wilks 1995) have been used as the measures of performance. The bias and variance components of the root-mean-square error are typically not identified separately. It is not clear whether the improved skill from the regression approaches is superior to the variance reduction that would be obtained by simply averaging across the individual models, and whether the improvement is a result largely of the bias correction for each model.

Quasi-objective methods for combining model ensembles have been developed (Mason et al. 1999) and used previously in the seasonal forecasts issued by the International Research Institute (IRI) for Climate Prediction. The state space of the variable of interest is divided into three categories or ''terciles'' using the historical station/regional data. Forecasts are then issued as probabilities of outcomes for each of three terciles for the variable. The method combines the ensemble forecasts from the models based on past performance of individual models, prior knowledge from empirical studies, and other information. The actual combination itself is done in a subjective manner. The belief is that probabilistic forecasts may be more useful for applications than the forecast of the conditional mean (as obtained by a regression approach), and that the amount of information available can perhaps justify only a three-category forecast.

The combination method developed here provides a fully objective counterpart to the previous probabilistic forecast methods at IRI and elsewhere. In principle, it allows for more extensive retrospective evaluations, and more flexibility for changes in models, ensemble sizes, and other parameters. These were primary motivations for the work. Through the analysis of individual model skills relative to combined model skill, we are able to identify the contribution of multimodel combination,

over and above straightforward bias correction, in improving forecast performance.

3. Approach

a. Preliminaries

Consider as a starting point the examination of a single state variable at a single location, with the following information available:

1) A historical time series of the state variable, x_t, $t = 1, \ldots, n$; where n is the length of record. In the present context of seasonal forecasts, n can be taken to be the number of years of record.
2) For each GCM, there is a corresponding ensemble time series of forecasts, denoted y_{jt}, $t = 1, \ldots, n$, $j = 1, \ldots, m$; where m is the number of ensemble members.
3) The user intends to classify the data into K categories. The categories may or may not be equally spaced. A familiar example is the tercile ($K = 3$, and the categories are derived from breakpoints, such that the marginal probability associated with each category is 1/3). One could also consider $K = n + 1$, and use the ranked observations as breakpoints for the categories. This is equivalent to a statement about the empirical cumulative distribution function of x, since each ranked observation demarcates a specific empirical percentile of the marginal distribution of x. As precision increases with increasing K, reliability of the assertion of the probabilities in each category decreases, for fixed n, due to sampling variability.

First, preprocessing of the observational data and model outputs into K categories is done by identifying the set of $K - 1$ breakpoints that represent specific percentiles of the respective distributions, the model counterpart derived from the aggregate of all m ensemble members. For example, the (approx.) 33d and 67th percentiles of the observation and model distributions would be used to categorize the respective data for the case of standard terciles. Then it is possible to consider the sequence of observation category outcomes for each of the n time points to be described by a discrete random variable X_t, which assumes integer values between 1 and K. Similarly, the sequence of model category outcomes, for each of the m ensemble members can be described by a random variable Y_t, assuming the same range of integer values.

If the model is unbiased (and n is large enough), the breakpoints of the two distributions will coincide. In general, there will be some model bias, and the breakpoints will therefore differ. But by dealing only with the categorical information as defined above, this overall bias is effectively removed. Of course, there remains the possibility of conditional bias, which would contribute significantly to degrading actual forecast performance.

b. Probabilities and uncertainty

Now, consider the following candidate probabilistic forecasts:

Climatology:

$$P_{kt}(x) = P_k(x), \qquad k = 1, \ldots, K, \quad t = 1, \ldots, n, \tag{1}$$

GCM:

$$P_{kt}(y) = m_{kt}/m, \qquad k = 1, \ldots, K, \quad t = 1, \ldots, n, \tag{2}$$

where $P_{kt}(x)$ is the probability of drawing a state variable value that falls in category k, in year t, using only the marginal distribution, $P_k(x)$; $P_{kt}(y)$ is the probability of drawing a state variable value that falls in category k, in year t, using only the GCM forecast ensemble; m_{kt} is the number of GCM ensemble members that fell in category k in year t; and m is the total number of ensemble members. $P_k(x)$ is equal to $1/K$ for categories based on equal-sized percentile ranges [e.g., for terciles, $P_k(x) = 1/3$].

If the category breakpoints are considered to be fixed at the specified values, and the probabilities $P_k(x)$ are estimated from different sample realizations (each of length n), the estimated probabilities would vary across realizations. For example, consider tossing a six-faced die. If the die is repeatedly tossed n (e.g., 40) times, and the probability of each face coming up is computed, the estimated probability of each face will be different for each set of n tosses. Clearly as the sample size, n, used to define the percentiles increases, the variance of estimation decreases. Thus, one can think of the $P_k(x)$ as random variables, given fixed category definitions (that were selected, of course, based on the percentiles of the n year sample). An appropriate prior distribution for the $P_k(x)$ given the multinomial process X_t, is the Dirichlet distribution $D(\mathbf{a})$:

$$f(\mathbf{P}) = D(\mathbf{a}) = (B(\mathbf{a}))^{-1} \prod_{k=1}^{K} P_k^{a_k - 1}, \tag{3}$$

where $\mathbf{P}$ is the vector of category probabilities, and $B(\mathbf{a})$ is a generalized beta function defined as

$$B(\mathbf{a}) = B(a_1, a_2, \ldots, a_K) = \frac{\prod_{k=1}^{K} \Gamma(a_k)}{\Gamma\left(\sum_{k=1}^{K} a_k\right)}, \tag{4}$$

where Γ is the Gamma function, and a_k is a scale parameter analogous to the number of outcomes in category k (i.e., $n/3$ for terciles). The multinomial process and associated Dirichlet distribution are straightforward extensions of the more familiar Bernoulli process and binomial distribution, where now rather than two (mutually exclusive) outcomes and one process parameter there are K, and $K - 1$, respectively (in our application $K = 3$).

Defining $\alpha = \sum_{k=1}^{K} a_k$,the moments of this distribution are (O'Hagan 1994, p. 279):

$$E[P_k] = a_k/\alpha$$

$$\text{var}[P_k] = a_k(\alpha - a_k)/[\alpha^2(\alpha + 1)]$$

$$\text{cov}[P_k, P_l] = -a_k a_l/[\alpha^2(\alpha + 1)] \quad (5)$$

For the case of tercile or other equiprobable categories, it is clear that for a fixed K, the properties of this distribution are determined solely by the parameter α, the sample size used for determining the probabilities, and that the variance decreases as α increases.

The above presentation is also relevant for the GCM forecast-derived probabilities $P_{kt}(y)$, but in this case based on a sample of size m. Thus, Eqs. (3) though (5) can be considered applicable for both situations, with the proper interpretation of the parameters a_k and α.

c. Forecast skill and regularization

Here, the first objective is to evaluate whether the ensemble forecast from a particular GCM has skill. This problem may be considered in the following manner. For each year there is a prior probability distribution for the $P_{kt}(x)$, defined through Eqs. (3) and (4), and the assertion that $a_k = nP_k$. This corresponds to the climatology forecast and its associated uncertainty. For the same process, a GCM ensemble forecast is presented for each year, expressed as $P_{kt}(y)$. The uncertainty in this forecast each year could be described using Eqs. (3) through (5), and the parameters m_{kt}, and m, corresponding to the a_k and a. Next, consider how the two sources of information might be merged objectively, and how the process can provide a measure of model forecast skill. It is useful to consider the various possibilities as follows.

Suppose a prior belief is that the GCM really draws from the marginal distribution of the data, and has no conditional or forecast information. In this case, the GCM ensemble in a given year simply provides additional versions of possible state variable values that can be used to improve prior (climatological) estimates of the category probabilities, but does not provide specific utility for a seasonal forecast. Jointly using the GCM ensemble and the climatological forecast then could increase the effective sample size α, and thus improve the precision of the estimates of the category probabilities. If under this noninformative GCM scenario, it is considered that each year of record is equal, in information, to one GCM ensemble member, then the optimal weight for combining the GCM and climatology forecasts would be proportional to the sample sizes: $m/(m + n)$ and $n/(m + n)$, respectively. These are designated as the ''prior odds ratio'' for GCM and climatology, respectively.

Alternatively, suppose that the GCM has considerable skill in conditional forecasting and its ensemble is capable of representing the underlying probabilities of events in different categories. However, given the relatively small sample size (m or α) for the ensembles, the associated probabilities, $P_{kt}(y)$, have high uncertainty. Recognizing that model performance reflects a trade-off between bias and variance, we may be willing to pool the GCM ensemble forecast (low bias and low variance) with the climatological forecast (high conditional bias but lower variance). This scenario would lead to a weight for climatology that is smaller than the prior odds ratio $n/(m + n)$.

Finally, suppose that the model-based forecasts are seriously flawed, in the sense that the conditional probabilities indicated by the ensembles have much higher variance than expected for a sample of size m, but on average they differ little from the climatological probabilities (by construction if the bias in the marginal distribution of the model is removed). In this case, if the climatology forecast and the model forecast are pooled in some way, it would be desirable to reduce the weight given to the model forecast, below its prior odds ratio of $m/(m + n)$, corresponding to the GCM providing a noninformative forecast where the ensemble variance is representative of climatology.

A method is sought to arrive at a formulation for combining the two forecasts, and assigning appropriate weight to each, taking into account the effect of sample sizes for both climatology and model forecasts. This can be done quite readily using a Bayesian approach, since the uncertainty in both sources of information is provided by a single parameter of the Dirichlet distribution. The reasoning proceeds as follows.

Climatology and its uncertainty provide the prior forecast distribution. This prior belief is updated each year with the GCM forecast probabilities and their associated uncertainty, to provide a posterior probability distribution for the forecast. Given the use of the Dirichlet distribution as a conjugate distribution for the multinomial process, it can be shown (O'Hagan 1994, p. 279) that the posterior distribution resulting from the combination of the two sources of information, with parameters $\mathbf{a}$ and $\mathbf{b}$ is also Dirichlet with parameter ($\mathbf{a} + \mathbf{b}$). Here, consider a weighted combination of the climatology and the GCM ensemble forecast, where the weight is to be determined as part of an optimization process.

The posterior probabilistic forecast for each year can be denoted thus

$$f(\mathbf{Q}_t|\mathbf{P}_t(y)) = D(\mathbf{c}_t) \qquad \mathbf{c}_t = \mathbf{a} + \mathbf{b}_t$$

$$a_k = nP_k(x) \qquad k = 1, \ldots, K$$

$$b_{kt} = wmP_{kt}(y) \qquad k = 1, \ldots, K,$$

$$t = 1, \ldots, n, \quad (6)$$

where $\mathbf{Q}_t$ is a vector of posterior probabilities for each of the categories for time step t; w is a weight or regularization parameter applied to the ensemble sample size m, such that wm represents the effective sample

size of the GCM ensemble forecast relative to the sample size (n) of the climatology forecast.

The mean and variance of the posterior categorical probability forecast may then be defined following Eq. (5), as

$$\begin{aligned} E[\mathbf{Q}_{kt}] &= c_{kt}/c \\ \text{var}[\mathbf{Q}_{kt}] &= c_{kt}(c - c_{kt})/[c^2(c + 1)] \\ c_{kt} &= (a_k + b_{kt}) = nP_k(x) + wmP_{kt}(y) \\ c &= \sum_{k=1}^{K} c_{kt} = \alpha + b \qquad \alpha = n \\ b &= wm \end{aligned} \tag{7}$$

Here, c can be thought of as the effective sample size associated with the combined forecast. It is the weighted sum ($n + wm$) of the sample size for climatology and of the GCM ensemble size. Correspondingly, c_{kt} reflects the effective number of counts in category k for year t, subsequent to the weighted combination of the climatology and the GCM forecasts. In practice, the raw ensemble probabilistic forecast $P_{kt}(y)$, would be replaced by the average posterior probabilities for each category, $E[\mathbf{Q}_{kt}]$, and its variance would be known from Eq. (7), as well as an uncertainty distribution from Eq. (6). For a tercile forecast the expected value of the posterior probability for category k is $(n/3 + wm_{kt})/(n + wm)$. Note that this is a function of both the sample size for climatology and the size of the model ensemble, as is desired. Since one can make either climatology or the model ensemble dominate the estimate through the choice of w, it is clear that the choice of w needs to reflect the information content associated with a climatology sample of size n, and a model ensemble of size m. Changing either n or m requires a new choice of w.

The selection of w constitutes an optimization problem, the result depending on the choice of skill measure that is to be optimized. An appropriate choice is the posterior likelihood function, defined over the N year common record available of historical and model data, at a particular grid location. This has the form

$$L(w) = \prod_{t=1}^{N} E(Q_{k^*t}), \tag{8}$$

where k^* represents the category actually observed to occur at each time t. Thus $L(w)$ simply reflects the product over all times (years) of the forecast probabilities assigned to the correct category. It represents an integration of the ''performance'' of the candidate model over a run of events. Two different models can be compared in terms of their likelihood ratio L_2:L_1 which represents the ratio of the probability that model 2 is appropriate relative to model 1. Consider a coin-tossing example ($K = 2$). Let us say that one has a fair coin (p(Heads) = p(Tails) = 0.5). For the first example say that one has misspecified the model and believes that p(Heads) is 0.8. Now over a large number of coin tosses, one expects to get roughly half heads, and the likelihood of the model is $0.8^{0.5n}\ 0.2^{0.5n}$, which is 1.05×10^{-4} for $n = 10$; 1.61×10^{-40} for $n = 100$, as compared to the null model ($p = 0.5$) for the fair coin, which gives 9.77×10^{-4} for $n = 10$; 7.89×10^{-31} for $n = 100$, leading to a likelihood ratio of 9.31 for $n = 10$; 4.91×10^{9} for $n = 100$ in favor of the fair coin model.

A general purpose nonlinear optimization algorithm called Feasible Sequential Quadratic Programming (FSQP) (Zhou and Tits 1993) was used to maximize $\log(L(w))$ subject to the constraint that w be positive. Zhou and Tits (1993) show that this algorithm is globally convergent and locally superlinear convergent. A succession of quadratic programs is solved to determine the optimal solution formed by Taylor series approximations of the functions at each solution point (see also Luenberger 1989).

The relative performance of two models may be compared either in terms of the magnitude of the regularization parameter, w, (or equivalently the ratio wm/n), or in terms of the ratio of their sample likelihood. The weight measure (wm/n) is interpretable in terms of the implied sample size associated with the forecast relative to using a n year climatology record. If the ratio wm/n is greater than 1, then on average the GCM ensemble forecast has information content that exceeds what would be provided by a n year record and the associated climatology forecast. As this ratio decreases, the information added by the GCM ensemble forecast decreases, and in particular for $w = 1$, the m member GCM ensemble forecast contributes simply m noninformative (in the sense of a conditional forecast) years to the climatology forecast. Values of w less than 1 suggest that individual ensemble forecast members contribute less information than a year drawn at random from the climatological record.

The likelihood ratio provides a measure of the relative probability with which two competing models are likely to represent the actual outcome in a sequence of tests of the models. Recognizing that users may often want to know how likely one model is to be accurate in a single event on average, rather than averaged over an arbitrary sample size, it is useful in this context to report the likelihood ratio normalized for the sample size over which it is evaluated:

$$\text{LR}_{ij} = \left(\frac{L_i}{L_j}\right)^{1/n} \tag{9}$$

For the coin-tossing example presented earlier, this translates into ratios of 1.25 in favor of the null model for both the sample sizes (10 or 100) considered.

While one can directly compare a GCM to climatology or to another GCM using Eq. (9), given the existing small ensemble size, this may not be feasible. If the model ensemble does not show any probability of a

solution in a particular category for a given year, and the actual outcome lies in that category, then the probability of the model being correct that year (and hence its full sample likelihood) will be zero. This precludes a useful comparison of two models without a further modification of either the likelihood ratio measure, or a prior smoothing [e.g., by kernel methods; see Rajagopalan and Lall (1995)] of the model ensemble probabilities. On the other hand, the optimal combined forecast always gives nonzero probabilities in all categories because of the finite contribution of climatology, so one can adopt the ratio of likelihood based on the posterior probability distribution to that of climatology as a measure of regularized model performance. Since the denominator in all such comparisons is the fixed likelihood associated with a climatology model, the likelihood ratio (LR) of two models is readily computed as the ratio of their base LR values with respect to climatology.

It is often of interest to estimate the changes in apparent model skill as the number of categories (K) is varied. As noted earlier, a small K gives a coarse representation of the underlying cumulative density function, while a large K attempts to more closely approximate the underlying density. However, for a fixed sample size available for testing and validation, a larger K may induce a much higher variability. Potentially, models formulated with different numbers of categories could also be compared with each other in terms of their likelihood ratio.

The selection of w through the maximization of the posterior likelihood function using FSQP has associated sampling variability. If the common period of record for forecasts and historical data used for validation (N) is small, one can expect a high degree of variability in the selection of w across different realizations or grid cells. For the coin toss example, one can readily visualize that if N is 10, due to sampling variability in the outcome of the coin toss, one may choose the biased probability model (p(Heads) = 0.8) as correct in repeated tests, more often (as a percentage) than if N was 100. As K increases, the number of parameters to be estimated increases, and hence the effective degrees of freedom of the scheme decreases, leading to increased variability in estimating w or LR. The use of a fully Bayesian approach based on hierarchical modeling (Gelman et al. 1995) that explicitly treats w as a random variable across all grid nodes, with a specified prior distribution, would be promising in this regard. This approach would allow one to build spatial and/or temporal structure into the choice of w. We expect to pursue such a strategy in the future.

d. Generalization to a combination of multiple forecast models

The procedure described above generalizes readily to the development of a posterior probability forecast through a combination of forecasts from J different models (including a climatology forecast). Now the posterior probability distribution can be rewritten by analogy to Eq. (6) as

$$f(\mathbf{Q}_t|\mathbf{P}_{tj}(y), j = 1 \ldots J) = D(\mathbf{c}_t) \qquad \mathbf{c}_t = \sum_{j=1}^{J} \mathbf{a}_{jt}$$

$$a_{jkt} = w_j m_j P_{kij}(y)$$

$$k = 1, \ldots, K,$$

$$j = 1, \ldots, J,$$

$$t = 1, \ldots, n$$

$$c = \sum_{j=1}^{J} w_j m_j$$

$$E[Q_{kt}] = \frac{\sum_{j=1}^{J} w_j m_j P_{ktj}(y)}{\sum_{j=1}^{J} w_j m_j}, \tag{10}$$

where m_j is the size of the ensemble for model j ($m_j = n$ for climatology), and w_j is the weight ascribed to model j.

The weights w_j are selected by maximizing the posterior likelihood function defined as before [Eq. (8)], under the constraint that each w_j is positive. The weights are normalized to sum to 1 for presentation purposes after the solution of the maximization problem. If the ensemble size for each model is the same, the weights directly reflect the relative information content of each model. An information measure for the combination of models equivalent to the w for the single model case is provided by w_{j^*}, where j^* is the index of the climatology model. Higher values of w_{j^*} will indicate lower information content in the models and viceversa. The posterior model formed as the combination of multiple forecast models can also be compared with simpler models using the likelihood ratio defined earlier.

4. Application

The model evaluation and combination procedures described in the previous section were applied to global precipitation and temperature forecasts obtained from three atmospheric general circulation models (AGCMs) that the IRI currently uses for its real-time forecasting. The three models are ECHAM3 (from the Max Plank Institute), MRF9 (from NCEP), and CCM3 (from the National Center for Atmospheric Research). Further descriptions are provided in Mason et al (1999). All three spectral models were run at a T42 (or T40) resolution (approximately 2.8° latitude and longitude) with vertical resolution of 18 (or 19) layers. The models were run in a simulation mode, forced with observed sea surface temperatures (SSTs) for the period 1950–91. An ensemble of 10 runs with identical SST forcing but differing initial conditions was available for each AGCM. Thus

we have 10 ensemble simulations of precipitation and temperature for each month from each model, at each location globally.

In the results presented here the focus is on two seasons, April–June (AMJ), and January–March (JFM). These represent rainy seasons for particular regions that will be emphasized. The results are representative of the overall results, except that the relative performance of individual models can differ from region to region. Tercile forecasts ($K = 3$) were considered, to be consistent with the practice in the IRI net assessment forecasts. For each season, for each variable of interest (precipitation, surface temperature), at each grid location the following was done:

1) Estimate the regularization parameter w_j for model j by maximizing the posterior likelihood function defined in Eq. (8), using the full record of 1950–90 (i.e., $n = N = 41$).
2) Estimate the associated LR, using Eq. (9).
3) Repeat steps 1 and 2 for the combination of all three models.

Spatial maps of these estimates were generated and compared to see the relative performances of the models over different regions, variables, and seasons. In the current study, the epochal variation (e.g., different decades or ENSO phases) in the performance of the different models has not been investigated.

The confidence levels for the LR are computed using a bootstrapping procedure that scrambles the observations in time while maintaining their spatial structure. The predictor data fields from the GCMs are not randomized. The time series of the predictand (i.e., the observed precipitation or temperature) is drawn at random with replacement from the original time series. For each year sampled, the entire spatial field corresponding to a historical year is drawn at random. This is done to preserve the spatial correlation structure in the predictor and predictand fields. The LR values at each grid point (2861 grid points in all) are then computed. Ten such realizations of 41 yr each are drawn. This provides 10 $\times$ 2861 LR values, from which the 90th, 95th, and 99th percentile are obtained. Although a larger sample size is typically preferred for bootstrapping, very small variability in the LR values corresponding to these percentiles across realizations at the level of 10 simulations were observed. The significance levels are estimated separately for each case—single predictive model versus the multimodel combination—to reflect the changed problem definition.

5. Results

Likelihood ratios and weights for the different models and for their combination, for both precipitation and temperature forecasts, and for all seasons, were obtained. A few selected results are presented here. All results are based on the time period 1950–91.

a. Individual model skill

Figure 1 shows the spatial map of LR for JFM precipitation for the three models. The three colors—blue, yellow, and red—indicate regions where the LR attains significance at the 90%, 95%, and 99% confidence levels, respectively. All three models have statistically significant LR values in regions with well-known ENSO teleconnections (the western Pacific, and northeastern South America). In addition the ECHAM model shows skill over southern parts of the United States. NCEP appears to show significant skill over South Africa, while CCM shows skill over eastern Africa and southern South America. It is noted that ECHAM exhibits rather broad areas where the LR value was marginally greater than 1, but below the threshold value for statistical significance. For all models, the majority of land area of the world shows no significant skill relative to climatology. This is consistent with many previous studies—the result of large internal variability, particularly in the extratropics, and only a limited set of externally forced signals, each with limited domain of influence.

Figure 2 shows the spatial map of model weights. Note that the weights are normalized, such that for $N = 40$ and $m = 10$, a weight of 10/50 [i.e., $wm/(n + wm)$] or 0.2 is equivalent to $w = 1$ in the original notation, which was the threshold for a model ensemble member to be treated as the equivalent of one year of climatology. The spatial maps of the weights are generally consistent with the LR maps. In areas where individual models have relatively high LR scores, a relatively high weight is assigned (e.g., South Africa for NCEP, southern South America for CCM, west coast of the United States for ECHAM). There are many areas, especially in the midlatitudes where the GCMs perform worse than climatology (weight below 0.2, the white space in the figure indicates a weight of 0).

Figures 3 and 4 show the spatial maps of LR and model ensemble weight for temperature for the AMJ season. All the models generally show significant LR values in the western Pacific, northern and northeastern parts of South America, southern United States, and southwestern Africa. Note that the LR values are relatively high over most near-coastal or island locations in all of the models. This result is expected due to the fact that the model simulations analyzed here are run with prescribed observed SST, which should influence significantly and preferentially the coastal and island regions. The skill levels should be expected to diminish in a true forecast situation, where SST must also be predicted. The model weights are generally consistent with the LR maps. The majority of regions outside the Tropics and away from the coasts once again show no significant skill in the models. Within the Tropics, notably larger areas with significant LR values and weights are obtained for temperature than for precipitation.

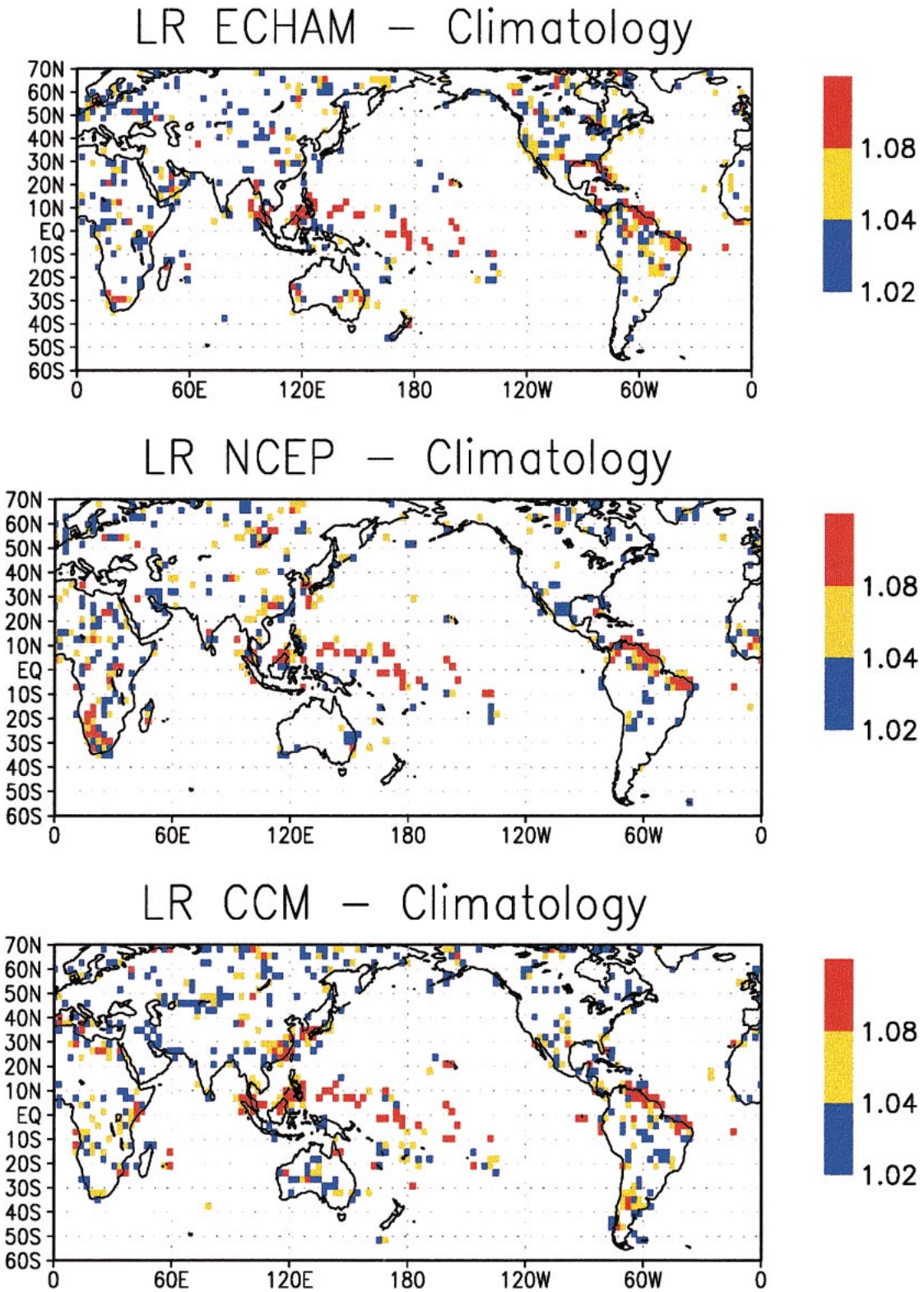

FIG. 1. Spatial map of LR values for precipitation for the JFM season, for the three models, when considered individually. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test.

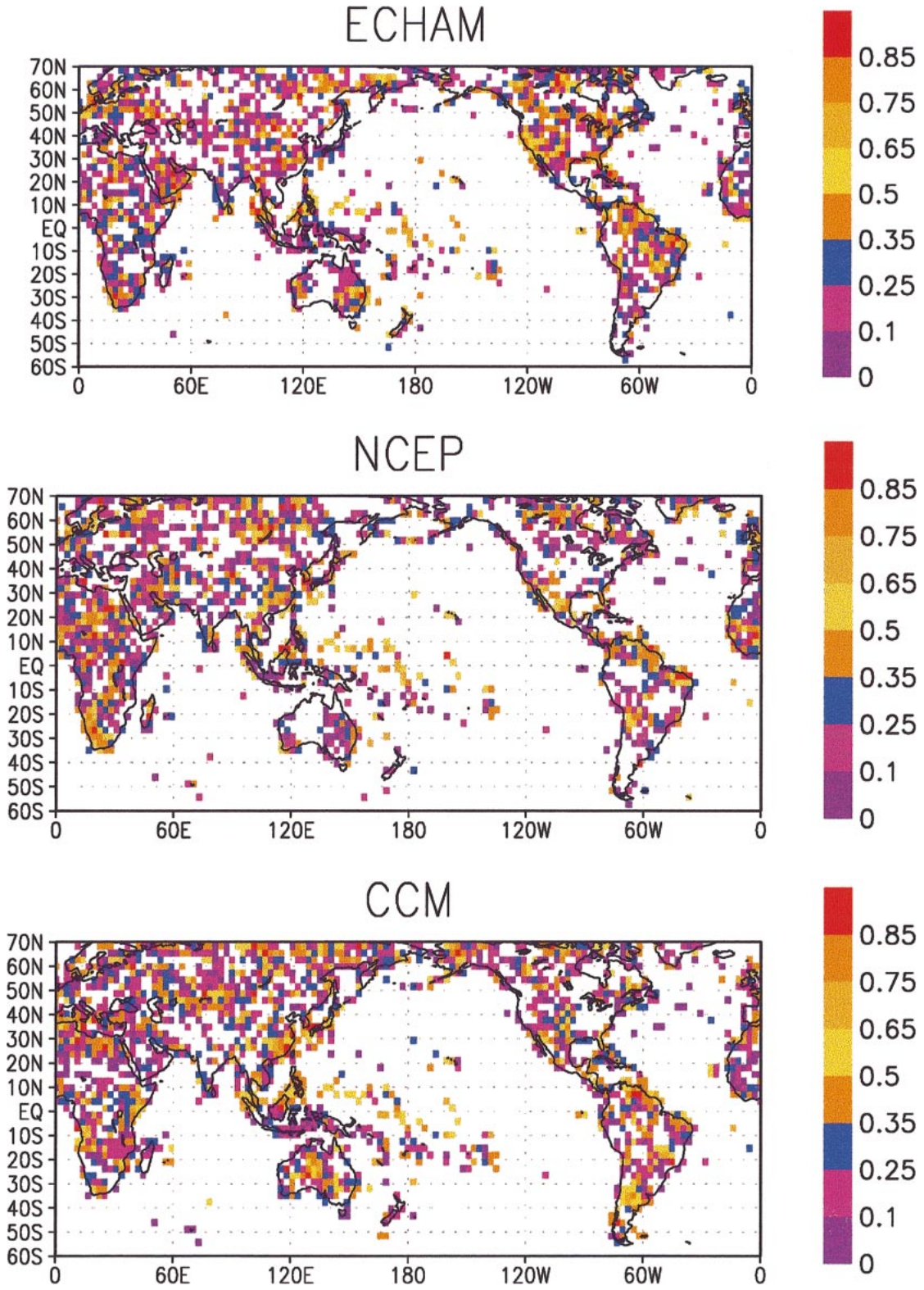

FIG. 2. Spatial map of model weights for precipitation for the season JFM, when models are regularized individually. The weights have been normalized so that the weight for the model and for climatology sums to 1. A normalized weight less than 0.2 indicates that each model ensemble member contributes less than 1 yr of climatology.

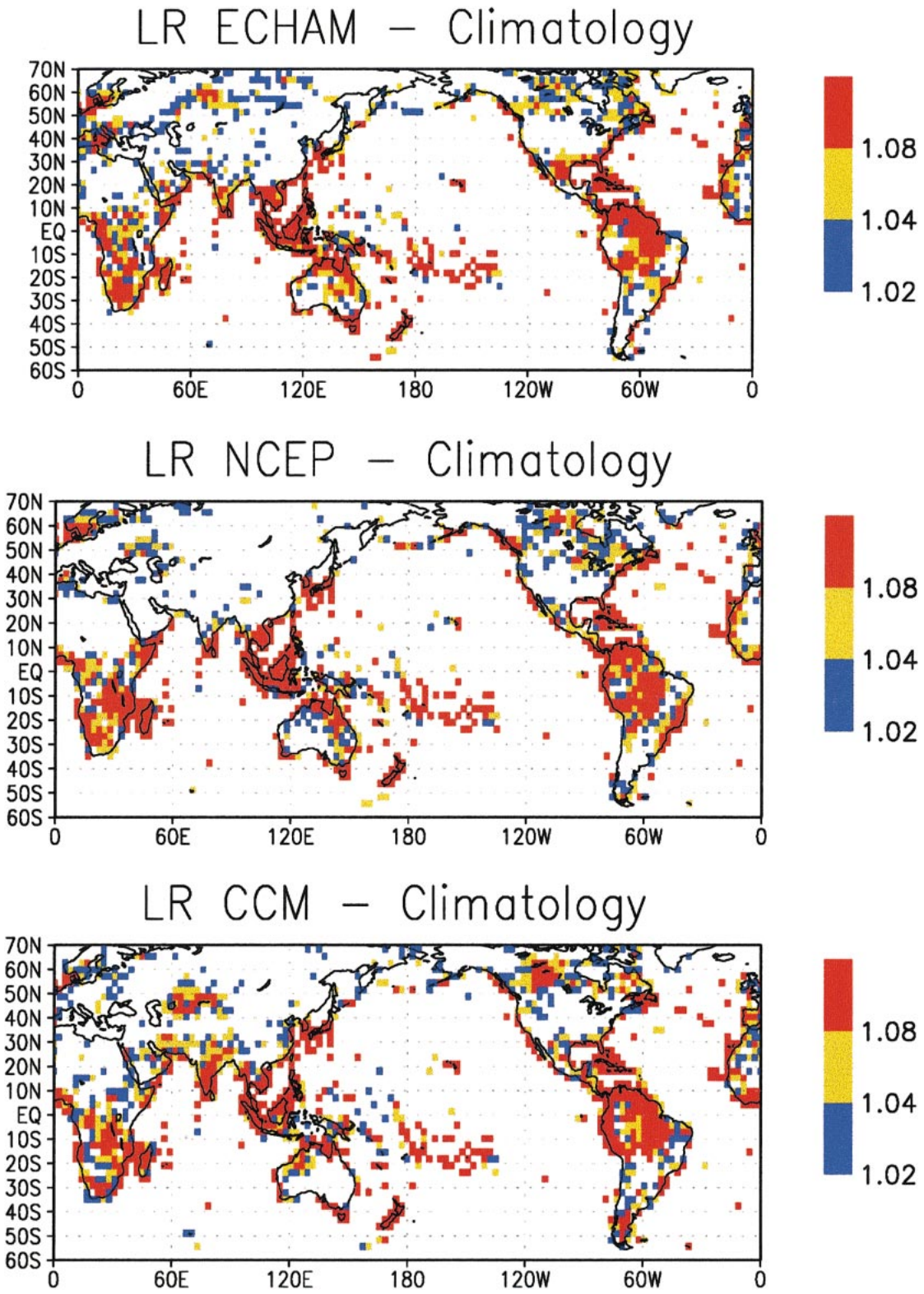

FIG. 3. Spatial map of LR values for temperature for the season AMJ, for the three models when considered individually. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test.

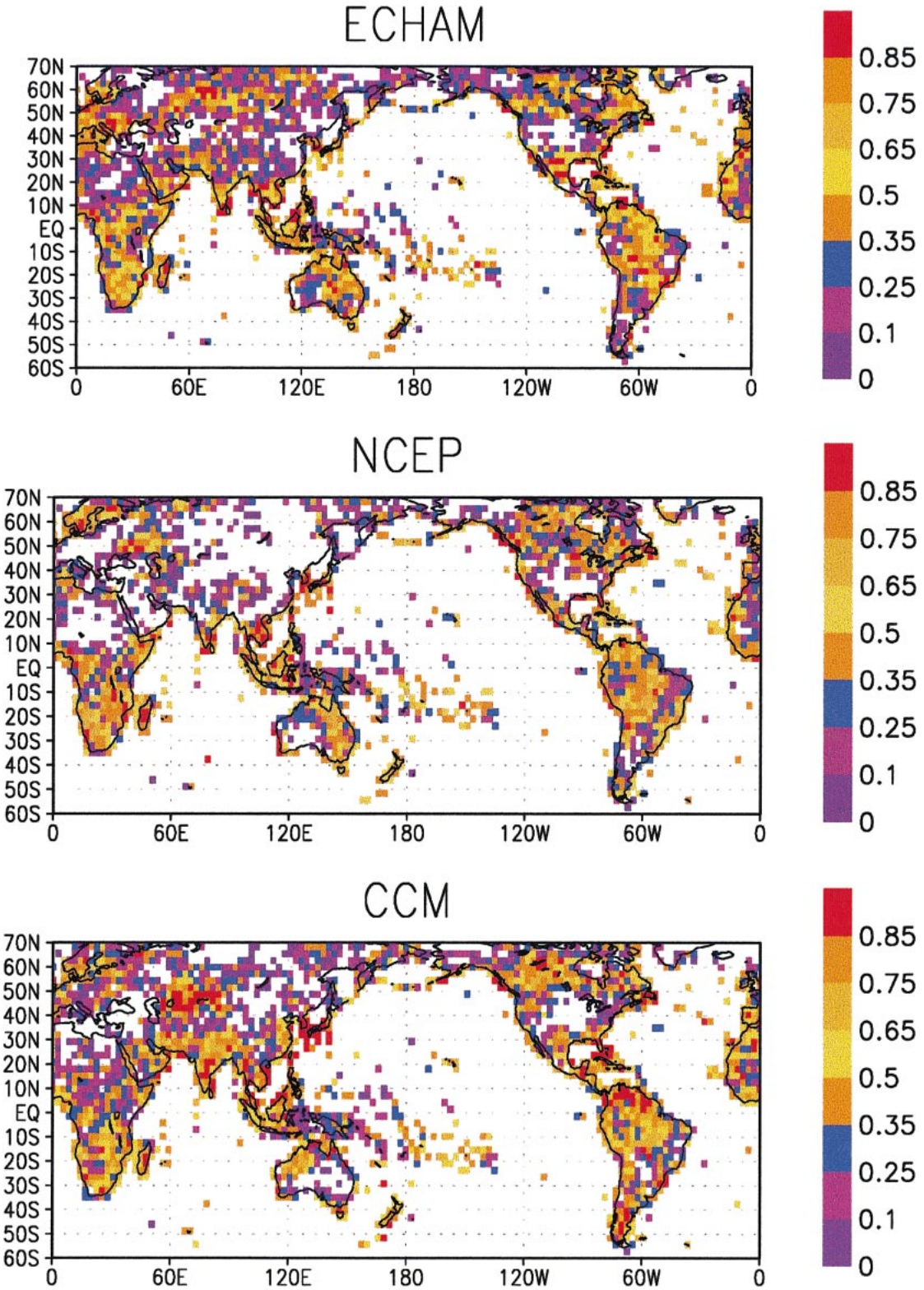

FIG. 4. Spatial map of normalized model weights for temperature for the season AMJ, when models are regularized individually. A normalized weight less than 0.2 indicates that each model ensemble member contributes less than 1 yr of climatology.

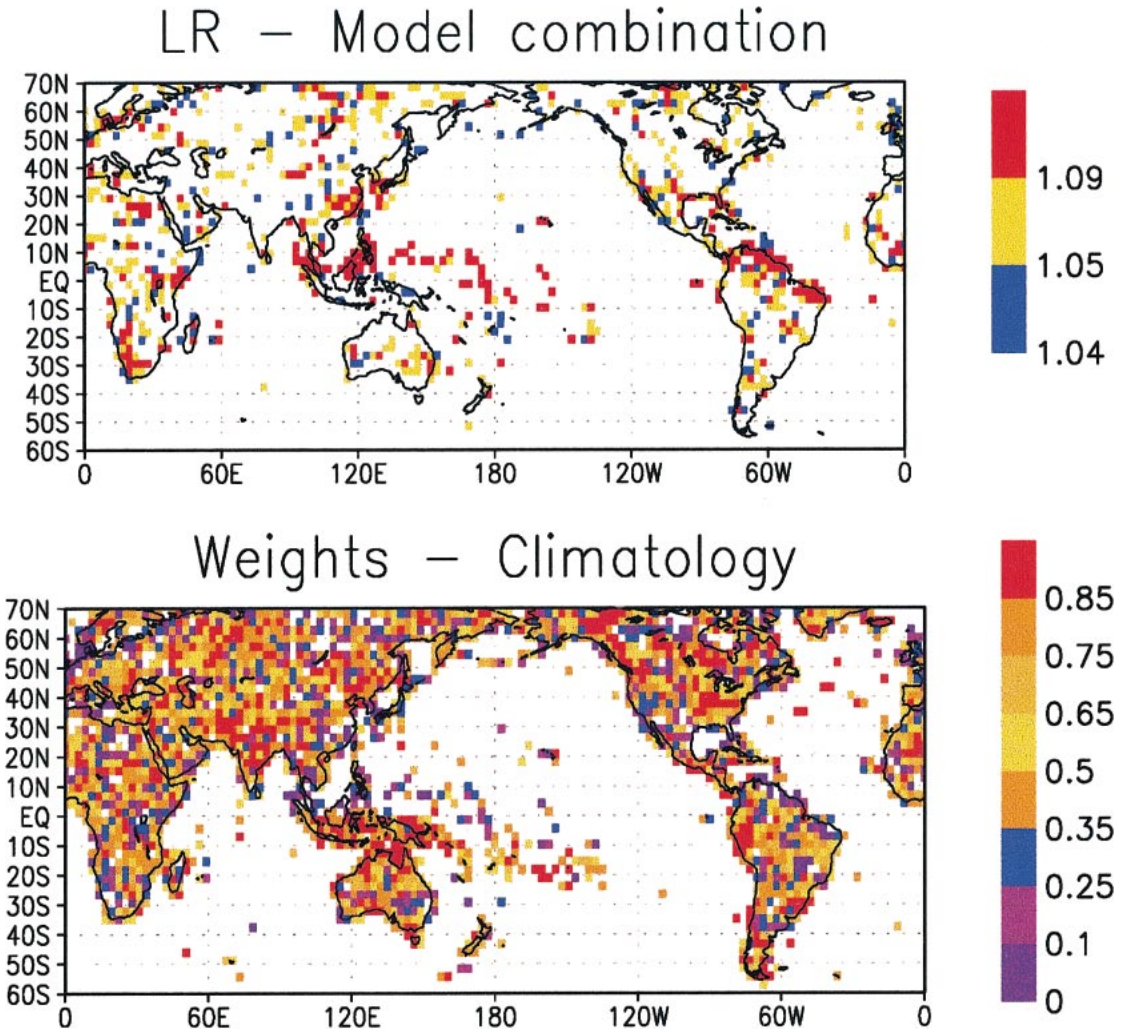

FIG. 5. (top) Spatial map of LR values for precipitation for the season JFM from model combination. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test. The LR values are higher than for the corresponding significance levels for a single model, reflecting the additional parameters used. (bottom) Spatial map of the weights for climatology in the model combination. A high weight (much greater than 0.25) for climatology reflects a situation in which the models are not informative.

b. Multimodel combination

Figure 5 shows the spatial map of LR, and the corresponding climatology weights (w_{j*}), for JFM precipitation based on the multimodel combination. Somewhat larger areas with significant LR values can be seen—indicating that by combining the models there is a general improvement overall, compared to any individual model (see Fig. 1). Note that the values of LR corresponding to the same level of significance are now higher than for the case of a single model. This is to be expected since a larger number of coefficients is now being estimated and hence a price must be paid for the reduced degrees of freedom. For areas where all the models had something to say, there is improvement in skill, reflecting the increased effective sample size of conditional information relative to the marginal information presented by climatology. However, the statistical significance levels have increased almost as much, reflecting in this case the variance reduction to be expected from correlated predictors.

The performance of a simpler, equal weights approach for combining the GCM forecasts, using a straight average of the categorical probabilities from the three GCMs, was also assessed. Results for the JFM precipitation forecasts are illustrated here. Maps of the LR for the equal weights combination relative to climatology and the LR for the optimal combination to the equal weights combination, respectively, are presented in Figs. 6a and 6b. Comparing Figs. 5a and 6a, there is seen an appreciable decrease in the regions with skill. This is to be expected since the skill of the individual models varies regionally. From Fig. 6b, it can be seen

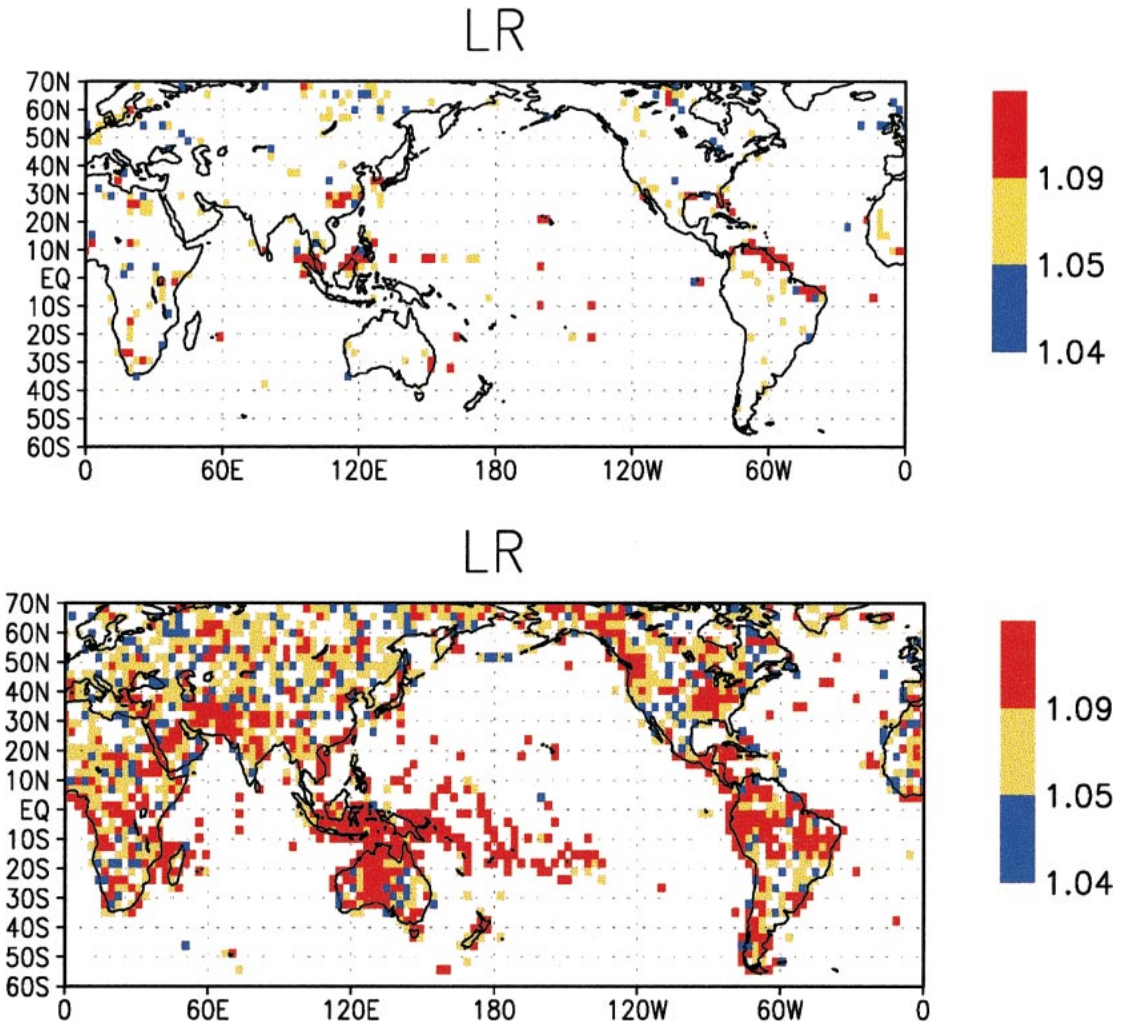

FIG. 6. (top) Spatial map of LR values for ''equal weights'' model combination relative to climatology, for JFM precipitation. (bottom) LR values for the optimal model combination relative to the equal weights combination, for JFM precipitation. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test.

that the likelihood ratio of the optimal model to the equal weights model exceeds the 90% significance level over most of the world. Indeed, the equal weights model underperforms climatology over a large area.

Figure 7 shows the spatial map of LR and the corresponding weights for climatology for AMJ temperature, based on the optimal multimodel combination. Compared to precipitation (Fig. 5), one can see a substantial increase in the size of contiguous regions with significant LR values across the globe, particularly in the Tropics, and along continental margins, as mentioned previously. Correspondingly, the weights for climatology in these regions are much smaller—indicating that the models have information content. It is well known that temperature exhibits a larger spatial correlation scale than precipitation. From this alone, the improved results with respect to temperature are not surprising. The results for other seasons follow a similar pattern.

The LR values at a given grid box from the model combination can be lower than the LR values from any one model—this is due to higher variability in the selection of the weights when solving for the additional parameters. Using the methods described here one can evaluate whether or not the results from a single model or the model combination are robust at a grid point or region of interest. Detailed comparisons of regional skill are also possible in terms of the LR and the *w* values. One could, for example, compute a spatial likelihood for the region for each year (as the product of the probability ascribed to the category with the ''hit'' at each grid box), and examine the time series of this skill measure to get some insights as to the conditions under which skill improves.

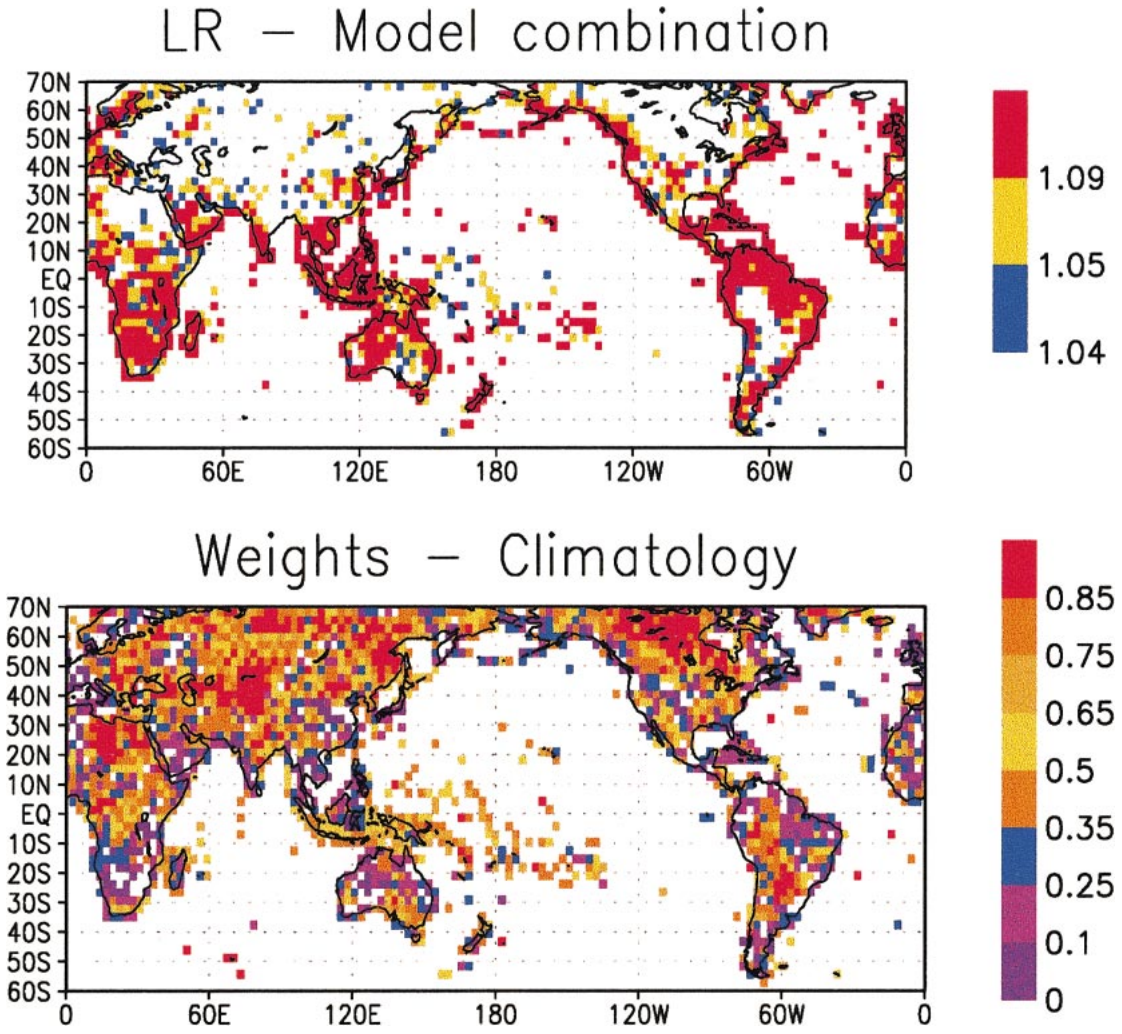

FIG. 7. (top) Spatial map of LR values for temperature for the season AMJ from model combination. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test. (bottom) Spatial map of the weights for climatology in the model combination. A high weight (much greater than 0.25) for climatology reflects a situation in which the models are not informative.

c. Sensitivity to objective function

The analysis presented was repeated using the widely used ranked probability skill score (RPSS; see, e.g., Kumar et al. 2001), as the objective to be maximized instead of the likelihood ratio. The RPSS is computed as follows. First, we obtain a ranked probability score for the model forecast:

$$\mathrm{RPS}_{\mathrm{mod}} = \sum_{t=1}^{N} \sum_{k=1}^{K} (Q_{kt} - O_{kt})^2 \qquad (11)$$

where Q_{kt} is the multimodel cumulative forecast probability for category k at time t, O_{kt} is the corresponding cumulative ''observed probability,'' where the category probability is taken to be 1 for the category that was observed to occur and 0 for the others. Next, we obtain a similar score for the climatology forecast:

$$\mathrm{RPS}_{\mathrm{clim}} = \sum_{t=1}^{N} \sum_{k=1}^{K} (C_{kt} - O_{kt})^2, \qquad (12)$$

where C_{kt} is the cumulative climatology probability for category k at time t (1/3 for a tercile category). Finally, the RPSS is obtained as

$$\mathrm{RPSS} = 1.0 - (\mathrm{RPS}_{\mathrm{mod}}/\mathrm{RPS}_{\mathrm{clim}}). \qquad (13)$$

We applied the model evaluation and combination procedures described in section 3, but using the RPSS [Eq. (13)] as the objective function as opposed to the log likelihood function [Eq. (8)]. The results using RPSS for JFM precipitation are presented in Fig. 8. The regions with significant RPSS (evaluated using a Monte Carlo procedure analogous to the one described earlier) match very closely to significant LR regions (Fig. 5). Figure 8b shows the climatology weights for the model

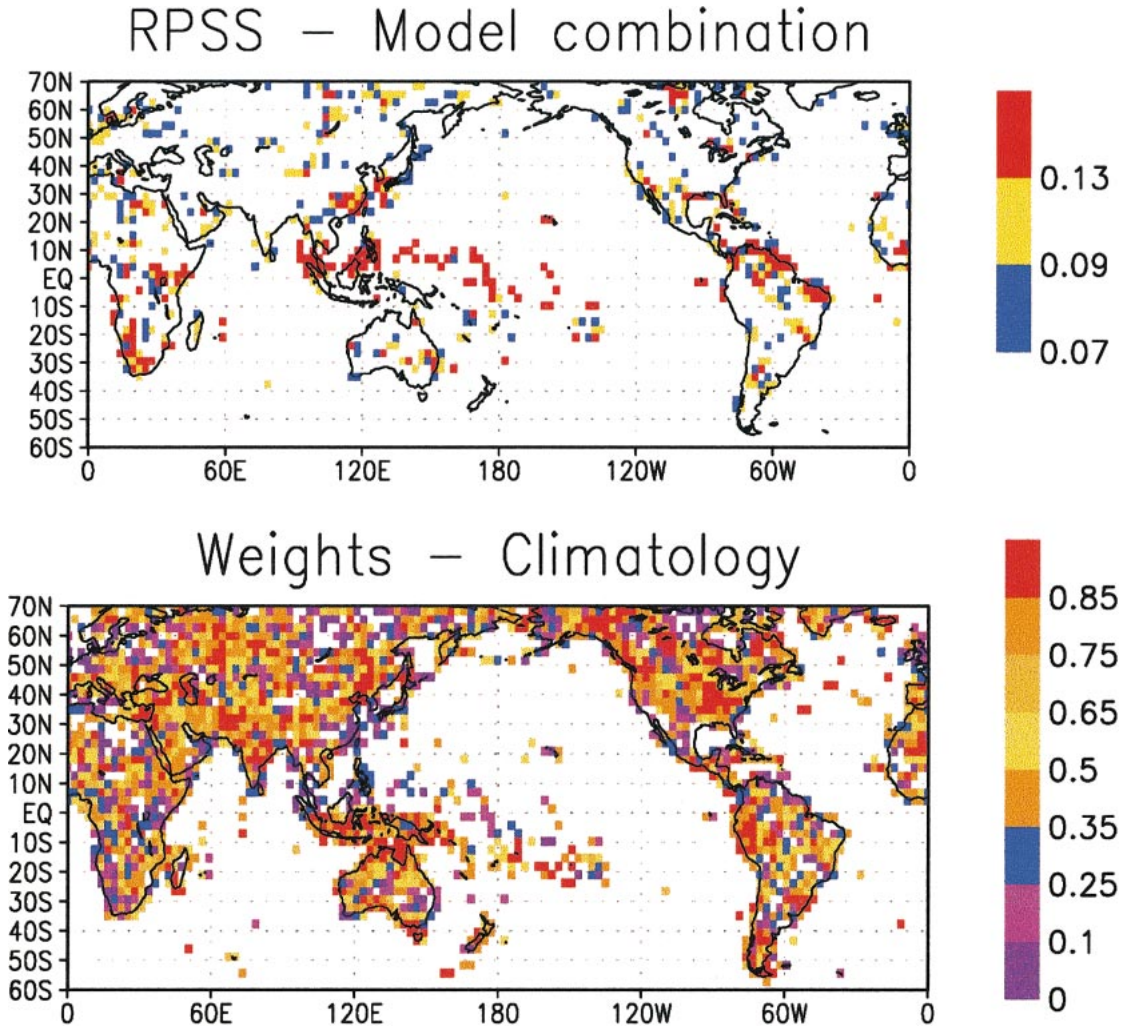

FIG. 8. (top) Spatial map of RPSS values for JFM precipitation, based on the model combination. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test. (bottom) Spatial map of the weights for climatology in the model combination.

combination that, again, are similar to those obtained from LR (Fig. 5b). Thus, at least in terms of these two objectives functions, the results are similar.

d. Forecast example

An example of a combined model forecast for the AMJ 2000 precipitation in Africa is provided in Fig. 9. The IRI net assessment precipitation forecast, which was derived subjectively using outputs and retrospective skill assessments for the same models, is shown in Fig. 10. It is broadly consistent with the forecast generated by the objective scheme shown in Fig. 9, but has somewhat less spatial structure, and somewhat more conservative category probabilities overall. This is consistent with the forecasters' intentions to reflect additional sources of uncertainty in the net assessment forecasts (e.g., the uncertainty in forecast SST, not accounted for in the simulation data analyzed here). The observed precipitation for the season (AMJ 2000) is shown in Fig. 11. The net assessment forecast called for increased probability of lesser precipitation over the greater Horn of Africa region and parts of the southwestern coast of Africa. It also called for a higher probability of wet conditions over southeastern and south-central parts of Africa, a small region on the west coast, and normal precipitation elsewhere (Fig. 10). The category probabilities obtained from the model combination broadly agree with the net assessment forecast over the greater Horn of Africa region and south-central Africa—that is, there is an increased probability of category 3 (higher precipitation) over southeastern and south-central Africa and a higher probability of category 1 (lower precipitation) over the greater Horn of Africa region (Fig. 9). Further, the model combination indicates a slightly higher probability of dry conditions over western Africa—

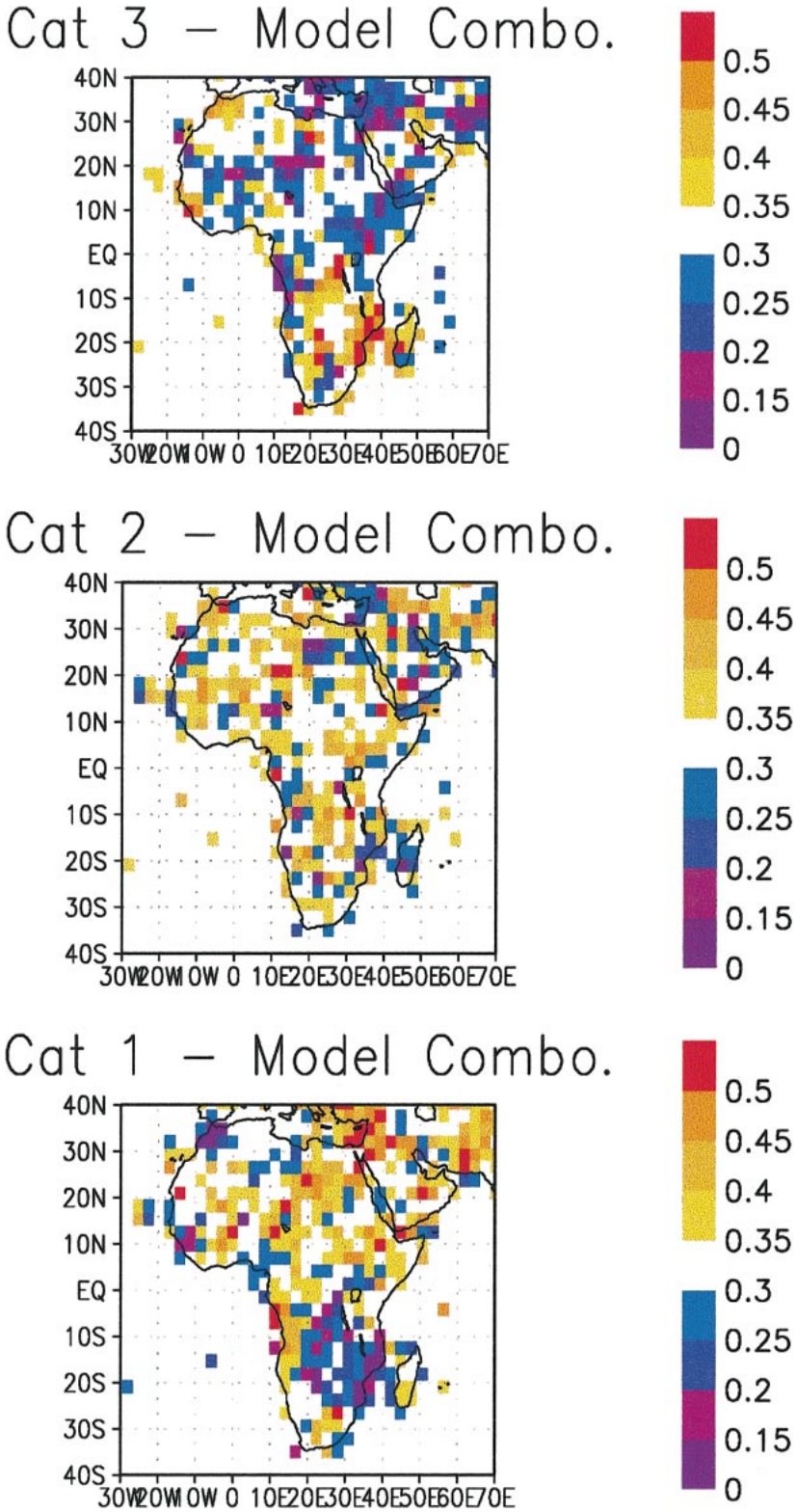

FIG. 9. AMJ 2000 forecast probabilities from model combination for (top) category 3 (upper tercile), (middle) category 2 (middle tercile), and (bottom) category 1 (lower tercile).

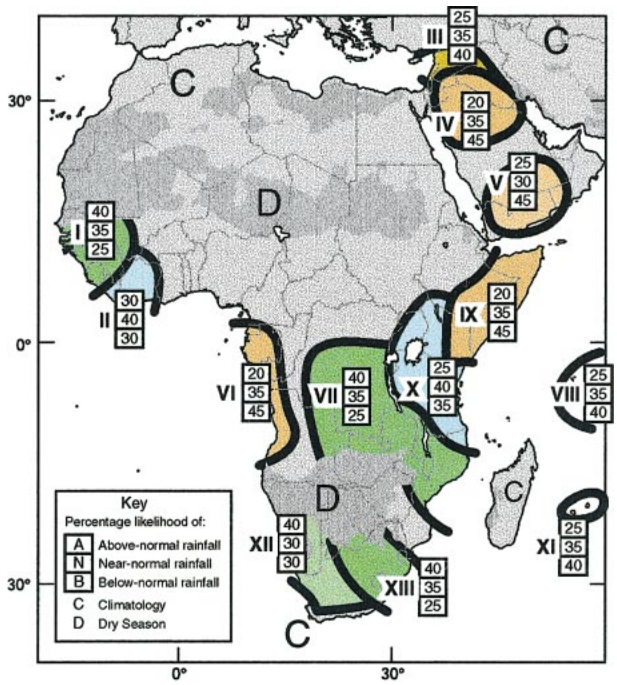

FIG. 10. IRI net assessment forecast for AMJ 2000 precipitation.

that is enhanced probability for category 1 (low precipitation) in Fig. 9c. The observed precipitation indicates increased precipitation over southeastern and parts of south-central Africa, and reduced precipitation over the greater Horn of Africa region, as called for by both the forecasts. Western Africa shows an overall normal to dry condition, consistent with the model combination forecast. For completeness, maps of historical AMJ precipitation forecast skill (likelihood ratio) and climatology weights for the model combination are presented in Fig. 12.

A definitive comparison of the performance of the objective forecast scheme and the net assessment forecasts cannot be made, given the small sample of the latter that are available (less than 20 at present). Qualitatively, we have found that the results are generally quite similar. This is encouraging—it suggests that objective methods can be competitive with current methods, while providing extra flexibility and an ability to address longer retrospective periods for validation.

6. Discussion

A new Bayesian methodology for assessing skill and combining forecasts from different models is presented.

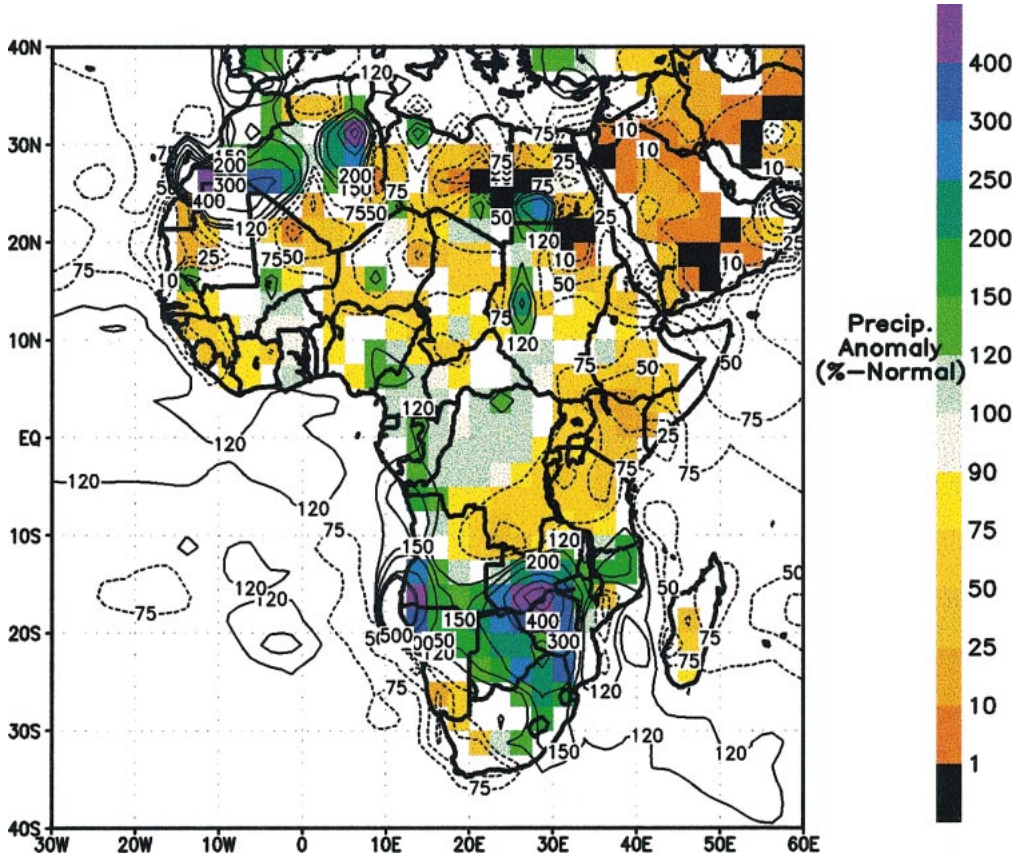

FIG. 11. Observed precipitation for AMJ 2000. Data is based on Climate Anomaly Monitoring System outgoing longwave radiation Precipitation Index (CAMS-OPI) analysis available at the NOAA Climate Prediction Center.

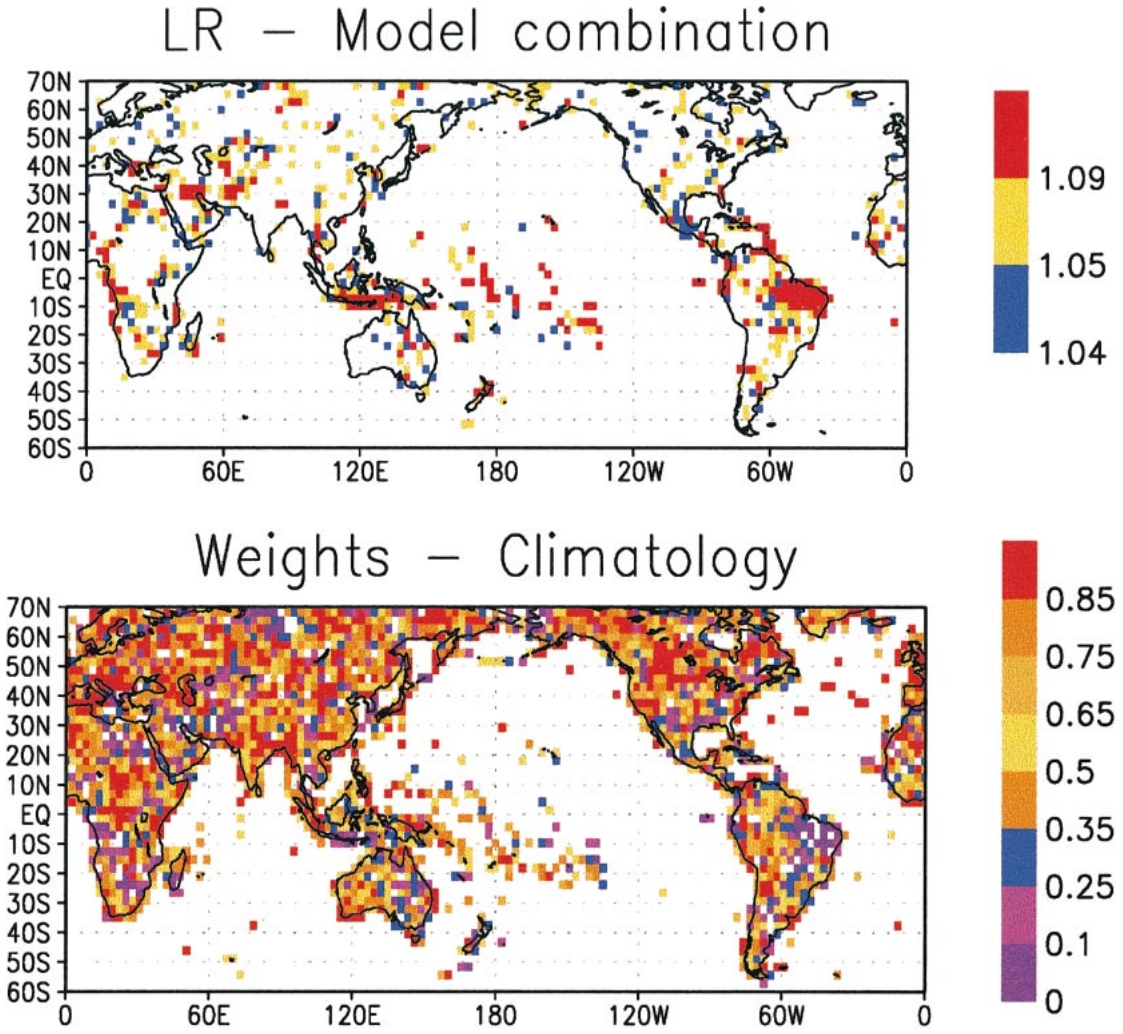

FIG. 12. (top) Spatial map of LR values for precipitation for the season AMJ from model combination. The three colors shown correspond to values that exceed the 90%, 95%, and 99% significance level determined through a Monte Carlo test. (bottom) Spatial map of the weights for climatology in the model combination.

The methods can be applied to ensemble forecasts from dynamical or statistical models. The categorical structure adopted allows a probabilistic forecast to be generated at a resolution (number of categories) desired by the user, and automatically considers the correction of bias in the marginal distribution of forecasts. In addition, it incorporates the effect of increasing uncertainty as the number of categories increases for fixed sample size or the baseline information on climatology at the site increases or decreases. An estimate of the uncertainty distribution of the estimated probabilistic forecast is also provided as a by-product of the algorithm. Selected objective criteria for evaluating model skill were also introduced. The least squares criteria used by classical regression-like schemes presume an underlying Gaussian error structure. Generalized linear models adapt to other distributions. However, these have not been explored in the particular context discussed here. Often, competing approaches use only the ensemble mean and hence lose some of the information that may be available. The nonparametric flavor of the categorical approach and the likelihood ratio is consequently attractive, particularly when dealing with variables such as precipitation whose marginal distribution may be highly skewed.

Unfortunately, the categorical structure of the model is also a limitation. The process of discretization induced by the categories does not allow one to properly consider the ordinal structure of the original data. Each of the categories is considered independent and the binary criteria (success if in category and failure if not) applied to judge performance is not cognizant of the distance between the forecast and the observation. In this sense, an approach based on the cumulative distribution function or on regression may do better since the full range of values of forecast and observation would be consid-

ered. This problem may also be ameliorated if one changed the manner in which the likelihood function is calculated. For instance, one could estimate the posterior likelihood as

$$L(w) = \prod_{t=1}^{N} \sum_{k=1}^{K} g_k(k^*) Q_{kt} \qquad (14)$$

where k^* is the category corresponding to the observation, $g_k(k^*)$ is a discrete kernel or weight function that sums to 1 and appropriately weights the category probabilities (e.g., the category k^* gets the highest weight, the adjacent categories get a smaller weight, and the weights decrease with distance from k^*). Examples of some kernel functions that are appropriate in this context can be found in Rajagopalan and Lall (1995). Whether the additional complexity of using an approach such as this is warranted needs to be investigated. There are many reasons to prefer an underlying continuous model to the discretization artificially imposed by the categories. We are exploring alternative formulations in the same Bayesian spirit to address this issue.

The algorithm presented here has been applied independently at each grid point. Information on the spatial structure of model forecasts or observations is consequently not used. Further, there are results suggesting that in many areas model skill is present only under certain conditions, for example, ENSO events. Thus, the temporal structure of the observation and forecast fields may also be of interest. The lack of consideration of spatial and temporal structure in these fields is a deficiency that also plagues competing algorithms (e.g., Krishnamurti 1999, 2000; Mason et al. 1999). A hierarchical modeling strategy (see Gelman et al. 1995), where the model parameters (e.g., w) are allowed to have prior distributions that embody spatial and temporal structure, may be one possible strategy for addressing this problem. Such a framework would also allow for a more realistic accounting of the uncertainty in model specification, and in properly computing the posterior odds considering parameter uncertainty. We are in the process of investigating such a strategy.

It is desirable to better understand the variability in model selection as a function of the sample size available for model fitting (N), the sample size for climatology (n), and the ensemble size (m). Numerical experiments to test the performance of the algorithm in this respect need to be conducted. We have deferred some of these experiments awaiting the development of a strategy for consideration of space–time structure. Actually, since we have access to the uncertainty distribution of the posterior probabilities as a by-product of the estimation process, it is possible to address this question immediately, in a limited way. A sophisticated decision maker who is interested in using the probabilistic forecast could generate an ensemble of categorical, probabilistic forecasts from the Dirichlet distribution specified in Eq. 6. These forecasts could then be used in a formal decision-making process (e.g., reservoir simulation) to derive optimal decision rules considering the uncertainty in the probabilistic forecast. In the hierarchical modeling context, one would have a prior distribution for w, which would be used to develop a posterior distribution for w using the posterior likelihood or other objective function. Then, instead of a point value for w, one would use this posterior distribution for w to develop a posterior distribution for the category probabilities. The uncertainty in selecting w from finite N, n and m, is naturally accounted for in this process.

The algorithm reported here is currently being used in the context of the real-time monthly climate forecasts that the IRI produces. It is hoped that in the near future this and related approaches will provide the opportunity to produce objective, probabilistic forecast products tailored to multiple applications. The objective component is crucial to the ability to produce hindcasts over time periods long enough to provide reliable assessments of performance in practical decision-making contexts, and to update these assessments with continually evolving models and forecast methodologies. The ease with which such combination algorithms can be adapted to alternative objective functions is seen as highly beneficial to producing specialized products of greater utility in specific problem areas such as agriculture, water resources, public health, fisheries, and disaster management.

Acknowledgments. We are grateful to Andrew Gelman, Lisa Goddard, Neil Ward, Tony Barnston and Vincent Fortin for their helpful suggestions in the course of the developments cited here. Financial support for this work was provided by International Research Institute for Climate Prediction and National Oceanic and Atmospheric Administration Grant NA67GP0299.

REFERENCES

Barnston, A. G., Y. He, and M. H. Glantz, 1999a: Predictive skill of statistical and dynamical climate models in forecasts of SST during the 1997–98 El Niño episode and the 1998 La Niña onset. *J. Climate,* **12,** 217–244.

——, A. Leetmaa, V. E. Kousky, R. E. Livezey, E. O'Lenic, H. Van den Dool, A. J. Wagner, and D. A. Unger, 1999b: NCEP forecasts of the El Niño of 1997–98 and its U.S. impacts. *Bull. Amer. Meteor. Soc.,* **80,** 1829–1852.

Cane, M. A., S. E. Zebiak, and S. C. Dolan, 1986: Experimental forecasts of El Nino. *Nature,* **321,** 827–832.

Epstein, E. S., 1969: A scoring system for probability forecasts of ranked categories. *J. Appl. Meteor.,* **8,** 985–987.

Fraedrich, K., and N. R. Smith, 1989: Combining predictive schemes in long-range forecasting, *J. Climate,* **2,** 291–294.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin, 1995: *Bayesian Data Analysis.* Chapman and Hall, 526 pp.

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

——, ——, ——, ——, ——, ——, ——, and ——, 2000: Multimodel ensemble forecasts for weather and seasonal climate. *J. Climate,* **13,** 4196–4216.

Kumar, A., A. G. Barnston, and M. P. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate,* **14,** 1671–1676.

Lall, U., and A. Sharma, 1996: A nearest neighbor bootstrap for resampling hydrologic time series. *Water Resour. Res.,* **32,** 679–693.

Luengberger, D. G., 1989: *Linear and Nonlinear Programming.* Addison-Wesley, 491 pp.

Mason, S. J., L. Goddard, N. E. Graham, E. Yulaeva, L. Sun, and P. A. Arkin, 1999: The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Amer. Meteor. Soc.,* **80,** 1853–1973.

O'Hagan, A., 1994: *Kendall's Advanced Theory of Statistics.* Vol. 2B, *Bayesian Inference,* Edward Arnold Press, 330 pp.

Palmer, T. N., C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multimodel ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2034.

Pavan, V., and F. J. Doblas-Reyes, 2000: Multi-model seasonal hindcasts over the Euro-Atlantic: Skill scores and dynamic features. *Climate Dyn.,* **16,** 611–625.

Rajagopalan, B., and U. Lall, 1995: A kernel estimator for discrete distributions. *J. Nonparametric Stat.,* **4,** 409–426.

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences.* International Geophysical Series, Vol. 59, Academic Press, 464 pp.

Zhou, J. L., and A. L. Tits, 1993: Nonmonotone line search for minimax problems. *J. Optimization Theory Appl.,* **76,** 455–476.