A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations

JEFFREY L. ANDERSON

GFDL/Princeton University, Princeton, New Jersey

(Manuscript received 23 March 1995, in final form 17 November 1995)

ABSTRACT

The binned probability ensemble (BPE) technique is presented as a method for producing forecasts of the probability distribution of a variable using an ensemble of numerical model integrations. The ensemble forecasts are used to partition the real line into a number of bins, each of which has an equal probability of containing the "true" forecast. The method is tested for both a simple low-order dynamical system and a general circulation model (GCM) forced with observed sea surface temperatures (an ensemble of Atmospheric Model Intercomparison Project integrations). The BPE method can also be used to calculate the probability that probabilistic ensemble forecasts are consistent with the verifying observations. The method is not sensitive to the fact that the characteristics of the forecast probability distribution may change drastically for different initial condition (or boundary condition) probability distributions. For example, the method is capable of evaluating whether the variance of a set of ensemble forecasts is consistent with the verifying observed variance. Applying the method to the ensemble of boundary-forced GCM integrations demonstrates that the GCM produces probabilistic forecasts with too little variability for upper-level dynamical fields. Operational weather prediction centers including the U. K. Meteorological Office, the European Centre for Medium-Range Forecasts, and the National Centers for Environmental Prediction have been applying this method, referred to by them as Talagrand diagrams, to the verification of operational ensemble predictions. The BPE method only evaluates the consistency of ensemble predictions and observations and should be used in conjunction with additional verification tools to provide a complete assessment of a set of probabilistic forecasts.

1. Introduction

Breaking with a long tradition of producing a single "deterministic" numerical forecast, operational prediction centers have recently begun to produce real-time ensemble numerical forecasts. While it now seems apparent that ensemble forecasts are one of the few currently tractable approaches to modeling the uncertainty inherent in predictions of the atmosphere–ocean system, a great number of questions about how best to produce, interpret, summarize, and evaluate ensemble forecasts remain.

Ensemble forecasts have been utilized in a variety of ways at operational centers. Perhaps the most fundamental application is to use the ensemble mean forecast as a substitute for a single traditional "discrete" forecast (Brankovic et al. 1990; Milton 1990; Tracton and Kalnay 1993). In general, ensemble average forecasts have reduced error compared to the mean error of the individual forecasts when evaluated with most traditional error metrics (Seidman 1981; Murphy 1988). As pointed out by Leith (1974), much but not all of this error reduction can be reproduced by statistical filtering techniques making use of previous forecast verifications.

Ensemble forecasts have also frequently been subjected to clustering algorithms (Brankovic et al. 1990; Ferranti et al. 1994), with the goal of producing a small, easily understood set of forecast states, usually characterized by the cluster means. Such methods attempt to make use of more information than is used by a single grand mean forecast, and there have been some promising results. Nevertheless, the proper definition of "cluster" in such ensemble results remains an unresolved question, and algorithms used for forming clusters contain a number of heuristic parameters that can have some impact on the resulting clusters. In addition, it continues to be difficult to establish whether clusters in this sense should really be expected to exist in the forecast probability distributions arising from current generation GCMs and observational error distributions (Trevisan 1995; Wallace et al. 1991; Cheng and Wallace 1993). Many researchers are continuing to address such issues in both simple models and GCMs.

A third application of ensemble forecasts has been to make a priori predictions of forecast skill (Murphy 1990; Milton 1990). The most straightforward of these

Corresponding author address: Dr. Jeffrey L. Anderson, Geophysical Fluid Dynamics Laboratory, Princeton University, P.O. Box 308, Princeton, NJ 08542.
E-mail: jla@gfdl.gov

methods postulates a relation between the spread of a forecast ensemble distribution and the skill of a forecast (Hoffman and Kalnay 1983; Mureau et al. 1993), usually the ensemble mean, derived from the ensemble. Even in relatively simple models, the prediction of skill has proved to be less straightforward than one might hope (Barker 1991), although there continue to be suggestions that such predictions can be made in operational models (Tracton et al. 1989; Barkmeijer et al. 1993; Brankovic et al. 1994). As pointed out by Anderson and Stern (1996), it is important that the measure of ensemble spread and the measure of forecast skill be appropriately related. Even when such efforts are taken, problems such as the interaction of systematic model error with spread and skill are not yet completely understood.

A final paradigm for utilizing ensemble forecasts is to examine the entire ensemble. Obviously, the most information can be extracted in this fashion, but the information may be in a form that is too complex to be readily utilized. Ensemble forecasts from a large variety of models have been examined in this fashion (see, e.g., Tracton and Kalnay 1993). Ensemble integrations with prescribed external forcing have also been evaluated in a similar fashion (Hoerling et al. 1992).

In what follows, a method for utilizing ensemble forecasts that makes direct use of all ensemble members is described. The goal of this method is to use as much information as possible from the ensemble by approximating the forecast probability distribution for some variable. This method produces probabilistic forecasts of variables and also leads to a method for verifying the consistency of such forecasts with observations. The consistency verification method is nonparametric and allows verification of large sets of ensemble forecasts. Unlike many tools that have been applied to verify ensemble forecasts to date, this method evaluates the entire probability distribution forecast, not just the mean or some set of low-order moments (Deque et al. 1994). The method can be applied to forecasts whose verifications have radically different probability distributions. The verification method can be used to study the systematic error characteristics of the entire forecast probability distribution. Similar methods have been developed independently by groups at several operational centers and are now being used operationally to verify the consistency of ensemble forecasts (Harrison et al. 1995).

It is assumed in the following sections that the state of the atmosphere used as initial conditions for numerical model forecasts can never be measured exactly because of observational limitations. Instead, the initial conditions can only be properly represented in terms of a probability distribution that is determined by the observed state and the distribution of observational errors; the significant difficulties in finding this observational error distribution are ignored here. A forecast consists of integrating the initial condition probability distribution with a numerical model and determining the resulting probability distribution at some later time. In practice, it is too expensive to integrate the initial probability distribution forward in time (Epstein 1969a), so traditional Monte Carlo ensemble forecasts must be used to attempt to produce an equitable sample of the forecast probability distribution. It should be noted that the methods currently in use by many operational centers do not explicitly try to create an equitable sample of the initial condition probability distribution (Tracton and Kalnay 1993; Harrison et al. 1995).

It is also useful to define a "true" forecast probability distribution. This is the probability distribution that would be obtained if the initial condition probability distribution could be integrated forward to the forecast time with a perfect model. In other words, it is the best forecast that can be made given that observational errors are inevitable. The actual true verifying state of the atmosphere (which can itself never be known exactly due to observational error) can be thought of as a random sample from the "true" forecast probability distribution.

Section 2 presents some basic results on sampling of random variables; the relation between ensemble forecasts and such random variables is then developed. Sections 3 and 4 present sample applications of the method. Section 3 examines the method in the context of a simple dynamical model, while section 4 uses a sophisticated atmospheric general circulation model forced by observed sea surface temperature (SST) distributions. Section 5 presents discussion and ideas for additional applications.

2. Binned probability ensemble forecasts

This section examines the use of binned probability ensemble (BPE) forecasts. The first subsection discusses the theory of BPE forecasts, while the second presents a simple idealized example of a forecast. The third subsection eventually relaxes the perfect model context that is assumed in the previous subsections in order to demonstrate use of the BPE method for validating ensemble forecasts and for investigating systematic errors in models and initial condition distributions.

a. Theory

Let X be a random variable and let the set x_i, $(i = 0, \cdots, n-1)$ be samples of this random variable. The x_i can be sorted by value, and there is a $1/n$ chance that sample x_0 is the smallest, a $1/n$ chance that it is the second smallest, etc. The remaining $n-1$ samples x_i $(i = 1, \cdots, n-1)$ partition the real line into n intervals, called bins hereafter. It follows that the probability that the sample x_0 falls into any given bin is $1/n$.

Now let $\mathbf{U}(t) = (u_1(t), u_2(t), \cdots u_m(t))$ be an m-dimensional vector that represents the state of a forecast model in phase space (Gleeson 1970). Because of un-

certainties in the initial condition of the forecast model integration, the state of a perfect forecast model at any time is represented as an m-dimensional random variable that is a function of time, $V(t)$, and the ith individual sample from V at a given time is represented by the state vector ${}_i\mathbf{U}(t)$. Suppose that there are n samples from $V(t)$, one of which is the truth, ${}_0\mathbf{U}(t)$, and an $(n - 1)$-member ensemble forecast, ${}_i\mathbf{U}(t)$, $i = 1, \cdots, n - 1$. There exists a random variable, $X_j(t)$, associated with the jth vector component of $V(t)$ (X_j is the marginal distribution of V for the jth vector component). Let ${}_iu_j$ represent the ith sample of X_j. The $(n - 1)$ forecasts of the jth component (${}_iu_j$, $i = 1, \cdots n - 1$) can be sorted to partition the real line into n bins, and as in the previous paragraph, the probability that the truth ${}_0u_j$ lies in any given bin is $1/n$.

When making an ensemble forecast, an explicit representation of the forecast probability distribution corresponding to V is not available for any but the initial time, so V cannot be sampled directly. Instead, when making ensemble forecasts (in some idealized world), one is given a single observation of the initial state, $\mathbf{U}_{\text{obs}}$, a probability distribution for the observational error corresponding to an m-dimensional random variable E, and a numerical forecast model (assumed to be a perfect model for the time being) that maps a state at the initial time into a forecast state at time t_f. The forecast model can be represented by a function, g, such that $\mathbf{U}(t = t_f) = g[\mathbf{U}(t = 0)]$.

Following Leith (1974), a probability distribution for the true initial state can be created by adding the observational error distribution to the observed point. Let the random variable $Y = V(t = 0) = E + \mathbf{U}_{\text{obs}}$, be associated with this initial condition distribution. Next, again assuming a perfect model, $V(t = t_f) = g(Y)$ is a random variable associated with the distribution of the forecast state, $\mathbf{U}(t = t_f)$. When making an $(n - 1)$-member ensemble forecast, $n - 1$ initial condition states are sampled from Y using the observed point and the observational error distribution. Each of these initial states is then integrated by the forecast model to produce an $n - 1$ member sample of $V(t = t_f)$. A binned probability forecast can then be made for any component of the state vector at the forecast time (or for any function of the state variables) following the example of the previous paragraph. The $n - 1$ ensemble values of the jth component of the state vector are sorted to divide the real line into n intervals; again, the probability that the true value of the jth component of the state falls into any given bin is $1/n$.

b. *Binned forecast example*

As a simple example, suppose a three-member ensemble forecast is made using an atmospheric forecast model. Three possible initial conditions for the ensemble are selected by randomly sampling from the initial condition distribution, and each is then integrated with the numerical model. Suppose that the forecast values of temperature at a given model grid point are 22, 23, and 26. Then the probability that the true temperature is less than 22 is one-quarter, the probability that it is between 22 and 23 is one-quarter, the probability that it is between 23 and 26 is one-quarter, and the probability that it is greater than 26 is also one-quarter. Without making use of additional information about the forecast model or the initial condition distribution, this is the most information that can be extracted about this temperature from the ensemble forecast. Providing probabilistic forecasts in this form is particularly useful since any user of the forecasts can extract whatever information they would like.

To avoid confusion in the following discussion, it is essential that a precise terminology be defined. Throughout the following sections, the term "member" is used to refer to an individual discrete forecast that is part of an ensemble of forecasts selected from the same initial condition probability distribution. The term "set" will be used to refer to a collection of n-member ensemble of forecasts. A set is composed of individual "samples," each of which is an n-member ensemble of forecasts. Each sample is assumed to be for a unique initial condition probability distribution.

c. *Model validation with binned forecasts*

Model validation can also be performed using BPE forecasts. Given an m-sample set of $(n - 1)$-member perfect model ensemble forecasts and the corresponding verifying true state vectors, the m samples of the bin number containing the true state should be uniformly distributed in the n bins. Since the BPE method is nonparametric, this result does not depend on any of the details of the probability distribution of the forecasts or the initial conditions, so a large set of ensemble forecasts can be grouped for validation without difficulty. The traditional chi-square test can be directly applied to the BPE results for a set of ensemble forecasts to see how likely it is that the bin distribution is uniform.

However, the whole premise of ensemble forecasts is based on the notion that the true state of the system being forecast cannot be exactly measured; instead, all observed states include the effects of the observational error, E. As pointed out above, the "true" state of the system at the verification time should fall into each of the n bins with equal probability (the perfect model assumption is still being made). However, this is not the case for the observed state at the forecast time. The observed state is sampled from a random variable $Z = V + E$. In order to validate ensemble forecasts using the BPE technique, each of the ensemble forecasts at the verification time is regarded as an independent estimate of the "true" state of the system. A random sample from the observational error distribution, E, must be added to each member of the ensemble forecast

so that the members will be independent estimates of the "observed" state at the forecast time. Once this has been done, the resulting $n - 1$ forecasts of the random variable Z can be used to form bins. This time, the *observed* state at the forecast time is equally likely to fall into any one of the n bins.

Now suppose that the perfect model assumption is no longer necessarily valid and that the specification of the error distribution E is also possibly incorrect. Validations for a large set of ensembles can be performed as outlined above. For each ensemble forecast in the set being verified, the ensemble members are used to form n bins. The number of the bin into which the corresponding observation falls is then found. The chi-square test can be used to determine if the observations are uniformly distributed in the bins and the significance of the chi-square test can be evaluated. A small value of chi-square significance means that the distribution of observations in the bins is significantly different from uniform so that either the forecast model or the observational distribution must be inconsistent with the truth.

In cases where the distribution of the observations in the bins is not uniform, an examination of the frequency with which the observation falls into each of the bins can help to analyze the structure of the model systematic error distribution. This is related to the traditional examination of model bias, however, in this case the chi-square test gives an a priori assessment of the significance of the model systematic error. Bias (in terms of the median, not the mean) in the forecasts can be seen if many of the verifications fall well to one side of the middle bin. Problems with variance of the forecast distributions can be seen if the verifications are clustered in a few neighboring bins, or if many of the verifications are located in bins far away from the center. It is important to recall that the chi-square test is measuring the statistical significance of the difference between the distributions, not the strength of the difference (see section 3). Given ensembles with many members and a sufficiently large sample of ensemble forecasts, problems with even higher order moments of the forecast distribution can be diagnosed.

The next two sections provide demonstrations of the BPE method for evaluating ensemble forecasts. The first section examines an initial value problem in a highly idealized model, while the second examines results for a modern general circulation model integrated for a long period in the presence of observed sea surface temperature forcing.

3. Binned forecasts in the Lorenz model

In this section, the three-variable Lorenz convective model (Lorenz 1963) is used to provide demonstrations of the BPE technique. This model is used because its small size allows very large samples of ensembles with many members to be generated. Although it is obviously far-fetched to assume that such large ensembles will ever be available for realistic atmospheric models, it is useful to use the simple model to understand the behavior of the BPE technique.

The Lorenz model is represented by the equations:

$$\dot{x} = -\sigma x + \sigma y \tag{1}$$

$$\dot{y} = -xz + rx - y \tag{2}$$

$$\dot{z} = xy - bz, \tag{3}$$

where the dot represents a derivative with respect to time. The parameters are set to $\sigma = 10$, $r = 28$, and $b = 8/3$, identical to those used in Lorenz (1963) with a nondimensional time step of 0.01 unless otherwise noted. With these parameters, the model displays chaotic behavior. A probability distribution for the initial conditions in this model is stretched and twisted in phase space as integration time increases. Since the system is ergodic, after a sufficiently long integration time the forecast probability distribution becomes identical to the distribution created by periodically sampling a single very long integration of the model.

A large number of true states, X_t, from the Lorenz model attractor are generated by integrating the model for a very long time from arbitrary initial conditions, discarding the first 100 000 time steps, and then sampling the rest of the integration every 1000 steps. For most of the results reported here, a set of 1000 initial states is used. An observational error distribution corresponding to the random variable E of the previous section is specified as multinormal with a standard deviation of 1.0 (about 2% of the range of these variables on the attractor) for each of x, y, and z. A set of 1000 "observed" states is generated by adding an independent random sample from E to each of the 1000 true states.

An ensemble of initial condition members are generated for each of the 1000 observed states by adding random samples from E to the observed states (Leith 1974). These ensemble initial conditions correspond to samples from the random variable Y defined in section 2. All experiments here use an ensemble size of nine unless otherwise noted.

For each of the 1000 true states, the true state and the nine ensemble initial conditions are integrated in the Lorenz model. At a given lead time and for each of the 1000 samples, the nine ensemble values of each of the Lorenz variables (x, y, and z) are sorted to form equiprobable bins. The bin number into which the true state falls is then determined. This results in 1000 true-state bin values at each forecast lead time; one can then test if the distribution of these 1000 samples is uniform in the ten bins. Figure 1 plots the significance of the chi-square test applied to the distribution of the true state in the bins for lead times out to 200 steps for the z variable. The chi-square significance can be interpreted as the probability that the distribution of the true

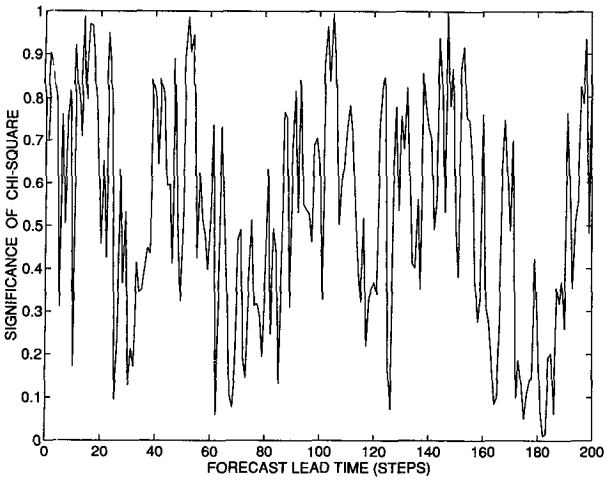

FIG. 1. Significance of chi-square test as a function of forecast lead time for z in the Lorenz model for perfect model ensembles, 1000 samples, nine-member ensembles.

values in the bins was selected from a uniform distribution. Small values indicate that the binned distribution is probably not uniform. For the results of Fig. 1, the chi-square tests support the conclusion that the real state falls into each of the bins with equal probability. This demonstrates that the BPE method provides consistent probabilistic forecasts. The large variability of the significance of the chi-square test with time is simply a reflection of the fact that for the null hypothesis of identical distributions, the chi-square significance values should have a uniform distribution on [0, 1].

In any more realistic situation, only noisy observations of the system would be available and the true state of the system could never be known exactly. From now on, it is assumed that only the observed values from the Lorenz model integration are available. As pointed out in section 2, in order to validate the ensemble forecasts in this case, it is necessary to add a random sample from the observational error distribution, E, to each ensemble forecast element before forming the bins into which the verifying observed value is placed. Results for this experiment are qualitatively similar to those of Fig. 1, again providing heuristic verification of the BPE technique as developed in section 2.

The ability of the BPE method to detect errors in the forecast model is investigated in a third experiment with the Lorenz model. The observed states and the ensemble initial conditions are selected as in the previous discussion. This time, however, the ensemble forecasts are integrated using a value of b in (3) that is 2% larger than in the control integration. In this case, the ensembles should become a progressively worse estimate of the true state as forecast lead time is increased. Figure 2 shows chi-square significance for the BPE technique applied to the observed values of z for this case. As lead time increases, the significance decreases to values close to zero, indicating that the BPE forecast is no longer a good estimate of the true state (i.e., the ensemble forecasts with added samples from the observational error are not good estimates of the observed state). At very long lead times (not shown in the figure), the chi-square values rebound to indicate that the BPE method provides a good probabilistic forecast. This is simply a result of the ergodicity of the Lorenz system and the fact that the attractors for the Lorenz system with the standard value of b and the 2% increased value are statistically extremely similar.

The BPE technique can also detect errors resulting from an improper estimate of the observational error when selecting ensemble initial conditions and validating the binned probability forecasts. An experiment to demonstrate this is constructed by using the same observed states as in the previous experiments. However, in this case, the error distribution used in forming the ensemble initial conditions and in modifying the ensemble forecasts before forming the bins is assumed to be multinormal with a standard deviation of only 0.90 (as opposed to 1.0, which is used to generate observed states from the true states). Figure 3 plots the significance of the chi-square test for z in this experiment. For small lead times, the chi-square significance is generally small, indicating that the BPE forecasts are not consistent with the observed states. However, after about 100 steps, the chi-square values begin to increase, indicating that the binned forecasts are becoming progressively better. This is a result of the symmetric nature of the error introduced in the observational error distribution. As the forecast lead time is increased, both the correct and the erroneous (i.e., the 1.0 and 0.90 standard deviation observational error, respectively) probability distributions in phase space are stretched and twisted as shown in Anderson (1996).

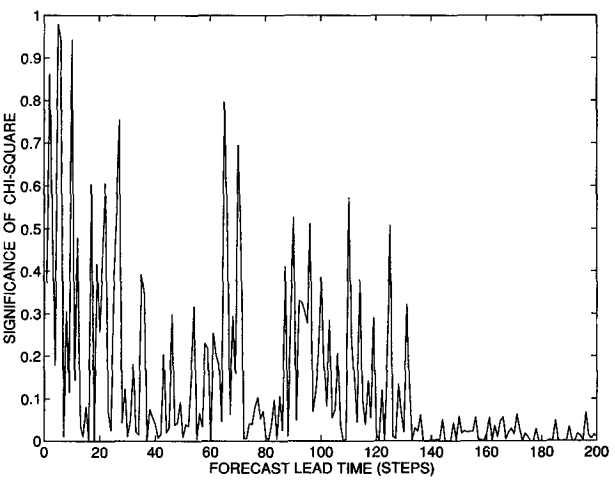

FIG. 2. Significance of chi-square test as a function of forecast lead time for z in Lorenz model with imperfect ensemble forecasts having a 2% larger value of b, 1000 samples, nine-member ensembles.

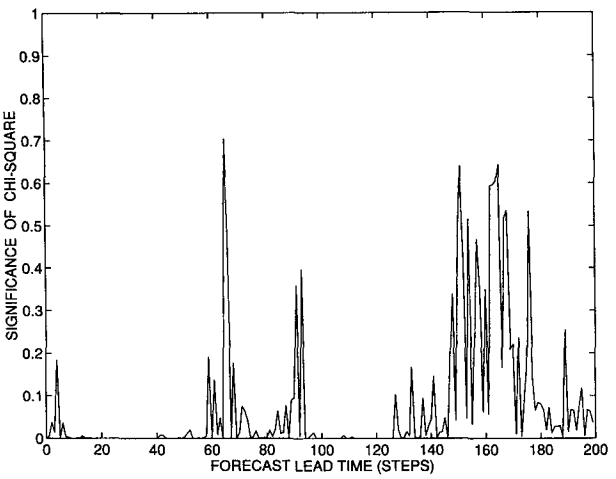

FIG. 3. Significance of chi-square test as a function of forecast lead time for z in Lorenz model with imperfect initial condition error distribution having standard deviation 0.9 times true value, 1000 samples, nine-member ensembles.

As the distributions become increasingly deformed, the probability density in the wings of the distributions becomes less, and it becomes more and more difficult to detect statistically significant differences in this part of the distribution.

This simple model context is convenient for a brief examination of the sensitivity of the results to the number of members in the ensemble and to the number of samples in the set (see discussion of terminology in section 2b) used to evaluate the BPE forecasts. As the number of ensemble members increases, the ability of a perfect model to resolve details of the forecast probability structure increases. However, if the BPE forecasts are being used to validate forecasts from an imperfect model, the results of the chi-square significance are not particularly sensitive to the number of bins (ensemble members) in this simple model. In this case the number of bins is not particularly relevant to determining if a forecast differs in some way from the verification because the mean of the two distributions is different. Nevertheless, if the details of the systematic error structure of the model are of interest or if differences only exist in higher-order moments of the distributions, using more than just a few bins may reveal additional information about the systematic error. In addition, more complex models might introduce more complicated structure in their forecast probability distributions. It might require relatively large numbers of bins to detect the details of such probability distributions.

Not surprisingly, the results of the chi-square test are sensitive to the number of samples in the set used to evaluate ensemble forecasts (i.e., the number of different observations for which ensemble forecasts have been made). As the sample size increases, the information yielded by the ensemble forecasts increases; large samples can reveal differences that cannot be distinguished from noise by small samples. Figure 4 shows the chi-square significance for the same conditions as in Fig. 2 (the flawed model example), except that the statistic has been evaluated for a smaller 100 sample and a larger 5000 sample set. The 5000 sample case shows that the significance of the chi-square becomes very small after just a few steps (the solid curve for the 5000 sample case is only visibly nonzero for the first few steps), while the 100 sample case is never able to distinguish between forecast and verification with statistical significance. Although the chi-square test is only a measure of the significance of differences between the forecasts and verifications and not the strength of the difference, indirect information about the strength of the difference can be gleaned from the sample size needed to find significant differences.

Only results for z have been shown in this section, since results for x and y lead to qualitatively similar conclusions. In the perturbed model and initial condition experiments, z is somewhat more sensitive to differences between the ensemble and the verification than are x and y.

4. AMIP ensemble integration

In this section, the BPE technique is applied to an ensemble of integrations in an atmospheric general circulation model (GCM) that is forced by observed sea surface temperatures (SSTs). The atmospheric model is an 18-level spectral model truncated at T42 and is described in Gordon and Stern (1974, 1982). The prescribed SSTs are those used for the Atmospheric Model Intercomparison Project (AMIP) (Gates 1992). A

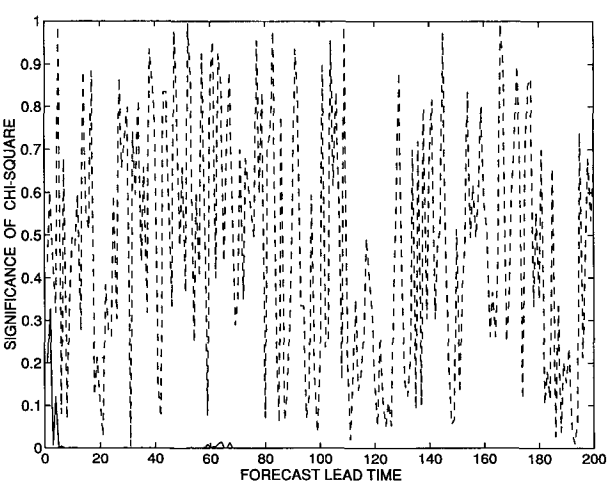

FIG. 4. As in Fig. 2 except for a 5000-element sample (solid line, only visible in lower left corner) and a 100-element sample (dashed line).

nine-member ensemble was integrated for 10 years from 1 January 1979 through 31 December 1988. The initial conditions for the ensembles were taken from analyses for 12 December 1978 through 21 January 1979, sampled every five days. Each of these analyses was then used as an initial condition as if it were the analysis for 1 January 1979. A more detailed description of the ensemble integrations can be found in Stern and Miyakoda (1995). Although this is not a true forecast experiment because of the imposed SSTs, the results will be referred to as "forecasts" throughout this section.

The first 14 months of the ensemble integrations were discarded in an attempt to eliminate direct effects of the initial conditions. The remainder of the integrations were divided into 35 3-month seasonal means extending from MAM (March, April, May) 1980 through SON (September, October, November) 1988, giving a total of 8 years for the DJF (December, January, February) season and 9 years for all others. An ensemble mean climatology was computed for each of the four seasons and removed from the individual seasonal fields to produce seasonal mean model anomaly fields. Verifying anomaly fields in this fashion is equivalent to making an a posteriori seasonal mean bias correction. Verifying anomaly fields neglects differences due to model bias but allows easier examination of differences in the variance of the forecast and verifying probability distributions. Since the BPE can reveal information about the variance that traditional parametric methods cannot, only the anomaly forecasts are verified here to highlight this ability.

a. Perfect model validation

As a first test, the BPE technique was used to validate the ensemble integrations in a perfect model context. One of the ensemble integrations was designated as the "truth," and the other eight integrations were treated as an eight-member ensemble. For a given season, at each grid point of the model's Gaussian grid, the eight-ensemble members were used to form nine bins for a given scalar field and the "true" value was placed in the appropriate bin. A separate chi-square test was performed at each grid point of the model Gaussian grid. In the first application, a set of nine samples consisting of each of the MAM cases was validated (in this case, each sample in a set corresponds to different external forcing, rather than to different observed states as in the initial value problem of section 3). When the first ensemble integration was designated as the "truth," the chi-square tests indicated that the remaining eight-member ensemble was not a consistent forecast of the "truth." Over 31% of the grid points produced chi-square results with significance less than 0.01. Additional tests revealed the cause of this inconsistency between the AMIP ensemble members. If integration 2 was designated as the truth, and integrations 3–9 were used as a seven-member ensemble forecast, just more than 1% of the grid points produced chi-square significance less than 0.01, and 11% of the points produced significance less than 0.1. This result indicates that integrations 2 through 9 are apparently consistent. A careful study of the AMIP integrations revealed that integration 1 had been performed with a slightly modified gravity wave drag parameterization. The differences in the integrations are subtle enough that they had not been noticed by a number of researchers who had previously used the data, but they were immediately detected by the binning technique.

b. Validation with observations

For the remainder of this section, the first integration of the AMIP model is discarded and the remaining integrations are used as an eight-member ensemble. Three model fields, 200-hPa height, 850-hPa temperature, and precipitation, were compared to observations. The upper air observations were produced from the daily NMC analyzed values. In the isolated cases where NMC daily observations were unavailable, an attempt was made to use the corresponding ECMWF daily analyzed values. The daily values were then averaged at each grid point to form seasonal means corresponding to the AMIP seasonal means described above. The precipitation data were prepared by Schemm et al. (1992) and are composed of data from the world monthly surface station climatology dataset. In all cases, validation was done by interpolating data from the model Gaussian grid to the data grid and then performing binning at each data grid point. For all of the observed data fields, seasonal mean climatologies were generated and removed from the individual seasonal means to produce seasonal anomaly fields.

Figure 5 shows the chi-square significance for the 850-hPa temperatures for MAM. This experiment has nine bins (since there are eight ensemble members) and, purely by coincidence, nine samples (determined by the number of available seasons for which the AMIP integrations are available). Figure 5 shows that the bias-corrected ensemble forecasts are significantly different from the observations over large areas of the globe, particularly over Antarctica, Africa, much of the tropical oceans, and the Himalayas and southeast Asia. In all other regions, the eight samples are too few to resolve differences between the ensemble forecasts and the observations.

Figure 6 shows the chi-square significance for the 850-hPa temperature, but this time for a 35-sample validation set including all seasons of the AMIP integration. As expected, the larger sample size leads to much larger areas for which the forecasts are significantly different from the observations. Despite this, large portions of the extratropics still have ensemble forecasts that cannot be distinguished from the observations. This suggests that the model with bias correction is

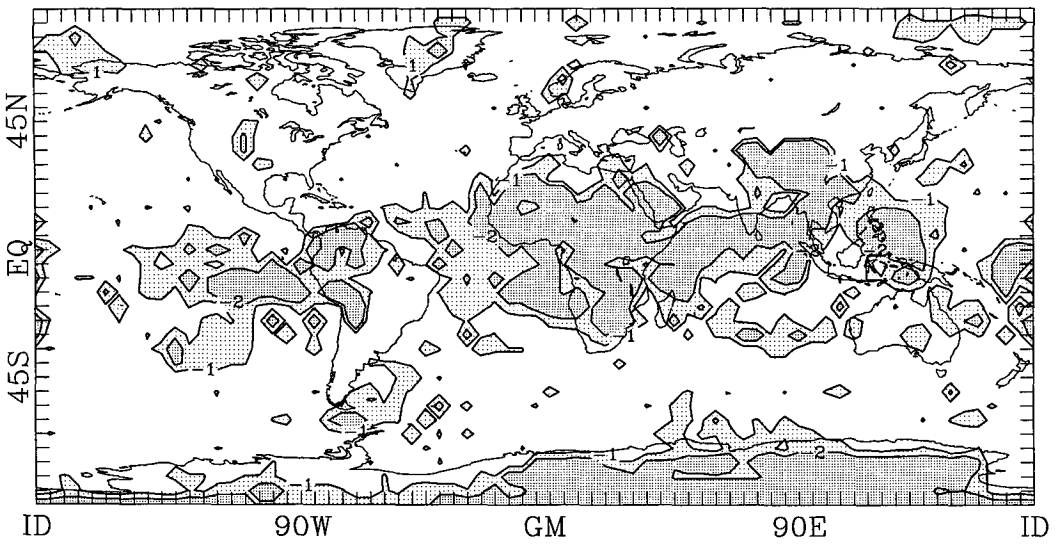

FIG. 5. Significance of chi-square tests for BPE validation of eight-member AMIP ensemble "forecasts" of 850-hPa temperature for MAM. Light (dark) stippled regions have greater than 90% (99%) significance that the ensemble forecasts are inconsistent with the observations.

doing a reasonable job simulating the midlatitude dynamics.

The BPE validation of the ensemble forecasts can be compared to the rms error of the ensemble mean forecasts, shown for MAM in Fig. 7. The dangers of evaluating forecasts based only on quantities such as rms that do not take into account variations in the probability distribution of the forecast variable are readily apparent. Many areas for which the ensemble forecasts are shown to be significantly different from the observations in Fig. 5 are found in areas that have small mean rms error in Fig. 7. Rms and local anomaly correlations for the ensemble mean forecast are poor predictors of ensemble forecast consistency with observations. For instance, some of the lowest chi-square significance for 850-hPa temperature occur over the central Pacific, in a region where rms is particularly low. It is vital that details of the complete ensemble distribution and its relation to the observations be taken into account. This is also true even for single member "deterministic" forecasts that can be verified using a two-bin BPE technique (although this is probably eas-

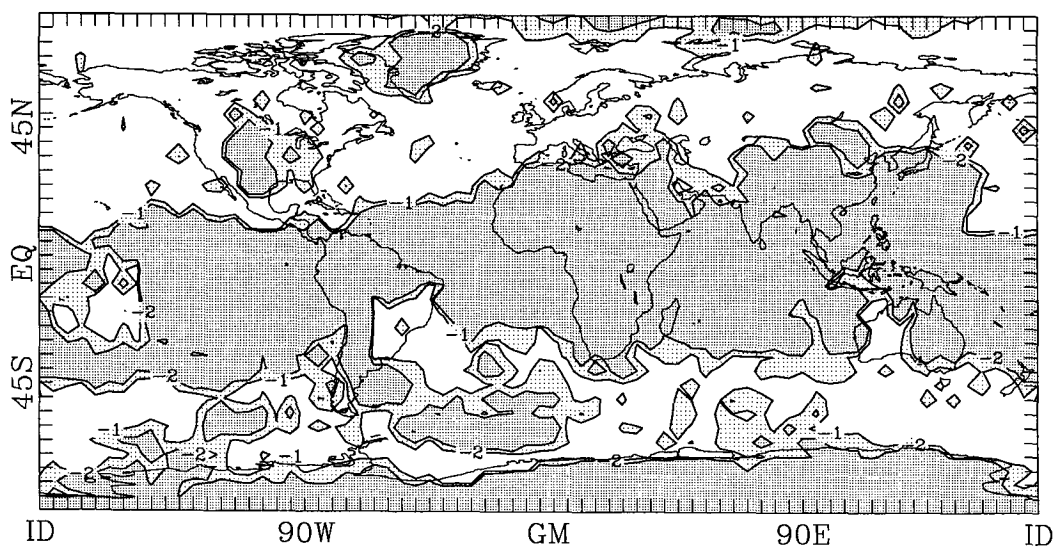

FIG. 6. As in Fig. 5 except validating against observations for all 35 seasons.

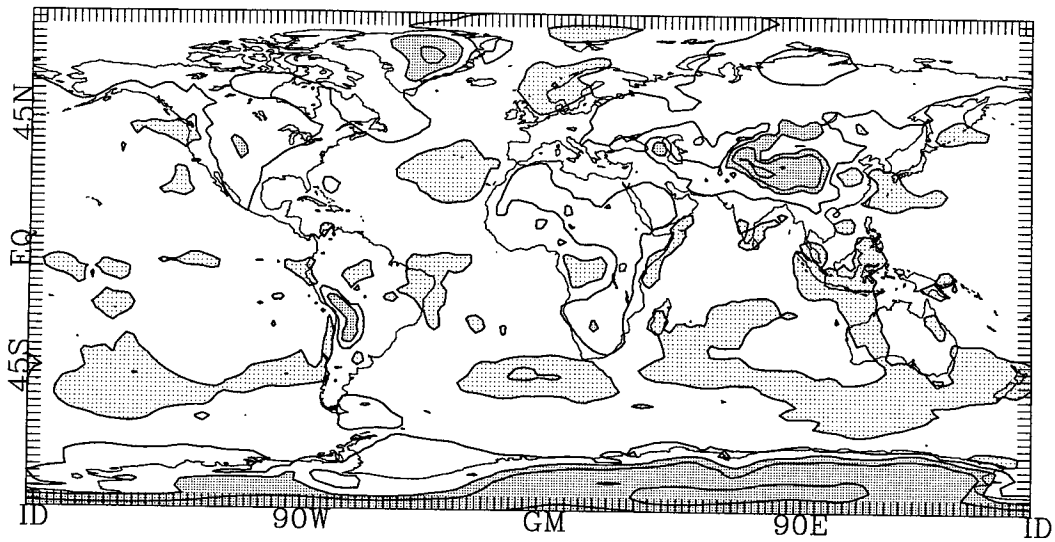

FIG. 7. Rms error of AMIP ensemble mean forecasts for MAM 850-hPa temperature. Contours are at 0.4°, 0.8°, 1.6°, 3.2°, and 6.4°, and 12.8°C; regions with less than 0.8°C are lightly shaded, and regions greater than 3.2° are heavily shaded.

ier to interpret for forecasts of the entire field not the bias-corrected forecasts examined here). The quality of a forecast cannot be judged purely by bias, rms, or anomaly correlation without more detailed knowledge of the structure of the true forecast probability distribution.

An examination of some of the individual gridpoint binning results for the 850-hPa temperature case can help to analyze the nature of the model's systematic errors as well as revealing why the binning validation and rms results are not always in good agreement. In almost all areas where Fig. 5 shows a significant difference between the ensemble forecasts and observations, an examination of the bins reveals that the outermost bins are heavily populated, while the innermost bins are almost empty. In many cases, for instance, for points over Africa and northern South America (Figs. 8a,b), almost all of the observations fall in one of the outermost bins, indicating that the observation is colder (warmer) than the coldest (warmest) forecast. This indicates that the ensemble forecasts do not have enough variability compared to the observations. In other words, the variability of the model ensemble response is not as large as would result if a perfect model were run for the same SST forcings. The rms is not a good predictor of forecast consistency with observations because the variability of the "true" forecast distribution can vary considerably as a function of geographical location. An rms normalized by a measure of this local variability would be an improved predictor of forecast consistency, but even this makes an implicit assumption that the verifying distribution is normal, which may not be a good approximation.

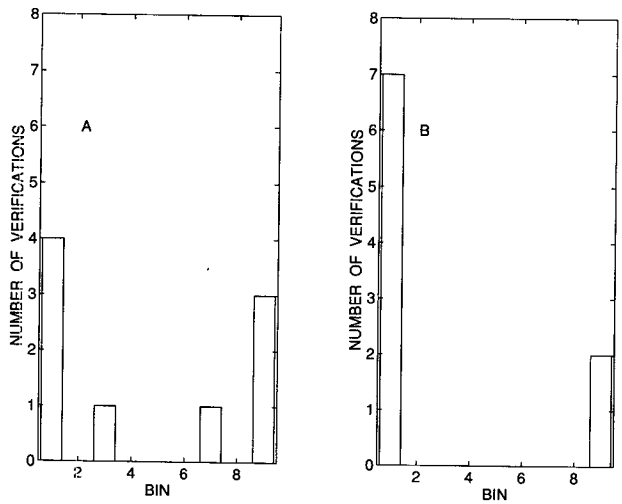

FIG. 8. Distribution of observations in bins for MAM 850-hPa temperature for grid points over Africa (A) and over northern South America (B).

The BPE results are similar for the 200-hPa geopotential height field. Figure 9 shows the chi-square significance for MAM BPE forecasts of this quantity. The forecasts are generally not significantly different from the observations in the extratropics, but in the Tropics, especially over the oceans, there are a number of areas where the forecasts are significantly different. The individual bins in these significant areas show the same behavior as for the 850-hPa temperatures; the outer bins are heavily populated by the observations. Again, this suggests that the model does not produce the observed amount of variability.

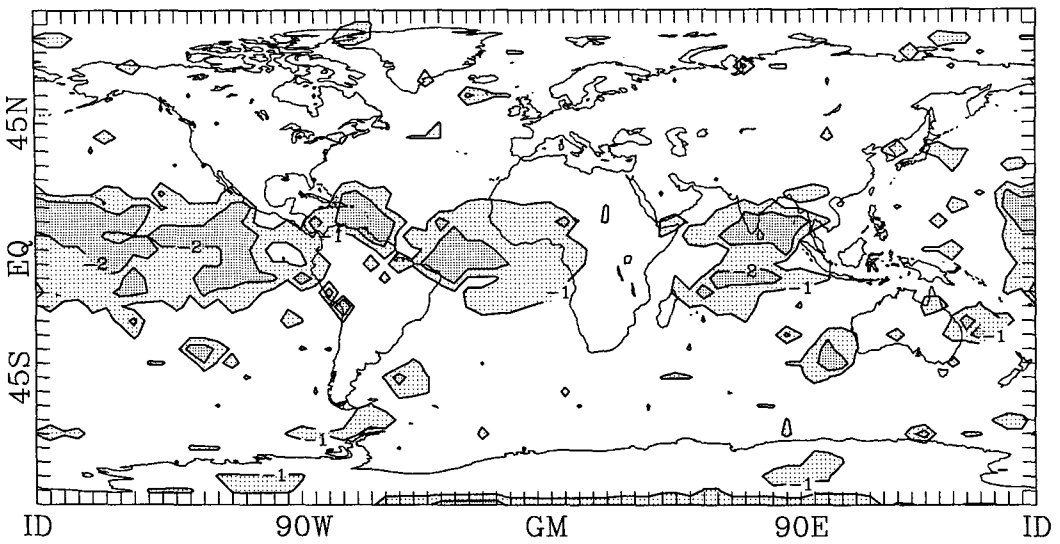

FIG. 9. Significance of chi-square tests for BPE validation of eight-member AMIP ensemble "forecasts" of 200-hPa height for MAM, stippling as in Fig. 5.

Qualitatively different behavior was found for verifications of the precipitation forecasts with the BPE technique. Figure 10 shows the chi-square significance for MAM precipitation forecasts; verification data are generally only available over land regions. There are a number of isolated regions for which significant differences between the ensemble forecasts and the observations are indicated. As is the case for all the results discussed here, the 99% significant areas have large field significance (Livezey and Chen 1983), with about 10% of the available grid points having significance at this level for Fig. 10.

Unlike the upper air forecasts, the precipitation ensemble forecasts have too little variability in some regions where the forecasts are inconsistent with the observations. This can be seen by examining the individual gridpoint bins in the significant areas (not shown). In many areas, for instance over North America, the observations are clustered into just a few of the interior bins.

c. *Relation to potential predictive utility*

Anderson and Stern (1996) used the same AMIP dataset to compute seasons and geographic regions for

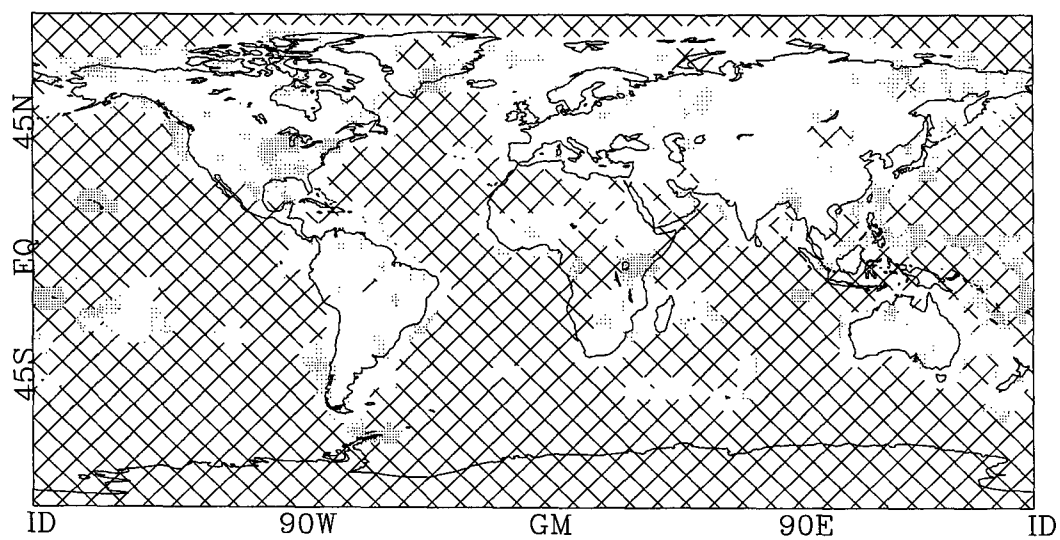

FIG. 10. Significance of chi-square tests for BPE validation of eight-member AMIP ensemble "forecasts" of precipitation for MAM, stippling as in Fig. 5. No verification data was available in cross-hatched regions.

which the ensemble forecasts were statistically significantly different from the forecasts for the same season in all other AMIP years; such regions were said to have potential predictive utility. Potential predictive utility is a nonparametric cousin of the more traditional potential predictability (Chervin 1986) and the reproducibility index of Stern and Miyakoda (1995). Presumably, regions of significant potential predictive utility identify times and places for which the external forcing is compelling enough that the AMIP response to the SSTs can be detected. Examining the results in Anderson and Stern (for MAM as an example) shows that most of the regions where the BPE technique shows significant *differences* between ensemble forecasts and observations correspond to regions that have significant potential predictive utility for both the temperature and height fields. Significant potential predictive utility in this context implies that the AMIP ensembles for individual MAM seasons can be distinguished from MAM ensembles for other years. In other words, these are regions where the height and temperature are significantly dependent upon the external SST forcing in certain years. The fact that these same regions generally have too little interannual variability suggests that the model is unrealistically constrained by the SST anomalies in these regions.

The fact that almost all dynamical quantities have insufficient variability while precipitation has too much variability gives additional clues to possible model deficiencies. It seems likely that the model's precipitation parameterizations are too sensitive in some sense, so that small changes in dynamical quantities can lead to unrealistic large changes in seasonal precipitation.

5. Discussion

The BPE technique is an extremely simple method for producing a probabilistic forecast from an ensemble of forecast model integrations. The ensemble forecasts are used to partition the real line into a number of bins, each of which can be assumed to have an equal probability of containing the "true" forecast. The method does not discard any information from the ensemble and thereby allows users to tailor their own applications in an optimal fashion. The method for producing probabilistic forecasts was successfully tested with perfect model experiments in both a simple initial value problem using the Lorenz convective model and in a boundary-forced problem using an ensemble of AMIP integrations. The method has also been applied at operational centers including the U. K. Meteorological Office, the European Centre for Medium-Range Forecasts, and the National Centers for Environmental Prediction (Harrison et al. 1995), where it is referred to as Talagrand diagrams.

The BPE technique can also be used for validating ensemble forecast probability distributions. Since observational errors are inevitable, the observed condition of the atmosphere will always be represented by a probability distribution. Given this fact, the forecast problem considered here can be phrased as: what is the resulting probability distribution when the observed initial probability distribution is integrated forward in time with a perfect atmospheric model. The probability distribution at the forecast time will be referred to as the "true forecast probability distribution" in the following discussion. It is convenient to regard the true state at the forecast time as being a random sample drawn from this probability distribution. Likewise, the observed state at the forecast time is a random sample from the true observed distribution. If one had the luxury of performing many experiments with the observed atmosphere, one could simply compare the members of an ensemble forecast to samples from the real atmosphere. Unfortunately, only a single sample from the true observed distribution is available for any given initial condition probability distribution. In fact, there do not even exist any reasonable analogs in the observational record (Gutzler and Shukla 1984; Van den Dool 1994), so there seems no way to get around the single sample problem.

Therefore, to validate ensemble forecasts, one must compare them to the observations for a large set of independent initial conditions. The problem is now complicated by the fact that the true observed probability distributions for the different initial conditions (or for the different boundary value forcing in the SST forced problem of section 4) may be drastically different. In fact, it is the prospect that the true observed probability distributions are significantly different from one another that makes the forecast problem of interest (otherwise some climatological forecast would always be a good forecast).

Because the true observed distributions may be quite different from one another, one may no longer be able to use a parametric method (Deque et al. 1994) to compare the ensemble forecasts to the observations. The BPE technique presents a robust, resistant, nonparametric tool for evaluating the consistency with observations of the entire probability distribution forecast produced from ensemble integrations. The method also produces an explicit evaluation of the statistical significance of differences found between the ensemble forecasts and the verifying observations. However, the BPE technique only provides information on the strength of differences indirectly through the forecast sample size needed to produce a given level of significance.

The BPE method can detect differences between ensemble forecast probability distributions and verifying observations that are a result of errors in both the forecast model and in the observational error distribution that is used to generate the ensemble initial conditions and the ensemble "observations." Examination of the bins for particular scalar quantities can give insight into

the systematic error behavior of the ensemble integrations. For example, if all of the verifying observations are clustered into bins on one side of the median bin, the forecast model is clearly producing biased forecasts. Traditional methods are already capable of examining systematic bias (in the mean), however, the BPE method allows examination of higher-order differences between the ensemble forecast distribution and the verification distribution. For instance, examinations of differences between the ensemble forecast variance and the verifying distribution variance can be readily identified, as pointed out in section 4.

The BPE technique could easily be applied to ensemble forecasts of arbitrary real-valued quantities to produce probabilistic forecasts. For instance, application to individual EOF weightings could allow examination of more dynamically relevant quantities than the single gridpoint values examined here. Statistical corrections such as bias correction, or even application of model output statistics (MOS), could be included as part of the forecast model before producing the BPE forecasts.

The BPE method can be applied even to a traditional "deterministic" forecast that can be viewed as a single-member ensemble. In this case, there are only two bins, and observations are either greater than or less than the forecast value. Use of the chi-square test can determine when a distribution of observations in the two bins is inconsistent with the uniform distribution, implying that the forecast and observations are inconsistent. This is related to the computation of model bias but automatically produces a statistical confidence.

The BPE technique evaluates whether ensemble probability forecasts are consistent with observations. However, consistency does not guarantee that the probability forecasts are useful. For instance, a 12-h ensemble forecast generated by randomly sampling observations from climatology would be consistent but not as useful as a forecast produced through operational ensemble forecasts systems. Similar issues involving the "sharpness" of conventional probabilistic forecasts have been addressed by Murphy and Winkler (1987) and Epstein (1969b). A complete ensemble verification system would have to include both consistency measures and additional measures of the size of the bins.

At present, ensemble forecasts produced at operational centers are not generally founded on the notion of equitable sampling of the forecast probability distribution that has been the foundation of the discussion here. Nevertheless, it is interesting to apply the BPE method to the results of ensembles of operational forecast models for either discrete gridpoint variables or for EOF weightings. It is quite possible that the continued improvement in operational models has reduced model systematic errors enough that an equitable sampling approach is not out of the question.

Acknowledgments. The author is indebted to K. Miyakoda, S. Griffies, V. Larichev, J. Lanzante, T. Opsteegh, and two anonymous reviewers who made valuable contributions to this research.

REFERENCES

Anderson, J. L., 1996: Selection of initial conditions for ensemble forecasts in a simple perfect model framework. *J. Atmos. Sci.*, **53,** 22–36.

——, and W. F. Stern, 1996: Evaluating the predictive utility of ensemble forecasts in a perfect model setting. *J. Climate,* in press.

Barker, T. W., 1991: The relationship between spread and forecast error in extended-range forecasts. *J. Climate,* **4,** 733–742.

Barkmeijer, J., P. Houtekamer, and X. Wang, 1993: Validation of a skill prediction method. *Tellus,* **45A,** 424–434.

Brankovic, C., T. N. Palmer, F. Molteni, S. Tibaldi, and U. Cubasch, 1990: Extended-range predictions with ECMWF models: Time-lagged ensemble forecasting. *Quart. J. Roy. Meteor. Soc.,* **116,** 867–912.

——, ——, and L. Ferranti, 1994: Predictability of seasonal atmospheric variations. *J. Climate,* **7,** 217–237.

Cheng, X., and J. M. Wallace, 1993: Cluster analysis of the Northern Hemisphere wintertime 500-hPa height field: Spatial patterns. *J. Atmos. Sci.,* **50,** 2674–2696.

Chervin, R. M., 1986: Interannual variability and seasonal climate predictability. *J. Atmos. Sci.,* **43,** 233–251.

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

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

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

Ferranti, L., F. Molteni, C. Brankovic, and T. N. Palmer, 1994: Diagnosis of extratropical variability in seasonal integrations of the ECMWF model. *J. Climate,* **7,** 849–868.

Gates, W. L., 1992: AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.,* **73,** 1962–1970.

Gleeson, T. A., 1970: Statistical-dynamical prediction. *J. Appl. Meteor.,* **9,** 333–344.

Gordon, C. T., and W. F. Stern, 1974: Spectral modelling at GFDL. The GARP Programme on Numerical Experimentation. *Rep. Int. Symp. on Spectral Methods in Numerical Weather Prediction,* Rep. 7, WMO, Copenhagen, Denmark, 46–80.

——, and ——, 1982: A description of the GFDL global spectral model. *Mon. Wea. Rev.,* **110,** 625–644.

Gutzler, D. S., and J. Shukla, 1984: Analogs in the wintertime 500-mb height field. *J. Atmos. Sci.,* **41,** 177–189.

Harrison, M. S. J., D. S. Richardson, K. Robertson, and A. Woodcock, 1995: Medium-range ensembles using both the ECMWF T63 and unified models—An initial report. UKMO Tech. Rep. 153, 25 pp.

Hoerling, M. P., M. L. Blackmon, and M. Ting, 1992: Simulating the atmospheric response to the 1985–87 El Niño cycle. *J. Climate,* **5,** 669–682.

Hoffman, R. N., and E. Kalnay, 1983: Lagged average forecasting, an alternative to Monte Carlo forecasting. *Tellus,* **35A,** 100–118.

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

Livezey, R. E., and W. Y. Chen, 1983: Statistical field significance and its determination by Monte Carlo techniques. *Mon. Wea. Rev.,* **111,** 46–59.

Lorenz, E. N., 1963: Deterministic nonperiodic flow. *J. Atmos. Sci.,* **20,** 130–141.

Milton, S. F., 1990: Practical extended-range forecasting using dynamical models. *Meteor. Mag.,* **119,** 221–233.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.,* **119,** 299–323.

Murphy, A. H., and R. L. Winkler, 1987: A general framework for forecast verification. *Mon. Wea. Rev.,* **115,** 1330–1338.

Murphy, J. M., 1988: The impact of ensemble forecasts on predictability. *Quart. J. Roy. Meteor. Soc.,* **114,** 463–493.

——, 1990: Assessment of the practical utility of extended range ensemble forecasts. *Quart. J. Roy. Meteor. Soc.,* **116,** 89–125.

Schemm, J., S. Schubert, J. Terry, and S. Bloom, 1992: Estimates of monthly mean soil moisture for 1979–1989. NASA Tech. Memo. 104571, Goddard Space Flight Center, Greenbelt, MD, 252 pp.

Seidman, A. N., 1981: Averaging techniques in long-range weather forecasting. *Mon. Wea. Rev.,* **109,** 1367–1379.

Stern, W., and K. Miyakoda, 1995: The feasibility of seasonal forecasts inferred from multiple GCM simulations. *J. Climate,* **8,** 1071–1085.

Tracton, M. S., and E. Kalnay, 1993: Operational ensemble prediction at the National Meteorological Center: Practical aspects. *Wea. Forecasting,* **8,** 379–398.

——, K. Mo, W. Chen, E. Kalnay, R. Kistler, and G. White, 1989: Dynamic extended range forecasting (DERF) at the National Meteorological Center. *Mon. Wea. Rev.,* **117,** 1604–1635.

Trevisan, A., 1995: Statistical properties of predictability from atmospheric analogs and the existence of multiple flow regimes. *J. Atmos. Sci.,* **52,** 3577–3592.

Van den Dool, H. M., 1994: Searching for analogues, how long must we wait? *Tellus,* **46A,** 314–324.

Wallace, J. M., X. Cheng, and D. Sun, 1991: Does low-frequency atmospheric variability exhibit regime-like behavior? *Tellus,* **43A,** 16–26.