# Evaluating the Potential Predictive Utility of Ensemble Forecasts

JEFFREY L. ANDERSON AND WILLIAM F. STERN

*Geophysical Fluid Dynamics Laboratory/NOAA, Princeton University, Princeton, New Jersey*

(Manuscript received 14 February 1995, in final form 30 June 1995)

ABSTRACT

A method is presented for determining when an ensemble of model forecasts has the potential to provide some useful information. An ensemble forecast of a particular scalar quantity is said to have potential predictive utility when the ensemble forecast distribution is significantly different from an appropriate climatological distribution. Here, the potential predictive utility is measured using Kuiper's statistical test for comparing two discrete distributions. More traditional measures of the potential usefulness of an ensemble forecast based on ensemble mean or variance discard possibly valuable information by making implicit assumptions about the distributions being compared.

Application of the potential predictive utility to long integrations of an atmospheric general circulation model in a boundary value problem (an ensemble of Atmospheric Model Intercomparison Project integrations) reveals a number of features about the response of a GCM to observed sea surface temperatures. In particular, the ensemble of forecasts is found to have potential predictive utility over large geographic areas for a number of atmospheric fields during strong El Niño–Southern Oscillation anomalous events. Unfortunately, there are only limited areas of potential predictive utility for near-surface fields and precipitation outside the regions of the tropical oceans. Nevertheless, the method presented here can identify all areas where the GCM ensemble may provide useful information, whereas methods that make assumptions about the distribution of the ensemble forecast variables may not be able to do so.

## 1. Introduction

Since the early days of numerical prediction of the atmosphere, there has been a continual drive to extend the lead times for which useful forecasts can be provided. Concurrently, there has been a rapidly growing understanding of the basics of nonlinear dynamical systems, which has led to the conclusion that practical forecasting is inherently a stochastic problem in which the ratio of "noise" to forecast "signal" generally increases with lead time. In order to deal with the stochastic nature of the forecast problem, both research and operational atmospheric modeling groups have begun to use ensemble model integrations. Questions of exactly how to utilize the information from an ensemble of model integrations have not yet been satisfactorily resolved. Most groups have examined the ensemble mean forecasts (Milton 1990; Mo and Kalnay 1991), some have examined the farthest outlier members of the ensemble (Mureau et al. 1993), and many have looked at a variety of algorithms for cluster analysis (Brankovic et al. 1990; Murphy 1990; Tracton and Kalnay 1993). Attempts to predict the skill of ensemble mean forecasts have also been made. This has been done by using the growth rates of some class of linear dynamical modes (Palmer 1993) or by using some measure of the ensemble spread as a predictor of the expected skill (Hoffman and Kalnay 1983; Murphy 1989; Brankovic et al. 1994).

Here, a method for evaluating when ensemble forecasts contain potentially useful information is presented. The question to be addressed is, When does an ensemble forecast provide more information than the appropriate (model) climatology? If the climate provides a forecast that is indistinguishable from the ensemble, then there is little sense in utilizing the ensemble integrations.

Traditional methods of evaluating the utility of ensemble forecasts are based upon a priori assumptions about the underlying continuous forecast probability distributions that are being compared. For example, one class of methods for evaluating the utility of an ensemble forecast has been based upon comparisons of the variance of the ensemble to the variance of the climate. If the ensemble variance is not significantly smaller than the climate variance, it has generally been assumed that the ensemble is not useful as a prediction (Shukla 1985). For instance, Stern and Miyakoda (1995) examined the ratio between an ensemble variance and a climate variance, their reproducibility, to assess the potential utility of their ensemble forecasts (technically simulations, see section 2).

*Corresponding author address:* Dr. Jeffrey L. Anderson, Geophysical Fluid Dynamics Laboratory/NOAA, Princeton University, P.O. Box 308, Princeton, NJ 08542.

A second class of methods for evaluating ensembles attempts to find significant differences in the means of the ensemble forecast and the climate control distribution. The Student's t-test of Chervin et al. (1976) and numerous subsequent authors falls into this category.

These traditional methods of assessing the utility of ensemble forecasts make implicit assumptions about what types of differences can exist between the distributions being compared. The variance methods generally assume that the means of the distributions are the same and that the distributions are normal, while methods such as the $f$ test assume that only the means differ while the variance and shape of the distributions are identical. Since ensemble forecasts are primarily being applied to situations in which there is a relatively small amount of forecast signal and a large amount of noise, it seems inappropriate to discard potentially useful information by making any a priori assumptions about the distributions being compared.

In what follows, the usefulness of ensemble forecasts is examined without making any a priori assumptions about the distributions of the forecast variables. A quantity that determines whether ensemble forecasts may provide more information than a climate forecast, the potential predictive utility (PPU), is defined. Statistical methods that are nonparametric, robust, and resistant are used to evaluate this quantity. The PPU is compared to more traditional measures of ensemble forecast usefulness; it is demonstrated that the PPU is a superior measure for determining when the ensemble forecasts can provide useful information.

Section 2 formally defines the concept of PPU, and section 3 presents a statistical method for evaluating this quantity. Section 4 provides an example application to an externally forced simulation (boundary value problem). Section 5 offers some discussion and suggestions for future research.

## 2. Definition of potential predictive utility

A number of previous studies have used the term "predictability" with an assortment of definitions (Lorenz 1965; Shukla 1981; Hayashi 1986; Murphy 1989; also see the review article by Shukla 1985). In order to avoid possible confusion with previous definitions, the term "potential predictive utility" will be defined here. An ensemble forecast is said to have PPU for some quantity $Q$ if the ensemble distribution of $Q$ can be distinguished from an appropriate "climatological" distribution of $Q$ in a statistically significant way. In other words, an $N$-member ensemble with PPU must provide a forecast that is significantly different from the forecast that could be produced by randomly selecting $N$ members from the model's climatological distribution.

An idealized application of the notion of PPU is illustrated in Fig. 1, which displays continuous probability density functions for a hypothetical ensemble forecast and the corresponding climatology. The ensemble forecast distribution at the initial time is determined by some "observational uncertainty." This initial distribution gradually expands and asymptotically approaches the model's climatological distribution as the forecast lead time increases. PPU exists as long as the ensemble forecast distribution can be distinguished from the climatological distribution. In the idealized example of Fig. 1, the ensemble forecast clearly has PPU at lead time 1, but the situation is not as clear by lead time 2. Although Fig. 1 depicts continuous distributions, all that is available in practical forecast situations are some finite samples drawn from the underlying continuous distributions of the forecast and the climate (throughout the following, "distribution" will refer to a discrete sample of the continuous probability distribution unless prefaced by "continuous").

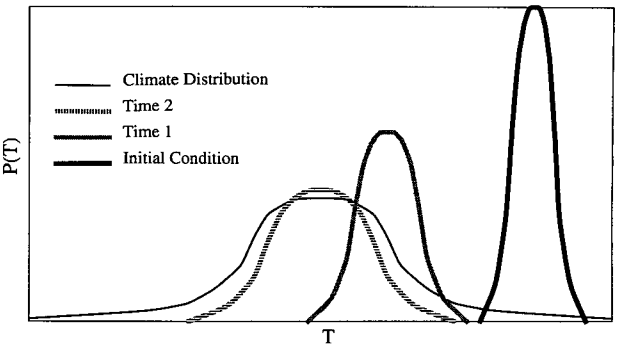


FIG. 1. An idealized depiction of potential predictive utility in an ensemble forecast problem.

It is appropriate at this point to discuss the concepts of significance and strength when examining whether two distributions are different. PPU only evaluates explicitly the significance of differences between two distributions. The strength of the difference (i.e., how large a difference is it) can be indirectly inferred from the size of the ensemble needed to find significant differences. As an example, in Fig. 1, a very small ensemble, for instance one with three members, would certainly be able to provide significant evidence that the lead time 1 and climate distributions are different. However, it might take a very large ensemble to provide significant evidence that the lead time 2 and climate distributions are different.

The idealized example of Fig. 1 was a traditional initial value forecast problem, similar to the weather forecasts produced by operational centers. Another application of PPU would be to a boundary value problem "forecast" similar to those produced for the Atmospheric Model Intercomparison Project (AMIP) (Gates 1992). In this case, time-dependent external forcing is specified and allowed to influence the evolution of the ensemble forecasts. After some sufficiently long inte-

gration time, the ensemble distribution no longer depends upon the details of the initial condition distribution but is instead controlled by the external forcing. In this case, the ensemble may demonstrate PPU only at certain times when the external forcing is unusually compelling. Technically, such boundary value problems are simulations and not truly forecasts; however, the terms "forecast" and "potential predictive utility" will be applied in this context. A boundary value problem of this kind will be examined in section 4.

One of the most delicate issues in studying PPU is the selection of the proper climate distribution. This distribution represents the null hypothesis forecast, which has no PPU by definition. An improper selection of the climate can make all ensemble forecasts appear to have PPU. In general, the climate distribution should be some "large" random sample from the same model that is used to produce the ensemble forecasts. An appropriate climate for the example in Fig. 1 would be a random sample from an extended integration of the forecast model.

The selection of the appropriate climate for the AMIP problem may be considerably more difficult. A sample from a long integration of the model might be appropriate, but this time the long integration must also randomly sample the distribution of external forcing. For instance, a long (climate) integration of the model forced by many ENSO warm event years and very few cold event years would probably be considered an inappropriate climate.

## 3. Statistical tests for comparing distributions

This section discusses a class of statistical tests for evaluating whether two discrete distributions can be distinguished. The null hypothesis for the tests is that the two distributions were randomly selected from the same continuous probability distribution. The tests produce the probability that two discrete distributions randomly selected from the same continuous distribution would be more different than the two given distributions. If the probability is small, then the two discrete distributions can be assumed to come from different continuous distributions. Here, a small value of the probability implies that the ensemble and climate distributions are different and that the ensemble has PPU.

The significance of the Kuiper statistic (Press et al. 1986) is selected as the measure of PPU here. The Kuiper's statistic evaluates differences in the cumulative distributions corresponding to two discrete distributions. There exists an asymptotic series expansion for the statistical significance of the Kuiper statistic that can be evaluated accurately using only a few terms. This is analogous to the use of the incomplete gamma function to evaluate the significance of the more familiar chi-square statistic (Press et al. 1986). The evaluation of the significance automatically takes into account the size of the distributions and the computed statistical significance is reasonably accurate for distribution sizes as small as 4. Small values of the significance correspond to ensemble forecasts with larger PPU. The confidence that the ensemble and climate distributions were sampled from different continuous distributions is equal to 1 minus the value of the Kuiper significance.

It is important to note that the Kuiper test is adept at disproving the null hypothesis, not at proving it (i.e., it can not prove that two distributions are sampled from the same continuous distribution). If one applies the Kuiper test to many pairs of distributions selected at random from the same continuous distribution, the set of test results is closely approximated by a uniform distribution on the interval 0 to 1, $U(0, 1)$. This is roughly analogous to behavior found for the more familiar chi-square test.

There are a number of other statistical tests related to the Kuiper test that can be used to evaluate the difference between two discrete distributions. The most well-known of these tests is the Kolmogorov–Smirnov (KS) test (Knuth 1981). Unlike the Kuiper test, the KS test is not equally sensitive to differences at all points on a distribution but is more sensitive to differences between two distributions that occur close to the median values. The Kuiper test has the more desirable property of being equally sensitive to differences at all points in the distributions, including differences far out on the distribution tails (Press et al. 1986).

A third related statistic is the Anderson–Darling statistic (Anderson and Darling 1954; Best 1994). Like the Kuiper test, it is equally sensitive to differences throughout the distributions. However, there is no simple way to compute the significance of the Anderson–Darling statistic.

All results presented below were evaluated using all three statistics, although results are shown only for the Kuiper test. In general, there was very little difference between the results using the three different methods. It is, however, relatively easy to construct cases in which the KS test is inferior to the Kuiper and Anderson–Darling tests. Given the relative ease of computing the Kuiper test, it was decided that this statistic was preferable to the Anderson–Darling statistic for the purposes of this presentation. Note that the concept of PPU is independent of the statistical tool used to distinguish between two distributions. It is possible that more powerful statistical tools could prove to be superior to the Kuiper test for evaluating PPU.

One traditional way for evaluating quantities similar to the PPU (often referred to as predictability or potential predictability) has been to measure the variance of the members of an ensemble forecast and to compare this to the variance of a sample of the climate (Shukla 1981; Hayashi 1986). If the variance of the ensemble is larger than that of the climate, predictability is generally said to be lost. This same concept is related to the use of ensemble spread to obtain an a priori estimate

of perfect model skill (Barker 1991; Hoffman and Kalnay 1983) and skill of real forecasts (Brankovic et al. 1990; Murphy 1990). The potential predictability of Chervin (1986) is also a ratio of variances between simulations with climate-mean boundary forcing and the observed climate.

The PPU as measured by the significance of the Kuiper statistic is a more powerful measure than those that only make use of information about the variance of the ensemble and the climate. Figure 2 shows two examples where comparing the variances can be misleading. In Fig. 2a, the solid curve is the climate distribution (the continuous curves here can be thought of as the continuous distributions from which the discrete distributions of the ensembles and the climate are sampled), while the dashed curves are two ensemble distributions that differ only in a translation of their median values. The traditional measures that make use of only the variance would find no difference between these two ensemble distributions; however, the PPU of these two ensembles is very different. Ensemble 1 has a distribution that is quite similar to the climate control; small discrete distributions would almost certainly fail to distinguish these distributions and would indicate no PPU. Ensemble 2 has a distribution that is very different from the climate; even a very small ensemble would be able to determine that these two distributions are different. In practical terms, if these were distributions of temperature at a model grid point, ensemble 1 would have a distribution virtually indistinguishable from the long-term climate while ensemble 2 would indicate that the forecast temperature was certain to be many standard deviations above normal. For the same size of ensemble, ensemble 2 would have a greater PPU than ensemble 1.

Figure 2b shows another extreme example in which tests making use of only the variance would be misleading. In this case, the variance of the ensemble forecast is much larger than that of the climate, which would traditionally be interpreted as indicating that the ensemble has no PPU. However, the distributions are clearly different, which would be reflected in the significance of the Kuiper statistic. The ensemble forecast of temperatures either far above or far below normal clearly has PPU. This is a particularly clear example of an instance in which the variance ratio technique's assumption of normality for the ensemble distribution is inappropriate.

As noted in the introduction, a second class of tests for evaluating ensembles has been based on detecting significant differences in the means of two distributions (Chervin 1976). It is obviously straightforward to construct examples analogous to those of Fig. 2 but in which distributions have the same mean despite having entirely different variances or shapes. An ensemble forecast distribution with the same mean but significantly different variance/shape from the climate distribution clearly has potential to provide useful forecast information.

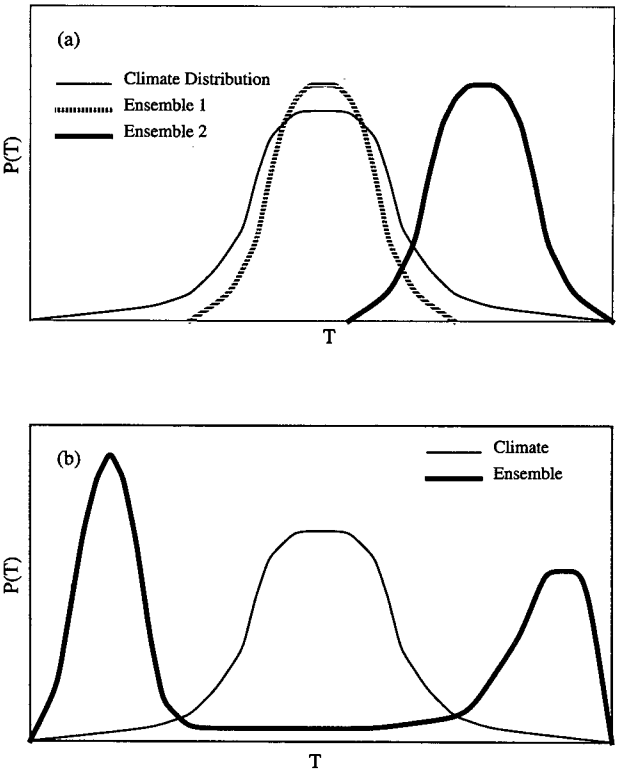


FIG. 2. Idealized distributions of ensemble forecasts and climate. In (a) the two ensembles have identical variances but markedly different PPU. In (b) the ensemble has a very large variance but a significant PPU.

## 4. Boundary value problem in a GCM

An example of the concept of PPU is presented using an ensemble of integrations in an atmospheric GCM forced by observed sea surface temperatures. 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 AMIP (Gates 1992). A 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).

The first 14 months of the ensemble integration were discarded in an attempt to eliminate direct effects of the initial conditions. The remainder of the ensemble integration was divided into 35 three-month seasonal means extending from MAM (Mar, Apr, May) 1980

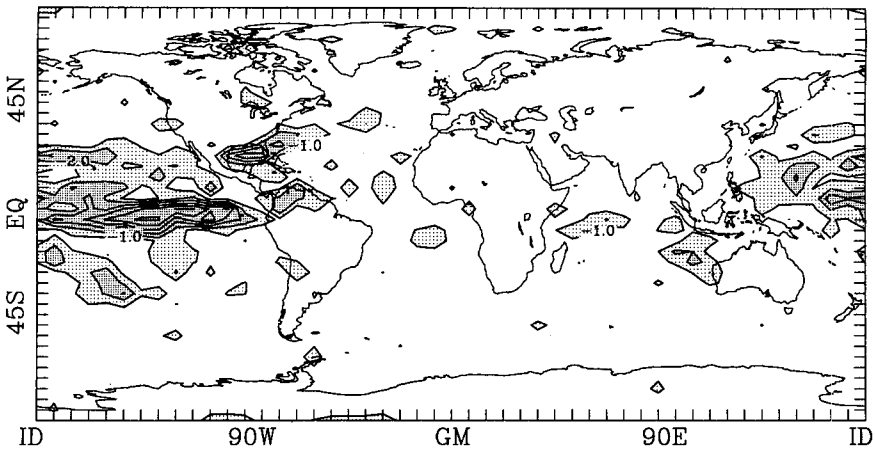


FIG. 3. Log of PPU for precipitation field from DJF 1982/83. Lightly shaded areas have values of less than 0.1 (confidence of statistical significance > 90%), while heavy shading is used for values of less than 0.01 (confidence of statistical significance > 99%).

through SON (Sep, Oct, Nov) 1988 giving a total of eight years for the DJF (Dec, Jan, Feb) season and nine years for all others.

The question of interest in this ensemble integration is whether the distributions of seasonal-mean atmospheric fields can be significantly influenced by the observed SST forcing. In other words, can the distribution of the nine-member ensemble for a single season in a given year be distinguished from the model's climate distribution for that season? In this case, the most appropriate available climate distribution is the set of eight (seven for DJF) nine-member ensembles for the same season but in all the other years (potential shortcomings of this choice of climate will be discussed later in this section). If the distributions can be distinguished, the ensemble forecast for the selected year is said to have PPU. The significance of the Kuiper statistic is used to evaluate the PPU of the ensemble integrations at each grid point on the 64° latitude by 128° longitude Gaussian grid that is used in this T42 model.

Figure 3 plots the log of the PPU (the Kuiper significance) of the model precipitation field for DJF 1982–83, in the middle of the unusually strong 1983 ENSO warm event. Areas that have PPU at the 90% confidence level are lightly shaded, while areas with greater than 99% confidence are heavily shaded. The vast majority of areas with significant PPU are located over the tropical Pacific Ocean, where the SST anomalies are consistently strongest. There are, however, a few continental areas that also have PPU, including the southeast United States and northern South America. The area over the United States is intriguing, since observational studies (Ropelewski and Halpert 1986) have suggested that this is a region where precipitation may be affected by ENSO. Figure 4 shows the same quantity as in Fig. 3 for the next season, MAM 1983.

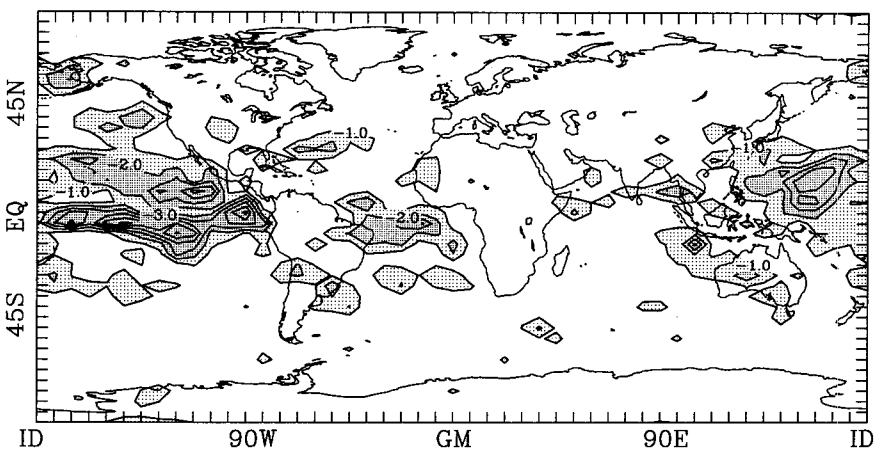


FIG. 4. Log of PPU for precipitation field from MAM 1983. Shading as in Fig. 3.

Most significant areas of PPU are still over the tropical oceans. However, new significant continental regions are found over Northeast Brazil and over Australia, again somewhat consistent with observational studies (Ward and Folland 1991; Joseph et al. 1991).

As pointed out in section 3, the significance of Kuiper's statistic would assume some small values by chance, even if the ensemble forecast had no significant PPU. Figure 5 shows the percentage of grid points for which the Kuiper significance was less than 0.01 (99% confidence level) as a function of the individual season for precipitation, 850-mb temperature, and 200-mb height. The percentage of points less than 0.01 is always considerably greater than 1%, demonstrating that there is "map significance" (Livezey and Chen 1983) for PPU. All three fields have significant PPU for a larger number of grid points during the 1982–83 ENSO warm event and for fewer grid points during 1985–86. Some additional interesting behavior can be seen in the 200-mb height field, which has significant PPU for a maximum number of points during SON 1988. The precipitation field generally has a much smaller area of significant PPU than the two upper air fields; unfortunately, precipitation is a much more interesting quantity to forecast as a seasonal mean.

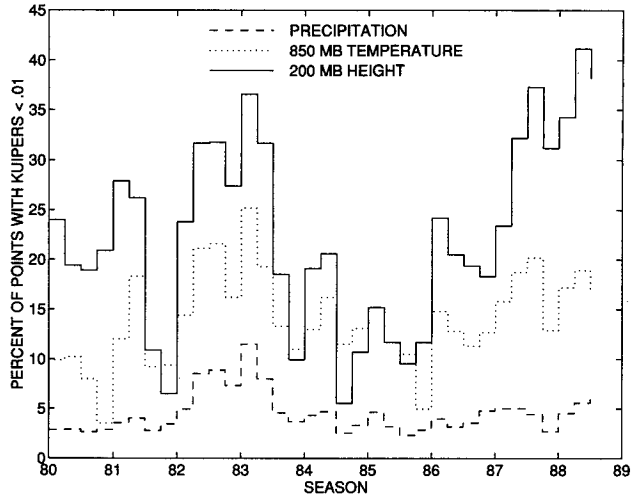


FIG. 5. Percentage of grid points with the Kuiper statistic significance less than 0.01 as a function of season for precipitation, 850-mb temperature, and 200-mb heights.

Several other plots of PPU can reveal additional interesting aspects of the model response to SST forcing. Figure 6 displays the PPU for 850-mb temperatures for MAM 1983, which has the largest number of significant points for this field. Although most of the significant PPU is over the oceans, there are also large areas over land that are significant, including much of South America and northwest North America. With the exception of the 1982–83 ENSO event, the 850-mb temperature fields have almost no significant PPU over land areas. These results can also be compared to findings from observational studies, for instance those for surface temperature in Halpert and Ropelewski (1992).

As noted in Ebisuzaki (1995), the tropical 200-mb heights in the AMIP integrations are strongly tied to the tropical SSTs; almost all seasons display a broad swath of significant PPU across the Tropics for this field. In general, the significant PPU of the 200-mb heights is limited to the Tropics and to some extratropical regions over the Pacific. However, during ENSO warm events, there is significant PPU over additional regions of the extratropics. Figure 7 shows the PPU for 200-mb heights for MAM 1987. Of particular interest are the areas of significant PPU over the eastern North Pacific and North America. These three centers are reminiscent of the centers of the PNA pattern that observational studies (Horel and Wallace 1981) have suggested is forced by anomalous tropical SSTs. This figure suggests that the model reproduces some aspects of the PNA and has PPU near the centers of the response.

The shortcomings of using traditional variance ratios, or similar functions of the variance, as measures of ensemble forecast utility are particularly noticeable for fields such as precipitation. Distributions of both total precipitation and anomalous precipitation, in both the model and the real world, are fundamentally nonnormal (especially in relatively arid regions) because of the positive definite nature of precipitation. Even worse, the shapes of the precipitation distributions for relatively wet and dry points are fundamentally different. For this reason, variance ratios, which implicitly assume normal distributions, are particularly inappropriate for such fields. Similar shortcomings are shared by measures of significant differences in the mean of distributions such as the Student's t-test. Figure 8 displays the ratio of the 9-member ensemble variance for the MAM 1983 precipitation to the variance for the remaining 72-member climate distribution from all other MAM seasons; this is the reproducibility of Stern and Miyakoda (1995). Figure 8 can be compared to Fig. 4, which displays the PPU for the same season. Although there are a number of similarities between the plots, there are also a number of significant differences. For example, Fig. 4 shows a large area of highly significant PPU along the equator in the eastern Pacific in an area where Fig. 8 shows very large variance ratios (traditionally interpreted as indicating that the ensemble forecast was of no utility). This region is consistently quite dry during MAM in the model, but during the 1983 warm event, all ensemble members demonstrate much above normal precipitation. In this case, there is a tremendous amount of useful information available in the ensemble forecasts that cannot be identified by quantities based on variance.

The southeast United States and western Australia are other regions where there are differences between the PPU of Fig. 4 and the variance ratio of Fig. 8. Both regions have weakly significant PPU in Fig. 4 but have

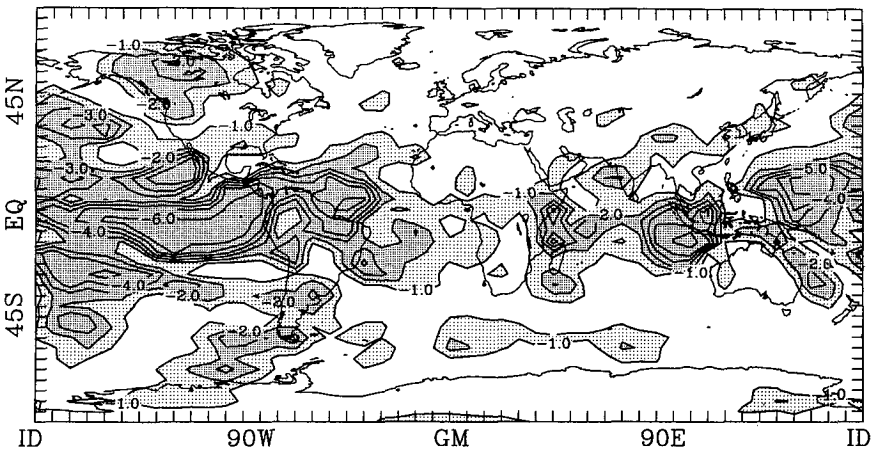


FIG. 6. Log of PPU for 850-mb temperature from MAM 1983; shading as in Fig. 3.

variance ratios close to unity. The anomalous response in these regions for MAM 1983 changed the distribution without significantly affecting the variance.

The discussion above has pointed out differences between the PPU and measures based on variance ratio for the precipitation field, which is not expected to have a constant distribution shape. There are also significant differences for fields such as 850-mb temperature, which might be more reasonably assumed to be normally distributed. Figure 9 shows the cumulative distributions for the MAM 1983 ensemble and the MAM climate distribution from all other years for 850-mb temperature at three individual grid points. These points were selected from regions where the PPU is large but where the variance ratio was larger than 1. Such points generally make up a considerable fraction of the points with significant PPU, about 25% in this 850-mb temperature example.

Figure 9a shows the distributions for a point over the eastern tropical Pacific; here, the response to the 1983 SST anomalies has been a large positive shift of the 850-mb temperature. The variance ratio suggests no useful information was available from the ensemble at this point.

Figure 9b demonstrates another type of behavior that can lead to significant PPU in regions with large variance ratios. In this case, a point over the western tropical Pacific, the ensemble forecast distribution is highly skewed with one extreme positive outlier that leads to the large variance ratio.

Finally, Fig. 9c shows an intriguing ensemble distribution from a region in the Indian Ocean. In this case, the ensemble distribution is bimodal, with three members indicating temperatures about 2°C below normal and the other six members indicating temperatures in excess of 1°C above normal. This leads to a large variance ratio, but the significance of the Kuiper statistic indicates that there is a significant difference between the ensemble and the climate. In such a case, one could justifiably make a forecast that

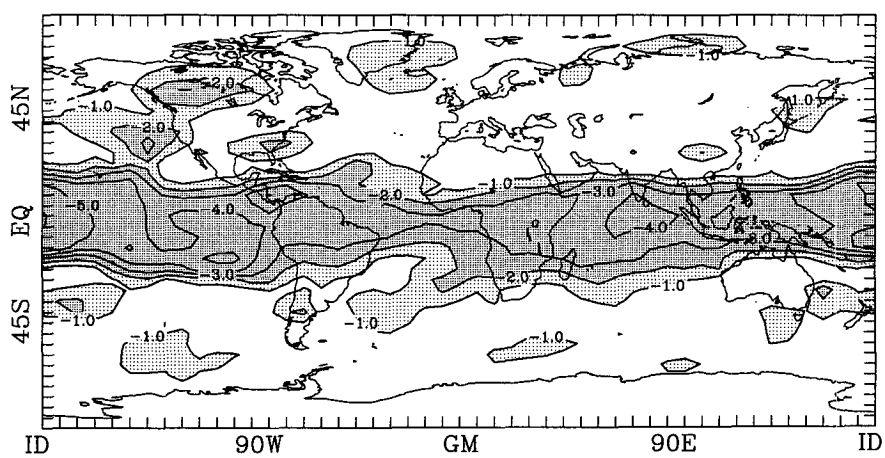


FIG. 7. Log of PPU for 200-mb heights for MAM 1987. Shading as in Fig. 3.

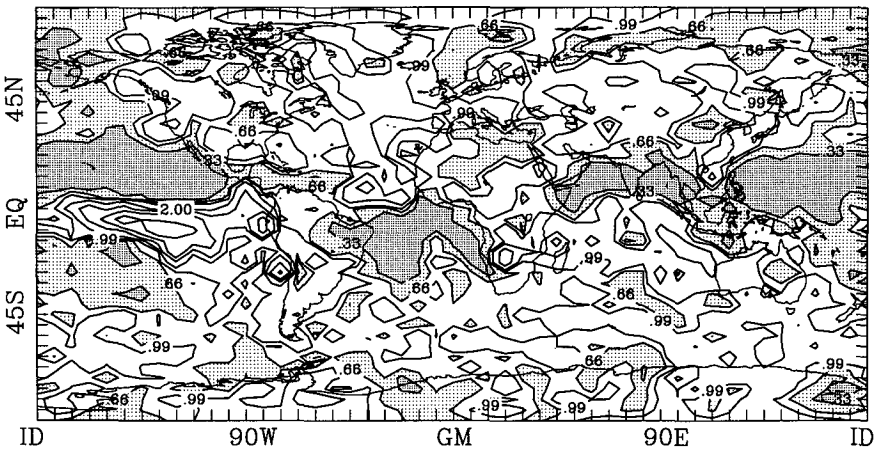


FIG. 8. Ratio of variance of 9-member 1983 MAM precipitation ensemble to variance in the remaining 72-member MAM climate distribution. Contours are at 1/3, 2/3, 1, 2, 4, 8, 16, and 32. Regions less than 1/3 are shaded dark and regions between 1/3 and 2/3 are lightly shaded.

temperatures would be either much above or much below normal at this point. It is an interesting and unanswered question whether this model produces significant areas of such apparently bimodal behavior and, if so, what type of dynamical mechanisms are responsible.

Methods like the Student's t-test are also inappropriate for comparing distributions like those shown in Figs. 9b,c (although the t-test would indicate that significant differences exist between the distributions in Fig. 9a). In the latter two figures, the shapes of the pairs of distributions are fundamentally different. Figure 9c is especially problematic since the distributions have very similar means despite having very different shapes.

The climate used as control in these experiments is from a relatively short period and may not be an adequate sample of the complete range of the model's behavior. A longer climatology would generally tend to increase the PPU. However, tests in which several years are removed at random from the climatology field while testing the PPU of a particular season suggest that these changes would probably be minor, assuming of course that the entire period of the climate sample is not somehow anomalous compared to the longer-term climate.

It is also important to recall that the notion of PPU defined here is dependent on the size of the ensemble. A larger ensemble would certainly be able to find more significant differences between the ensemble and climate distributions. Of course, the differences being detected can become increasingly subtle as the ensemble and climate distribution sizes increase. One could question whether differences in distributions that would require a very large ensemble to detect would ever have a substantial impact for anyone using a forecast.

## 5. Conclusions

Potential predictive utility has been offered as a quantity for measuring the ability of an ensemble forecast to provide more information than some appropriate climatological control forecast. The PPU examines the differences between discrete samples from an ensemble forecast and the climatological control. The Kuiper test, which is one means of numerically evaluating the PPU, automatically provides a confidence level for discrete distributions of arbitrary size. While this statistical test only evaluates the significance of the difference between the ensemble distribution and the climate distribution, it indirectly provides information about the strength of the difference. If significant differences can be found with a small ensemble, the difference between the distributions is larger than if a very large ensemble is needed to find significant differences.

In the previous section, the PPU has been applied to model gridpoint values. It could also be applied to other scalar quantities, for instance to the coefficients of EOFs from a long model integration. Applying the PPU to EOFs might help to identify particular patterns that could be successfully forecast. This in turn might lead to a better understanding of the dynamical processes involved.

In general, ensemble forecasts are likely to be applied at the extreme limits of forecast problems where the forecast signal is becoming small compared to noise. In such situations, it seems natural to use as much of the available information from the ensemble as possible. Traditional techniques for evaluating such quantities as potential predictability or reproducibility implicitly assume normal distributions with identical means for both the ensemble and climate distributions by using only the variance of the distributions. Techniques such as the Student's t-test implicitly assume

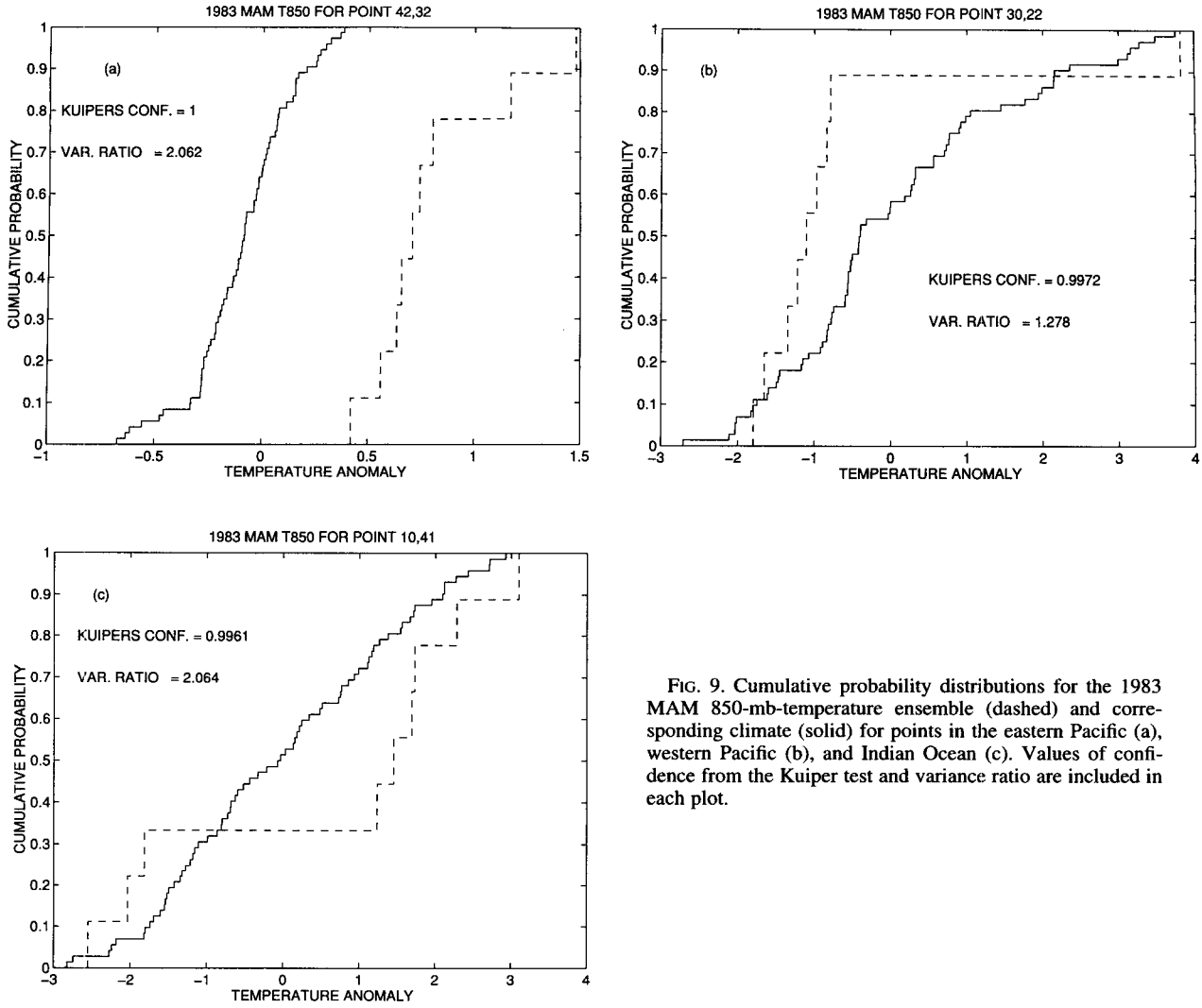


FIG. 9. Cumulative probability distributions for the 1983 MAM 850-mb-temperature ensemble (dashed) and corresponding climate (solid) for points in the eastern Pacific (a), western Pacific (b), and Indian Ocean (c). Values of confidence from the Kuiper test and variance ratio are included in each plot.

that the variance and shape of the distributions being compared are identical. Since the PPU, when computed using the Kuiper test, is a nonparametric test, it does not make any such assumptions and hence does not discard information as do the traditional techniques. In the AMIP simulations, the PPU is able to identify instances where an ensemble provides useful information, but where traditional tests would indicate the ensemble was not useful. The traditional tests are a particularly inappropriate choice for fields such as precipitation, which are inherently not normally distributed. However, even for fields that are likely to be more nearly normally distributed, the PPU can identify instances of useful ensemble forecasts that would be overlooked by traditional tests. The conclusion is that the PPU, by its very definition, is a more appropriate method for establishing whether an ensemble forecast provides more information than a control climate forecast. Once regions with significant PPU have been identified, a closer examination of these regions with additional tools (for example, the traditional variance ratio and Student's t-tests) can identify the details of the differences between the ensemble and climate distributions.

Extending the notion of PPU to real forecasts portends to be, not surprisingly, fraught with additional difficulty. However, when making ensemble forecasts, one can begin by using the PPU as defined above to identify those regions where the model forecast is significantly different from its climatology. It is only in these regions that the forecast can possibly be expected to provide more information than the real (observed) climatology.

The application of PPU to the AMIP ensemble has some implications for the use of models to predict seasonal atmospheric response given a perfect (or good) forecast of the SST forcing. In this particular T42 spectral model, there are a number of fields, such as 850-mb temperatures, for which the model has consistent PPU over the tropical oceans. The ensemble has PPU

for 200-mb heights over even broader areas spanning most of the Tropics. There are periods of particularly anomalous SST forcing, such as the 1982–83 ENSO warm event, when the areas with PPU expand for all fields. In these cases, there is PPU for 850-mb temperature over broad areas of the tropical continents and also some areas over land in the extratropics. The PPU is significant over much smaller areas for precipitation and for other near-surface model quantities such as surface temperature and soil moisture. Nevertheless, during periods of anomalous SST forcing, significant PPU for these fields exists over limited areas of both the tropical and extratropical continents. If the model dynamics are a somewhat faithful simulation of those in the real world, this is additional support for the observational evidence that anomalous temperature and precipitation occur in response to strong SST anomalies. It also suggests that current models, despite their shortcomings, might be able to provide useful predictions for some variables of interest over limited areas of the extratropical continents in cases of unusually anomalous tropical SST.

*Acknowledgments.* The authors are grateful to John Lanzante, Gabriel Lau, Edward Epstein, and two anonymous reviewers for their valuable critiques of earlier versions of this paper.

## REFERENCES

Anderson, T. W., and D. A. Darling, 1954: A test of goodness of fit. *Amer. Stat. Assoc. J.*, **49,** 765–769.

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

Best, D. J., 1994: Nonparametric comparison of two histograms. *Biometrics,* **50,** 538–541.

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.

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

——, and S. Schneider, 1976: On determining the statistical significance of climate experiments with general circulation models. *J. Atmos. Sci.,* **33,** 405–412.

Ebisuzaki, W., 1995: The potential predictability in a 14-year GCM simulation. *J. Climate,* **8,** 2749–2761.

Gates, W. L., 1992: AMIP: The atmospheric model intercomparison project. *Bull. Amer. Meteor. Soc.,* **73,** 1962–1970.

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, Copenhagen, WMO, Rep. No. 7, 46–80.

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

Halpert, M. S., and C. F. Ropelewski, 1992: Surface temperature patterns associated with the southern oscillation. *J. Climate,* **5,** 577–593.

Hayashi, Y., 1986: Statistical interpretations of ensemble-time mean predictability. *J. Meteor. Soc. Japan,* **64,** 167–181.

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

Horel, J. D., and J. M. Wallace, 1981: Planetary scale atmospheric phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.,* **109,** 813–829.

Joseph, P. V., B. Liebmann, and H. H. Hendon, 1991: Interannual variability of the Australian summer monsoon onset: Possible influences of Indian summer monsoon and El Niño. *J. Climate,* **4,** 529–538.

Knuth, D. E., 1981: *The Art of Computer Programming. Vol. 2, Seminumerical Algorithms.* Addison-Wesley, 688 pp.

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., 1965: A study of the predictability of a 28-variable atmospheric model. *Tellus,* **17,** 321–333.

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

Mo, K. C., and E. Kalnay, 1991: Impact of sea surface temperature anomalies on the skill of monthly forecasts. *Mon. Wea. Rev.,* **119,** 2771–2793.

Mureau, R., F. Molteni, and T. N. Palmer, 1993: Ensemble prediction using dynamically conditioned perturbations. *Quart. J. Roy. Meteor. Soc.,* **118,** 299–323.

Murphy, J. M., 1989: 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.

Palmer, T. N., 1993: Extended-range atmospheric prediction and the Lorenz model. *Bull. Amer. Meteor. Soc.,* **74,** 49–66.

Press, W. R., B. P. Flannery, S. A. Teulosky, and W. T. Vetterling, 1986: *Numerical Recipes: The Art of Scientific Computing.* Cambridge University Press, 702 pp.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Nino/Southern Oscillation (ENSO). *Mon. Wea. Rev.,* **114,** 2352–2362.

Shukla, J., 1981: Dynamical predictability of monthly means. *J. Atmos. Sci.,* **38,** 2547–2572.

——, 1985: Predictability. *Advances in Geophysics,* Vol. 28b, Academic Press, 87–122.

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

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

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north nordeste of Brazil. *Int. J. Climate,* **11,** 711–743.