Seasonal Predictions, Probabilistic Verifications, and Ensemble Size

ARUN KUMAR

Environmental Modeling Center, NCEP/NWS/NOAA, Camp Springs, Maryland

ANTHONY G. BARNSTON

International Research Institute for Climate Prediction, Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York

MARTIN P. HOERLING

Climate Diagnostics Center, NOAA–CIRES, Boulder, Colorado

(Manuscript received 29 February 2000, in final form 24 July 2000)

ABSTRACT

For the case of probabilistic seasonal forecasts verified by the rank probability skill score, the dependence of the expected value of seasonal forecast skill on a hypothesized perfect atmospheric general circulation model's ensemble size is examined. This score evaluates the distributional features of the forecast as well as its central tendency. The context of the verification is that of interannual variability of the extratropical climate anomalies forced by sea surface temperatures in the tropical Pacific associated with ENSO. It is argued that because of the atmospheric internal variability, the seasonal predictability is inherently limited, and that this upper limit in the average skill is the one that can be achieved using infinite ensemble size. Next, for different assumptions of signal-to-noise ratios, the ensemble size required to deliver average predictive skill close to inherent skill is evaluated.

Results indicate that for signal-to-noise ratios of magnitudes close to 0.5, the typical ensemble size currently used for the seasonal prediction efforts (i.e., 10–20 members), is sufficient to ensure average skill close to what is expected based on infinite ensemble size. For smaller standardized seasonal mean atmospheric anomalies, the ensemble size required to obtain predictive skill close to the inherent limit increases dramatically. But for these cases the expected skill itself is very low and the use of larger ensemble size has to be judged against the marginal level of prediction skill. Further, for small signal-to-noise ratios, forecasting the climatological distribution becomes nearly as effective as accurately defining the slight deviations from climatology.

1. Introduction

In a recent study Kumar and Hoerling (2000) analyzed the effect of ensemble size on the average seasonal prediction skill for different signal-to-noise ratios. In their study two metrics for seasonal prediction skill, that is, spatial correlation and the mean-square error for the ensemble mean as the forecast, were analyzed. It was shown that the gain in the spatial anomaly correlation skill score as a function of ensemble size depends strongly on the ratio of the boundary forced signal and the internal variability (or the climate noise) of seasonal mean atmospheric states. For both small and large signal-to-noise ratios the size of the ensemble was shown to have little impact on the spatial correlation skill. The largest gain in the predictive skill with increasing ensemble size occurred for the intermediate signal-to-noise ratios. We should point out that within the context of seasonal predictions, the signal-to-noise is generally defined as the ratio of the strength of the boundary forced atmospheric signal and the standard deviation of the internal variability of the seasonal atmospheric means.

It is reasonable to argue that a shortcoming of ensemble mean prediction methodology is that it does not take into account the information about the spread of individual members within the ensemble. This information can be used if a probabilistic seasonal forecast, in contrast to a deterministic seasonal forecast as in the case of ensemble mean, is employed. An obvious question then is how the average skill of the probabilistic seasonal forecasts derived from the ensemble of atmospheric general circulation model (AGCM) integrations may depend on the size of the ensemble. Analysis of this dependence is the focus of this paper.

Corresponding author address: Dr. Arun Kumar, Climate Modelling Branch, NOAA/NCEP, Room 807, 5200 Auth Road, Camp Springs, MD 20746.
E-mail: arun.kumar@noaa.gov

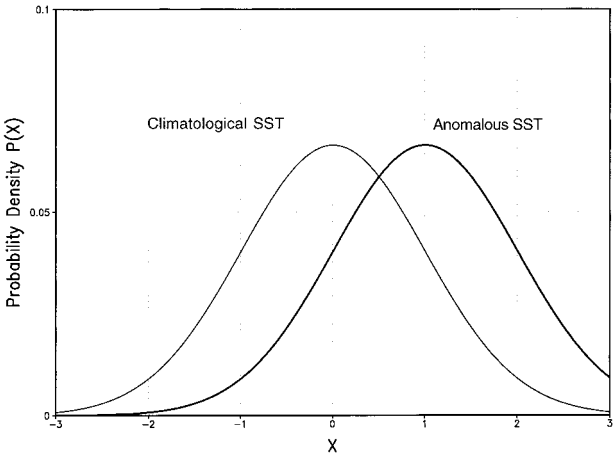

FIG. 1. Comparison of the PDF for the seasonal mean atmospheric states for climatological tropical Pacific SSTs vs that for anomalous SSTs. Anomalous SSTs bias the PDF such that the mean (or the first moment) of the PDF is nonzero. Here we assume that SST anomalies do not impact the spread of the PDF.

Out of several possible metrics available for quantifying the skill for the probabilistic seasonal forecasts, we analyze in the variation of the rank probability skill score (RPSS) with the ensemble size for different signal-to-noise ratios. Details of this particular skill measure are reviewed in section 2. The assumptions for the seasonal mean atmospheric variability and analysis procedure are also discussed in this section. Results on the dependence of RPSS on the ensemble size are described in section 3.

2. RPSS and the analysis procedure

a. Model of atmospheric seasonal variability

We first assume that for a fixed boundary forcing, for example, sea surface temperatures (SSTs), the variability of the seasonal mean atmospheric states in the extratropical latitudes can be characterized by a Gaussian distribution with mean μ and standard deviation σ. A nonzero value for standard deviation of the probability density function (PDF) characterizing the seasonal mean states signifies the fact that for fixed SSTs, the extratropical seasonal mean state is not deterministic and can vary from one realization to another.

Interannual variation in the SSTs can impact different moments of the PDF. A particular example of this is shown in Fig. 1 where two PDFs for the seasonal mean states, one for the climatological SSTs and the other for the anomalous SSTs, are shown. The first moment of the PDF of the seasonal mean atmospheric states for the climatological SSTs is, by definition, zero. One possible impact of the anomalous SSTs is that they may bias the PDF, and hence, for the anomalous SSTs the first moment of the PDF, that is, μ, has a nonzero value. In this particular example, without any loss of generality, we have assumed that the interannual variations in SSTs do not impact the spread of the seasonal means about their respective expected values. We should point out that these PDFs can be thought to represent the characteristics of seasonal mean variability of some atmospheric variable (e.g., geopotential height) at a particular geographical location. Alternatively, these PDFs can also represent the interannual variability of indices of different low-frequency modes of extratropical atmospheric variability, for example, the Pacific-North American mode.

For the above assumptions for the seasonal atmospheric variability, the signal due to SSTs is manifested as the shift in the first moment of the PDF. Further, this shift can also have geographical dependence. For example, it is a well known fact that in the extratropical latitudes and during the boreal winter, the influence of the interannual changes in the tropical Pacific SSTs are mostly confined to the Pacific-North American (PNA) region (Horel and Wallace 1981; Trenberth et al. 1998). The strength of the change in the mean of the PDF with SSTs relative to the magnitude of the spread of the PDF can be considered as a basic measure of the predictability, or signal-to-noise, in the observed system.

In principle, PDFs for the seasonal mean atmospheric states can be constructed for different anomalous SSTs. Such a procedure requires that, for a given SST anomaly, a large sample of atmospheric states is available. However, the instrumental record is only a century long and within that record no two SST states are identical. This shortcoming can be circumvented using an ensemble of AGCM simulations where for an identical SST forcing, but starting from different initial conditions, a complete spectrum of seasonal mean atmospheric states can be sampled. For a perfect AGCM, the accuracy of the PDF estimated using an ensemble of AGCM simulations will depend on the size of the ensemble. Within the context of our analysis a hypothesized perfect AGCM is replaced by a sampling procedure such that the statistical properties of seasonal mean states in the ensemble thus obtained are the same as those for the collection of observed seasonal mean atmospheric states.

b. Analysis procedure

Given the PDF of seasonal mean atmospheric states for climatological SSTs, probabilistic forecasts can be made for a number of classes (or categories) that are chosen a priori. For example, the range of the atmospheric variable under consideration, x, can be divided into different classes such that the probability for the seasonal means of x to be in any particular class is equal. A common practice is to divide the climatological PDF into below, above, and normal categories such that the probability of seasonal means of x to fall within each category is one-third. For anomalous SSTs and from the knowledge of the corresponding PDFs, anomalous prob-

abilities relative to the climatological probability for different classes can then be found.

For an ensemble of AGCM integrations the probability of the seasonal mean of x to fall in each category can be easily computed by a simple counting procedure. These estimated probabilities then form the basis for the probabilistic seasonal forecast for the observed seasonal mean of x to be in different classes.

For an infinite ensemble size, the probabilities for x to be in various classes for different SST states can be precisely determined. Further for a fixed SST forcing, these probabilities also remain fixed. However, as pointed out by Kumar and Hoerling (2000), for a fixed SST forcing the observed seasonal mean atmospheric states can vary, and since they are only partially constrained by the boundary forcing, different realizations can fall in different categories. The inherent limit for the expected value of any skill measure, therefore, depends on the spread in the seasonal mean observed states.

For finite ensemble sizes, the estimated probabilities for x to be in different classes can vary from one ensemble realization to another. As a consequence, incorrect estimation of the class probabilities also factors into the accuracy of the prediction. Due to this, the expected forecast skill can be lower than its inherent limit.

A particular measure of the predictive skill for the probabilistic seasonal forecast is the rank probability skill score (Epstein 1969; Murphy and Daan 1985; Wilks 1995). The rank probability score (RPS), for a probabilistic forecast for n equiprobable forecast categories is defined as

$$\mathrm{RPS} = \sum_{m=1}^{n} (Y_m - O_m)^2, \tag{1}$$

where Y_m and O_m, respectively, are the predicted and observed cumulative probabilities for the category m and are defined as

$$Y_m = \sum_{j=1}^{m} y_j, \quad \text{and} \quad O_m = \sum_{j=1}^{m} o_j.$$

In the above expressions y_j and o_j, respectively, are the predicted and observed probabilities for the jth forecast category. For a particular realization, the observed probabilities for all the classes except the one in which the observed state falls, are zero. RPS is a measure of the squared distance between the forecast and the observed cumulative probabilities. A more comprehensive discussion about RPS can be found in Wilks (1995).

We can also define an RPS resulting from the use of the climatological probabilities as the forecast (hereafter denoted by RPS_c). RPS_c is found by replacing the cumulative forecast probabilities for each category in (1) by the corresponding cumulative climatological probabilities.

Equation (1) defines the RPS for the a single forecast-observed event. The expected value of RPS averaged over l such events, $\langle \mathrm{RPS} \rangle$, is defined as

$$\langle \mathrm{RPS} \rangle = \frac{1}{l} \sum_{i=1}^{l} \mathrm{RPS}_i, \tag{2}$$

where RPS_i is the rank probability score for the ith forecast–observed pair and is given by (1).

The RPS skill score, or RPSS, is next defined as the ratio of expected value of RPS relative to the expected value of RPS obtained using climatological probabilities as the forecast (i.e., $\langle \mathrm{RPS}_c \rangle$).

Therefore, RPSS, defined by

$$\mathrm{RPSS} = 1.0 - \frac{\langle \mathrm{RPS} \rangle}{\langle \mathrm{RPS}_c \rangle}, \tag{3}$$

is a measure of the percent change in the RPS over the RPS based on climatological probabilities as the forecast. A negative value of RPSS implies that the skill of estimated probabilities as the forecast is worse than the use of climatological probabilities as the forecast.

For different shifts in the first moment of the PDF, the variations in the expected value of the RPSS with the ensemble size are next studied. As pointed out earlier, for a particular value of μ and for infinite ensemble size, the predicted cumulative probabilities for different classes are unique and do not vary from one forecast to another. For finite ensemble size, however, the predicted cumulative probabilities can vary from one ensemble realization to another, and thus can also impact RPSS.

The variation in the expected value of the RPSS for different μ and for different ensemble sizes is studied using a Monte Carlo procedure. Following this technique, for each μ a time series of observed realizations with a sample size of l is first generated. The statistical properties of the sample of observed realizations follows the properties of the corresponding Gaussian PDF. That is, the first and second moments of the sample are given by μ and σ, respectively. For each observed realization an ensemble of forecast realizations is also generated based on a similar sampling procedure. The ensemble of realizations thus generated is statistically equivalent to an ensemble of seasonal mean states obtained using a perfect AGCM simulation. This is so since the statistical properties of x for seasonal means generated using a perfect AGCM will have statistical properties similar to the observed PDF.

The forecast and corresponding cumulative probabilities for each class are obtained from the ensemble realizations by a simple counting procedure. Knowing in which class the corresponding observed realization falls, from Eq. (1), the RPS for a particular prediction–observation pair is evaluated. RPS_c is also evaluated. The RPSS summed over all l pairs of observed realizations and the corresponding forecasts is then given by (3). To ensure statistical robustness of our results, the sample size of the observed time series is chosen to be 10^6.

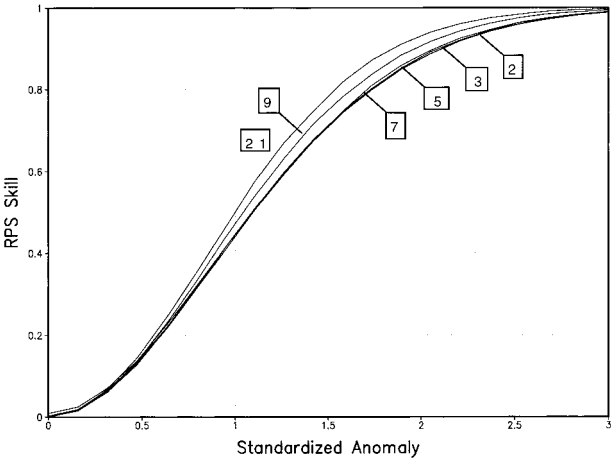

FIG. 2. The RPSS as a function of standardized values of the absolute value of the mean shift. Each curve corresponds to the RPSS for a different number of equiprobable classes, which is indicated by numerals. These scores are for the case of infinite ensemble size when the PDFs for the seasonal mean states can be exactly determined.

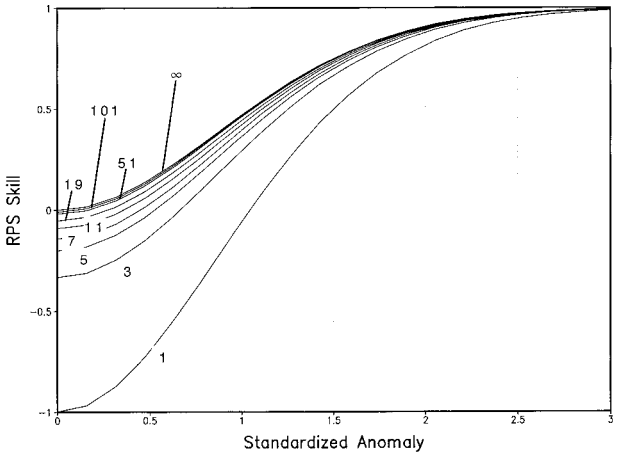

FIG. 3. The RPSS as a function of standardized values of the absolute value of the mean shift, for the case of three equiprobable categories. Each curve corresponds to the RPSS for a different ensemble size, as shown by the numerals in the plot. The highest curve represents the result for an infinite sized ensemble, and the other curves for sizes of 101, 51, 19, 11, 7, 5, 3, and 1.

3. Results

We start with an analysis of the variation in the expected value of the RPSS with different shifts in the PDF for the case of infinite ensemble size. This variation quantifies the RPSS over a range of signal-to-noise ratios of the seasonal mean atmospheric states and also represents the upper limit of the average seasonal predictive skill as measured in terms of RPSS. We also consider these variations for different numbers of classes of the predictand for which probabilistic forecasts are made.

The RPSS for different standardized values of mean shift is shown in Fig. 2. Different curves correspond to the RPSS for different numbers of equiprobable classes. It is apparent that the RPSS itself is not very sensitive to the number of forecast categories. For this reason we restrict subsequent discussion to the three category forecast system alone, since this system is widely used [e.g., forecasts from the Climate Prediction Center (Barnston et al. 2000) and from the International Research Institute (Mason et al. 1999)].

For small values of standardized shift, the RPSS for the forecast based on an infinite ensemble is close to zero, that is, the RPS of a prediction based on forecast probabilities derived from a large ensemble is the same as the RPS obtained from climatological probabilities as the forecast. This is to be expected since for an infinite ensemble and for a small signal-to-noise ratio, the forecast probabilities estimated from the infinite ensemble are approximately the same as the climatological probabilities for each class.

For large values of standardized shift RPSS approaches unity. For this case the RPS as defined by (1) itself approaches zero. This is so since while on the one hand the predicted cumulative probabilities for the below-normal, normal, and above-normal classes approach (0, 0, 1), on the other hand the cumulative probabilities for each observed event is also very likely to be (0, 0, 1), reflecting the fact that each observed seasonal mean state is almost certain to be in the above-normal category.

The RPSSs shown in Fig. 2 are the upper bound of the average seasonal predictive skill for probabilistic forecasts measured in terms of RPSS. This upper bound in RPSS can be contrasted with the upper bound of the average predictive skill measured in terms of the spatial anomaly correlation skill of the forecasts based on ensemble means as discussed in Kumar and Hoerling (2000). A comparison between the two skill scores indicates that RPSS tends to be smaller than the spatial correlation score.

We next study the impact of predictions based on finite ensemble size on the RPSS. As pointed out before, for the case of finite ensemble size, the predicted probabilities can vary from one realization to another and are no longer constrained to be the true probabilities as in the case of an infinite ensemble. This component of randomness in the forecast probabilities leads to a reduction in the RPSS. For different ensemble sizes, this reduction in RPSS for a 3-category forecast is illustrated in Fig. 3.

The topmost curve in Fig. 3 repeats the RPSS for the forecast probabilities obtained using an ensemble of infinite size and as pointed out before, is the upper bound for the average predictive skill. For each curve, the standardized shift at which RPSS becomes positive is the value below which the RPS using climatological probabilities is smaller (i.e., indicating higher skill) than the RPS based on probabilities estimated from the finite ensemble. The standardized shift at which this zero

crossing occurs decreases as the size of the ensemble increases.

From Fig. 3 it is apparent that with increasing ensemble size the RPSS rapidly converges to its upper bound. For ensemble size of 19 and for the signal-to-noise ratio larger than 0.5, the expected value of the RPSS is very close to its upper bound. For smaller signal-to-noise ratio the small value for the upper limit of RPSS can be computationally recovered using larger and larger ensemble sizes. The reality is that the probabilistic forecast for individual events is only marginally better than the use of the climatological probability as the forecast. For larger signal-to-noise ratios, for example, 1.5, probabilistic forecasts based on small ensemble sizes can efficiently recover the inherent predictability in the observed system.

The change in RPSS for finite ensemble size for different signal-to-noise ratios can be contrasted with the variation in the spatial correlation skill discussed in Kumar and Hoerling (2000). For small signal-to-noise ratio, the average correlation skill based on a finite and infinite ensemble equals the upper limit of predictability, which is close to zero. In contrast, RPSS based on a single AGCM realization is approximately -1, whereas the corresponding value for the RPSS based on infinite ensemble size is zero. This difference is because while on the one hand the RPSS is a quadratic measure, the spatial correlation, depending on the relative phase of the predicted and the observed anomaly, can either be positive or negative. This, for small signal-to-noise ratio, has an expected value of zero.

Negative RPSS for small ensemble sizes and for small signal-to-noise ratios indicates that when no useful forecast information is available but the forecast must be made, the better forecast would be the one that mimics the climatological PDF. This implies that in the 3-category system, equal probabilities should be forecast in the absence of SST forcing even if the ensemble distribution indicates otherwise, since the PDF based on finite ensembles can generate incorrect estimates for class probabilities. For small signal-to-noise ratios the forecast based on climatological probabilities, instead of having negative RPSS, will produce a zero RPSS over the long run.

4. Conclusions and discussion

We have presented analyses aimed at determining how large an ensemble size is required to attain average skills sufficiently close to the skill expected for an infinite ensemble size. In this paper we discussed this problem with respect to the rank probability skill score (RPSS) to complement the findings with respect to the spatial anomaly correlation skill discussed in Kumar and Hoerling (2000). The RPSS, a verification measure suited for probability forecasts, is sensitive to the distributional features of a forecast as well as its central tendency.

The results indicate that for the signal-to-noise ratios of 0.4–0.5, the typical ensemble sizes currently used for seasonal predictions (i.e., 10–20 members) are sufficient to ensure an average level of skill close to that expected with larger ensembles. To what anomalous tropical Pacific SSTs (for example SST anomalies in Niño 3.4 index region) this signal-to-noise ratio corresponds is next estimated. Given that the correlation between Niño 3.4 SSTs and different atmospheric variables in the regions of substantial teleconnections (e.g., boreal winter 500-mb heights and surface temperatures over south-central Canada, or precipitation in the southeastern United States) is approximately 0.6 (Horel and Wallace 1981), the tropical Pacific SST anomaly needs to be about 0.7 for the extratropical seasonal mean anomalies to be above the signal-to-noise ratio of 0.4–0.5 (i.e., 0.7×0.6). Further, given that the median absolute value of the standardized Niño 3.4 SST index over all cases, by statistical definition is 0.67, and the mean absolute value is 0.75, the above result implies that for mild-to-moderate ENSO episodes, the commonly used ensemble sizes for seasonal prediction are sufficient to capture the maximum possible expected value of predictive skill. For smaller signal-to-noise (or for weak SST anomalies), the size of ensemble required to capture the maximum possible skill increases dramatically. But for these cases, the expected value of maximum possible skill itself is low and the use of larger ensemble size has to judged against the marginal level of skill.

In the above analysis we have considered RPSS for a 3-category forecasting system alone. However, we have shown that RPSS is fairly insensitive to the number of equiprobable categories. Finally, our results are based on the assumption that PDFs of seasonal means are Gaussian. This assumption is probably adequate for broad forecast categories (e.g., three equiprobable categories) that do not focus on the distribution extremes. We should also point out that the present analysis only refers to the seasonal prediction problem where the characteristics of the PDF depend only upon the anomalous boundary forcing and are independent of atmospheric initial states.

Acknowledgments. The support offered by NOAA's Climate Dynamics and Experimental Prediction (CDEP) Program is gratefully acknowledged. We would also like to thank Drs. D. A. Unger, Wanqiu Wang, and two anonymous reviewers for their constructive and thoughtful comments.

REFERENCES

Barnston, A. G., Y. He, and D. A. Unger, 2000: A forecast product that maximizes utility for state-of-the-art seasonal climate prediction. *Bull. Amer. Meteor. Soc.,* **81,** 1271–1297.

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

Horel, J. D., and J. M. Wallace 1981: Planetary-scale atmospheric

phenomena associated with the Southern Oscillation. *Mon. Wea. Rev.,* **109,** 813–829.

Kumar, A., and M. P. Hoerling, 2000: Analysis of a conceptual model of seasonal climate variability and implications for seasonal prediction. *Bull. Amer. Meteor. Soc.,* **81,** 255–264.

Mason, S. J., L. M. Goddard, N. E. Graham, E. Yulaeva, L. Sun, and P. A. Arkin, 1999: The IRI seasonal climate prediction system and the 1997/98 El Niño event. *Bull. Amer. Meteor. Soc.,* **79,** 1853–1873.

Murphy, A. H., and H. Daan, 1985: Forecast evaluation. *Probability, Statistics, and Decision Making in the Atmospheric Sciences,* A. H. Murphy and R. W. Katz, Eds., Westview Press, 545 pp.

Trenberth, K. E., G. W. Branstator, D. Karoly, A. Kumar, N.-C. Lau, and C. F. Ropelewski, 1998: Progress during TOGA in understanding and modeling global teleconnections associated with tropical sea surface temperature. *J. Geophys. Res.,* **103,** 14 291–14 324.

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