On the Interpretation and Utility of Skill Information for Seasonal Climate Predictions

Arun Kumar

NOAA/Climate Prediction Center, Camp Springs, Maryland

(Manuscript received 19 August 2005, in final form 10 March 2006)

Abstract

In recent years, there has been a steady increase in the emphasis on routine seasonal climate predictions and their potential for enhancing societal benefits and mitigating losses related to climate extremes. It is also suggested by the users, as well as by the producers of climate predictions, that for informed decision making, real-time seasonal climate predictions should be accompanied by a corresponding level of skill estimated from a sequence of the past history of forecasts. In this paper it is discussed whether conveying skill information to the user community can indeed deliver the promised benefits or whether issues inherent in the estimates of seasonal prediction skill may still lead to potential misinterpretation of the information content associated with seasonal predictions. Based on the analysis of atmospheric general circulation model simulations, certain well-known, but often underappreciated, issues inherent in the estimates of seasonal prediction skill from the past performance of seasonal forecasts are highlighted. These include the following: 1) the stability of estimated skill depends on the length of the time series over which seasonal forecasts are verified, leading to scenarios where error bars on the estimated skill could be of the same magnitude as the skill itself; 2) a single estimate of skill obtained from the verification over a given forecast time series, because of variation in the signal-to-noise ratio from one year to another, is generally not representative of seasonal prediction skill conditional to sea surface temperature anomalies on a case-by-case basis. These issues raise questions on the interpretation, presentation, and utilization of skill information for seasonal prediction efforts and present opportunities for increased dialogue and the exploration of ways for their optimal utilization by decision makers.

1. Introduction

In recent years, there has been a steady increase in the emphasis on routine seasonal climate predictions and their potential for enhancing societal benefits and mitigating losses related to climate extremes. Informed decision making based on such predictions requires real-time seasonal forecasts to be accompanied by some measure of skill. To date, one measure of skill that is used is based on a past history of forecasts. The skill information is also deemed necessary for constructing consolidated forecasts whereby a consolidation technique combines different forecast tools relative to their past performance (e.g., Derome et al. 2001). Such necessity of skill information for seasonal predictions requires a careful appraisal of its influence on the forecast itself and its ultimate use in a decision making process.

In this note, specific examples are given to illustrate key issues associated with the estimates of seasonal climate prediction skill and the implications they may have for the interpretation of individual seasonal forecasts by the user. These include the following: 1) the stability of estimated skill depends on the length of the time series over which forecasts are verified, and therefore, skill information itself is associated with error bars that are seldom presented along with the skill information; 2) in a generally accepted predictor–predictand paradigm of seasonal climate prediction (e.g., Niño-3.4 SSTs predicting U.S. surface temperature), the skill is proportional (or conditional) to the strength of the Niño-3.4 SSTs. If this is the case, then each individual “predictor” event is associated with its own unique level of “conditional” skill, and information contained
in the skill obtained as verification over a time history of events may not be optimal on a case-by-case basis. Differences in the “average skill” versus conditional skill, in turn, could lead to assigning an erroneous level of confidence for individual seasonal forecasts.

The issues related to the estimation of skill are highlighted based on atmospheric general circulation model (AGCM) simulations that are described in section 2. Computational procedures for estimating skill are also described in section 2. In section 3, examples and the interpretation of issues pertaining to skill estimates are presented. A summary, discussions, and suggestions on possible alternatives for the presentation of skill information appear in section 4.

2. Data and analysis procedure

A discussion of the features inherent in the skill estimates is highlighted based on data from AGCM simulations. Use of model data facilitates our purpose; however, the results are not dependent on the model characteristics or the particular variable used. Indeed, similar results can be obtained analytically (e.g., estimation of confidence interval) or by using Monte Carlo techniques (Kumar and Hoerling 2000). An advantage of using AGCM data is that no a priori assumptions on the distribution of seasonal mean atmospheric states need to be made, and further, AGCM simulations also provide us with real examples illustrating issues inherent to skill estimates.

For the AGCM simulations, multiple realizations of seasonal means with identical SST forcing can be made, thus allowing one to separate seasonal atmospheric variability into a potentially predictable component forced by the SSTs and an unpredictable component internal to atmospheric dynamics. AGCM simulations in our analysis are from the previously operational Seasonal Forecast Model (SFM) at the National Centers for Environmental Prediction (NCEP). Details of this model and the prediction system are described by Kanamitsu et al. (2002a).

For SFM dynamical seasonal forecasts at NCEP, each month a set of seasonal hindcasts for the 1979–99 period was made. The purpose of the hindcast simulations was to generate climatologies from which the anomalies for the real-time seasonal forecast can be computed. An extensive history of hindcasts also serves the purpose that an assessment of AGCM’s past performance can be made. All the hindcasts were of approximately a 7-month duration and were forced with the observed SSTs (Reynolds and Smith 1994). An ensemble of hindcasts from 10 atmospheric initial conditions, but forced with the same observed evolution of SSTs, was made. For example, for each November from 1979 to 1999 (i.e., 21 yr), an ensemble of 10 AGCM simulations starting from 12-hourly apart atmospheric initial conditions between the first and fifth of the month was made, and hindcast simulations were made until the end of May. Atmospheric initial states were taken from the reanalysis-2 dataset (Kanamitsu et al. 2002b).

From the above set of hindcast simulations, for all DJFs from 1980 to 2000, 10 AGCM simulations from November initial conditions are available. December–February (DJF) seasonal means are approximately one month from the start of the simulations. Additional sets of 10 DJF simulations each, although with longer lead times, are also available from October, September, and August initial conditions. If all the simulations from different initial conditions are pooled, then for each DJF in the period of 1980–2000, an ensemble of 40 AGCM simulations is available. Furthermore, as the SFM was operational for more than 2 yr, and the hindcasts repeat each month, we also pool the hindcasts for DJFs over the full 2-yr period, resulting in an 80-member ensemble for each DJF between 1980 and 2000. We have analyzed and confirmed that the statistical characteristics of DJFs from different lead times are similar (see also Phelps et al. 2004; Peng and Kumar 2005; Chen 2004), and therefore AGCM simulations from different lead times can be pooled together.

To summarize, we have access to an 80-member ensemble for each DJF in the 1980–2000 period. A schematic illustrating DJF seasonal means is shown in Fig. 1. This dataset is used to estimate skill using a “perfect prog approach” and to illustrate issues inherent in the skill information; the anomaly correlation skill measure for the 200-mb height is used. In the perfect prog approach, out of many AGCM realizations, one randomly chosen realization is treated as the proxy for the “observations” while the remaining realizations are used to construct a model-based forecast. As both the “observed” and the “predicted” time series are from the same AGCM, both time series possess the same properties. Although this analysis relies on anomaly correlations as the skill metric, a similar analysis for any other measure of skill can be repeated. To illustrate well-known shortcomings in the estimated skill, two different kinds of anomaly correlation assessments are made, which are described below.

a. Skill assessment based on verification time series spanning many years (or unconditional skill estimates)

In this approach one randomly chosen AGCM-simulated time series from 1980 to 2000 is treated as a
To compute anomaly correlation, the observed and predicted anomalies are first defined. Let \( X_{\alpha}(x, y) \) represent the DJF seasonal mean for the ensemble member \( i \) and the year \( \alpha \) (see Fig. 1). For the sake of brevity, the longitudinal and latitudinal variation in \( X_{\alpha} \) is not included in notations any further. The anomalies for the DJF seasonal mean for a particular year \( \alpha \) are computed with respect to model climatology \( <X> \) obtained from the AGCM simulations for all the remaining years:

\[
\langle X_{\alpha} \rangle = \frac{1}{N(M - 1)} \sum_{j} \sum_{\beta \neq \alpha} X_{j\beta},
\]

where \( N = 80 \) is the number of DJF realizations for a particular year and \( M = 21 \) is the number of years in the time series. Seasonal mean anomalies \( X_{\alpha} \) are then defined as

\[
X_{\alpha} = X_{\alpha} - \langle X_{\alpha} \rangle.
\]

The proxy for the observed anomaly time series \( O'_{i\alpha} \) is any random choice of one of the model-simulated time series; for example, for \( i = 1 \)

\[
O'_{i\alpha} = X_{i\alpha}
\]

and is a randomly chosen column in Fig. 1. The corresponding predicted time series \( F'_{i\alpha} \) is defined as the ensemble mean of the remaining 79 realizations:

\[
F'_{i\alpha} = \frac{1}{(N - 1)} \sum_{j \neq i} X_{j\alpha}.
\]

The anomaly correlation between the observed and predicted time series is defined by

\[
AC_I = \frac{\sum F'_{i\alpha}O'_{i\alpha}}{\sigma_I^F \sigma_I^{O'}}.
\]

where \( \sigma \) is the standard deviation for the observed and predicted time series.

The spatial map of anomaly correlation using (5) is not unique, however, as the same procedure could be repeated 79 additional times providing a set of 79 different spatial maps of anomaly correlations. Because of finite verification time series and associated sampling variability, skill maps differ from one another. The variability in the anomaly correlation from one estimate to another is mainly due to the sampling variability in the observed time series in (3) and much less due to the predicted time series in (4), as the predicted time series is based on the ensemble mean of 79 DJF realizations with the same SST forcing, primarily reflecting the atmospheric response to the imposed SST anomalies.

From 80 different estimates for the anomaly correlation, a more robust estimate of “expected” anomaly correlation skill for the 21-yr time series is the average of 80 anomaly correlation maps, with error bars for individual anomaly correlation maps provided by the spread between different anomaly correlation estimates.

b. Skill assessment based on verification for a particular year (or conditional skill estimates)

For the same data one can also estimate anomaly correlation skill for each individual year (hereafter referred to as Method 2). Anomaly correlation estimates in this case are conditional to the imposed SST forcing.

---

1 If required, by random shuffling, additional samples of observed time series can be easily created (Graham and Mason 2005).
and depend on the atmospheric signal-to-noise ratio specific to the SST forcing (Kumar and Hoerling 2000; Sardeshmukh et al. 2000). Skill for specific years differs from the expected level of skill obtained using Method 1. For example, one would anticipate skill to be higher during years of strong SST anomalies and low when SST anomalies are near normal.

For a particular year \( \alpha = A \), the anomaly correlation is computed by pairing one model-simulated DJF realization as a proxy for the observations and the ensemble mean of the other 79 realizations as the predictions. If we choose the observed realization to be \( i \), then

\[
O_{iA} = X'_{iA}
\]

is the observed realization, the prediction for which is defined as the remaining 79 DJF simulations for the year \( \alpha = A \); that is,

\[
F_{iA} = \frac{1}{(N-1)} \sum_{j \neq i} X'_{jA}
\]

is the corresponding forecast. For different choices of \( i \) there are 80 \( (O_{iA}, F_{iA}) \) pairs of observed and predicted DJF anomalies. The conditional anomaly correlation is then given by

\[
AC_{iA} = \frac{\sum_{i} O_{iA}F_{iA}}{\sigma_{O}^{2A}\sigma_{F}^{2A}}.
\]

Following this approach, the expected value of anomaly correlation can be derived for every DJF in the 1980–2000 period and depends on the amplitude of the ensemble mean anomaly and the intraensemble variability, both of which can be influenced by the interannual SST variations.

3. Results

a. Analysis of skill estimates based on Method 1

Recall that the anomaly correlations based on Method 1 are estimates over a 21-yr time series, and from the data used in our analysis, 80 different estimates of anomaly correlations are made. Shown in Figs. 2a,b, respectively, is the mean anomaly correlation obtained from averaging 80 different estimates as well as the spread in their estimates.

The average value of the anomaly correlation skill is the largest in the tropical latitudes and decreases in the extratropical latitudes. This is a well-documented behavior known from a long history of predictability studies of seasonal atmospheric means (Madden 1976; Lau 1985, 1997; Chervin 1986; Kumar and Hoerling 1995; Barnett 1995; Harzallah and Sadourny 1995; Stern and Miyakoda 1995; Anderson and Stern 1996; Rowell 1998; Renshaw et al. 1998; Saravanan 1998; Trenberth et al. 1998; Wang and Zwiers 1999; Martineau et al. 1999; Zheng et al. 2000; Zwiers et al. 2000; Hunt 2000; Shukla et al. 2000; Peng et al. 2000; Kunosaki et al. 2001; Hoerling and Kumar 2002; Straus et al. 2003; Kumar et al. 2003; Palmer et al. 2004; among others). This feature is a consequence of the fact that the unpredictable internal variability of seasonal means tends to be low (high) in the tropical (extratropical) latitudes.

The spread among different estimates of anomaly correlation skill, on the other hand, is smallest in the tropical latitudes and increases in the extratropics. It is useful to reemphasize that variability in the anomaly correlation is almost entirely because of the sampling variability in the construct of the observed time series and not because of the sampling variability in the predicted time series that is based on the 79-member ensemble mean anomaly that varies little from one sample to another (not shown). As the internal variability of DJF seasonal means is small (large) in the tropical (extratropical) latitudes, constructs of observed time series also have small (large) sampling variations in the tropical (extratropical) latitudes. This leads to the latitudinal behavior in the spread of estimated anomaly correlation skill shown in Fig. 2b.

Latitudinal variations in the expected value of anomaly correlations in Fig. 2a and in the spread among different estimates of anomaly correlations have an inverse relationship; that is, large (small) values of expected anomaly correlation have small (large) spread. This behavior is further illustrated in Fig. 3 where a scatterplot between the expected value of anomaly correlation and the corresponding spread at each grid point is shown. This inverse relationship can be explained in different ways: 1) as the anomaly correlation skill has an upper bound of one, deviations above it are not allowed, leading to a smaller spread among different estimates when anomaly correlation is at its extremes; and 2) following a more physical reasoning, the expected value of anomaly correlation skill is high in the regions where SST-related atmospheric variability (often referred to as the external variance) is large compared to its internal variance. While a high (low) external-to-internal seasonal mean atmospheric variance ratio leads to a high (low) expected value of anomaly correlations, it also leads to a smaller (larger) spread because of sampling variations. A relationship similar to that shown in Fig. 3 between the expected value of the anomaly correlation and the uncertainty in its estimate can be obtained analytically. Indeed, Fig. 3
closely approximates the estimate of spread in the anomaly correlation for a bivariate normal distribution (Anderson 1984). One can also follow the Monte Carlo technique (Kumar and Hoerling 2000; Kumar et al. 2001b) to obtain the relationship shown in Fig. 3 as well as extend a similar analysis to other measures of skill.

The inverse relationship between the expected value of anomaly correlations and spread in such estimates due to sampling variations is unfortunate in the sense that in the extratropical latitudes where seasonal prediction skill is small, estimates based on short time series also have large errors. This is highlighted in Fig. 4 where two specific examples of anomaly correlation skill chosen out of 80 different estimates of anomaly correlation are shown. These two estimates are very different from each other and are also very different from the more robust estimate of anomaly correlation skill in Fig. 2a.

In the analysis based on the AGCM simulations one has the luxury to compute multiple estimates of anomaly correlation skill and also to compute sample
spread in such estimates; however, such is not the case when hindcast skill for a seasonal prediction method is verified against a single observed time series, as only a single map of skill estimate can be constructed. At various operational centers, however, it is a common practice to provide such skill maps based on verification over a past history of forecasts. Often times, such skill estimates are also used in the display of seasonal forecast (e.g., seasonal forecasts are only displayed when the past estimate of skill exceeds a predetermined threshold; 0.3 in many cases) or are used for combining forecasts from different prediction systems.

For a single estimate of skill information, it is hard to provide a corresponding estimate for the error bar associated with the skill estimate. Referring back to Fig. 4, a single estimate of skill could be any one of the maps, and by chance could be very different from the true expected value of skill in Fig. 2. Within such a framework, potentially large error bars on the estimated skill also exist. Furthermore, in the extratropical latitudes, error bars on the estimates of skill can be of the same magnitude. Although the spread in the skill shown in Fig. 2 (bottom panel) and illustrated further in Fig. 3 is specific for the AGCM used in this study, the fact remains that a large fraction of interannual variations in the extratropical seasonal atmospheric variability is not predictable (see the list of references cited above), and large errors in the estimate of skill for any seasonal prediction methodology do exist.

A straightforward approach for reducing the error bars for the estimated skill is to increase the length of the time series over which skill estimates are computed. However, even if this approach is implemented, another limitation of the skill information (illustrated in the next section) is harder to overcome.

b. Analysis of skill estimates based on Method 2

Estimates of anomaly correlation skill using Method 2 are based on the signal-to-noise considerations for each individual season. The case-by-case expected value of anomaly correlation skill differs from the expected value of anomaly correlation skill computed based on Method 1, which averages the information inherent in signal and noise across a mix of SST states. Examples of the expected value of anomaly correlation for two specific years obtained using Method 2 are shown in Fig. 5. One year (DJF 1993) corresponds to a weak tropical SST forcing (Fig. 5, top panel), while the other year (DJF 1998) corresponds to the strongest warm ENSO year in the 21-yr record (Fig. 5, bottom panel). As expected, the anomaly correlation skill for DJF 1993 is small because of weak tropical SST forcing, while the expected skill for DJF 1998 is large, consistent with large SST forcing. Further, the spatial patterns of anomaly correlation also differ from the expected anomaly correlation skill obtained using Method 1 (Fig. 2, top panel).

In light of the fact that the expected value of anomaly correlation changes from one year to another, another
issue inherent in the skill information obtained using Method 1 should now be apparent. For instance, when tropical SST forcing is large, the use of skill information based on Method 1 would lead to understatements about the expected skill. On the other hand, during years when the tropical SST forcing is weak, a similar use of skill information based on Method 1 would lead to undue confidence in the seasonal forecasts. Are users of climate information best served by the skill information estimated based on a history of forecasts? Or does the skill information based on the past history of forecasts need to be adjusted on a case-by-case basis?

4. Summary and discussion

Using atmospheric general circulation model simulations, issues inherent in the estimates of skill, and consequently, the usefulness of their information content
for seasonal prediction methodologies, are highlighted. The first is the stability (or the statistical significance) of skill estimates based on short time series leading to error bars associated with such estimates. In the extratropical latitudes with low seasonal predictive skill, error bars on the estimated skill could be of the same magnitude as the estimated skill itself. Even if error bars could be reduced from the use of longer and longer time series for verification, on a case-by-case basis, reliance on the skill information from past verifications would be limited.
could still lead to the erroneous use of such information.\(^2\)

Problems associated with skill information based on past performance are further exacerbated in the presence of nonstationary climate, or because of the low-frequency variations in the predictors. In other words, skill information obtained from the past evolutionary history of certain predictors may not be a useful indicator of the level of predictive skill that might occur in the future. Examples include low-frequency variability in the tropical ENSO variance and corresponding influence on the extratropical predictability (Nakaegawa et al. 2004), and the presence of tropical SST forcing that is not common in the historical record [e.g., the warming of SSTs in the Indian Ocean and warm pool region (Kumar et al. 2001b; Goddard et al. 2003; Fawcett et al. 2004)], leading to a spatial distribution of skill inconsistent with the expectations based on skill derived from past verifications.

A motivation for presenting the analysis described in this paper is to illustrate some well-known issues inherent in the skill information and the possibility that reliance on them can lead to suboptimal utilization of seasonal predictions. It is hoped that the analysis presented will lead to increased discussion on the requirement, as well as the interpretation of the information provided by the estimates of historical skill. Toward that end, the following exploratory discussions are also offered:

1) One suggestion is to first explore, and also question, the very need for skill information to accompany seasonal predictions. Because of the atmospheric internal variability inherent in the seasonal means, seasonal forecasts are traditionally cast in terms of probabilities. One could ask if providing estimates of skill conveys any information beyond what is already contained in the forecast probabilities. If predicted probabilities are reliable, a seasonal forecast implicitly contains its own “estimate of skill.” If, for example, at a geographical location, the seasonal mean surface temperature is predicted to be above normal with 60% probability and the forecast is reliable, in the long run approximately 60% of the observed anomalies should verify to be above normal (Hartmann et al. 2002). In other words, a reliable seasonal forecast also implicitly carries the “skill information,” and no further need for yet another estimate of skill obtained from past verification should be required. In this paradigm, the analysis of past skill is a tool for the forecast provider to achieve the goal of making reliable predictions, with information required by the user regarding the uncertainty in the forecast being contained in the forecast probabilities themselves. To justify a requirement for skill estimate, therefore, one needs to first establish if both predicted probabilities, as well as the estimate of skill, are indeed needed (or are currently being utilized) by different user applications. To decide this issue, instances of user applications that indeed utilize forecast uncertainty inherent in the forecast probabilities, as well as utilize the skill estimated from the past performance, should provide useful case studies.

2) One obvious suggestion would be to provide error bars alongside the skill information when providing seasonal forecasts to the user community. This needs to be true for whatever measure(s) is (are) chosen to convey the skill information. Providing an estimate for the error bars, however, requires a careful analysis. The question is not whether a numerical value of the skill is statistically different from zero but what the probabilities are that a given numerical value of skill corresponds to different possible values of “expected” skill.

3) Even though the uncertainty factor in skill estimates can be reduced by verifications over a longer and longer history of forecast, it is inconceivable for the forecast history to be long enough that conditional estimates of skill for individual prediction scenarios could be developed. As an alternative, a suggestion would be to explore scenarios under which a single skill estimate derived from the past forecast history can itself be used on a case-by-case basis. For example, if the atmospheric response to SSTs is primarily in the shift of the probability density function of the seasonal mean atmospheric states, and not in its spread, a single skill estimate can be adjusted according to the amplitude of the predicted anomaly. However, the extent to which interannual variability in SSTs influences the interannual variability of the atmospheric internal variability remains a contentious and unresolved problem [for a discussion, see Peng and Kumar (2005)]. However, in evidence to the contrary, recalling a single skill estimate based on the predicted amplitude of the seasonal anomaly may be a viable approach. In the meantime, based on AGCM simulations, it will also be useful to document how much the spatial pattern

---

\(^2\) Although verifications over longer time series result in a more representative estimate of the skill, the relative gain in accuracy quickly starts to saturate, and verification over a larger and larger time series has diminishing returns. This is analogous to ensemble size and errors in the estimate of correlation skill whereby larger and larger ensembles result only in incremental gains in the correlation skill [see Fig. 3 of Kumar and Hoerling (2000)].
of the “conditional” skill varies on a case-by-case basis relative to a single spatial pattern of “unconditional” skill. By doing so, one would at least know the magnitude of error that might be incurred because of the use of a single spatial pattern of skill.

4) From the user perspective, it might also be useful to let users be aware of the fact that the skill of the seasonal forecast can originate from different sources; for example, it could be due to ENSO or could be due to long-term trends, and further, such skill sources can vary from case to case. By providing this information and by educating the users, it is likely that a better understanding and awareness of the skill estimates can be developed. This will help users interpret limits of information contained in a single skill estimate and move them away from undue reliance on the use of single skill estimates in the development of relevant applications.

In conclusion, the question of skill information and its utility does not relate to the assessment of forecast skill for a particular seasonal prediction methodology, an activity that is useful in its own regard. Rather, it is questioning the use of skill information to convey uncertainty in seasonal predictions and understanding how (and how effectively) this information is utilized by the users in their decision making process. The questions are particularly relevant in light of the problems associated with estimating error bars for the skill estimates in the weak predictive skill regimes and the fact that a single skill estimate may not be appropriate on a case-by-case basis.

Acknowledgments. The support offered by the NOAA Office of Global Programs CDEP program is gratefully acknowledged. Constructive and careful comments by Drs. Amy Snover, M. P. Hoerling, W. Wang, and P. Peng, and two anonymous reviewers were very helpful in preparing the final version of the manuscript.

REFERENCES


