Performance of Recent Multimodel ENSO Forecasts

MICHAEL K. TIPPETT,* ANTHONY G. BARNSTON, AND SHUHUA LI

International Research Institute for Climate and Society, Palisades, New York

(Manuscript received 2 May 2011, in final form 28 September 2011)

ABSTRACT

The performance of the International Research Institute for Climate and Society “ENSO forecast plume” during the 2002–11 period is evaluated using deterministic and probabilistic verification measures. The plume includes multiple model forecasts of the Niño-3.4 index for nine overlapping 3-month periods beginning the month following the latest observations. Skills decrease with increasing lead time and are highest for forecasts made after the northern spring predictability barrier for target seasons occurring prior to the forthcoming such barrier. Forecasts are found to verify systematically better against observations occurring earlier than the intended forecast targets, an effect that becomes greater with increasing lead time. During the study period, the mean forecasts of dynamical models appear to slightly (and statistically insignificantly) outperform those of statistical models, representing a subtle shift from earlier studies. The mean forecasts of dynamical models have overall larger anomalies but similar errors to those of statistical models. Intermodel spread is related to forecast error in an average sense with changes in forecast error due to changes in lead and verification season being properly reflected in changes in spread. The intermodel spread underestimates the forecast error variance, to a greater extent for statistical forecasts than for dynamical ones. Year-to-year changes in plume spread provide little additional information relative to climatological ones.

1. Introduction

The purpose of this paper is to document the performance of a multimodel real-time ENSO forecast product over the period 2002–11. Since February of 2002, a number of groups have provided their ENSO forecasts to the International Research Institute for Climate and Society (IRI). Those forecasts are the basis for probabilistic ENSO category forecasts as well as an “ENSO prediction plume” like the one shown in Fig. 1 from February 2011. Here we limit our analysis to the ENSO prediction plume. Issued each month, the ENSO plume shows forecasts of seasonal (3-month average) Niño-3.4 anomalies for the upcoming nine overlapping 3-month seasons from a number of forecasts models; the number of models represented in the plume ranges from 15 at its inception to 23 at present. Forecasts are identified by production center and by whether the model is dynamical or statistical. The dynamical models range from fully coupled ocean–atmosphere GCMs to less complex Pacific-only coupled models.

The most common questions asked by users of this product are ones regarding its skill. Such questions are difficult to answer unequivocally on the basis of hindcasts alone for at least two reasons. First, not all of the models included in the prediction plume have extensive hindcasts. Second, hindcast skill may not be representative of real-time skill. The difference between hindcast and real-time skill may be due to difficulties that occur in real time (e.g., human error, data availability) but not in hindcast, as well as to model development choices that enhance skill only during the hindcast period (DelSole and Shukla 2009). Since real-time forecasts are the ones that users must use, it is valuable to evaluate their performance directly.

Specific questions about the ENSO forecast plume that are addressed here include, what is the skill of the multimodel mean? How does skill vary according to lead and season? How is the spread of the prediction plume related to forecast error? How do statistical and dynamical
models compare? Since the sample size is small, just 9 yr of forecasts, we limit our questions to these fairly broad ones. Addressing more detailed issues such as the merits of one model relative to another, or the conditional biases of a model, would require significantly longer records.

The paper is organized as follows: section 2 describes the models, observations, and skill scores used. Results are presented in section 3. Section 4 provides a summary.

2. Data and methods

a. Data

The ENSO forecasts and corresponding observations used here are 3-month averages of the Niño-3.4 index (Barnston et al. 1997). The first ENSO plume forecasts were issued in February 2002 and continue to the present; the latest forecast considered in this study was issued in March 2011. Forecasts from each start time extend to lead times covering nine running 3-month periods; the first forecast plume covers forecasts ranging from February–April (FMA) 2002 to September–November (SON) 2002 and is based on observations through the end of January 2002. The set of models contributing forecasts to the forecast plume has changed somewhat over time because of the introduction of new models, model replacement, and model discontinuation. Table 1 lists the contributing models and developing institutions, and separates the models into dynamical and statistical types.

b. Skill scores

The squared error skill score (SESS) is

\[ \text{SESS} = 1 - \frac{(f - o)^2}{o^2}, \]

where \( f \) and \( o \) are the forecast and observation (expressed as anomalies with respect to the climatology), respectively. The SESS is 1 if the forecast is equal to the observation—a perfect forecast. The SESS is 0 if the forecast error is equal in magnitude to that of a climatological forecast. The SESS is negative for forecasts whose error is greater than that of a climatological forecast. The SESS can be considered a one-sample estimate of the mean-squared error skill score (MSESS) given by
where the angle brackets denote the average over forecasts. In the absence of biases, MSESS is the explained variance.

The other deterministic skill measures used are anomaly correlation and Kendall’s tau (Kendall 1938). The skill of probabilistic categorical forecasts is measured using the ranked probability skill score (RPSS; Epstein 1969) and reliability diagrams (Murphy 1973).

c. Significance testing

The null hypothesis used for significance testing is that each model forecast is an independent sample from the climatological distribution. Samples from the climatological distribution were generated assuming that the seasonally stratified climatological distribution of Niño-3.4 is Gaussian. For each season, the mean and variance of Niño-3.4 was estimated from the Kaplan EXTENDED dataset using anomalies from the period 1950–2001. The null hypothesis that the n-model forecasts are independent samples from the climatological distribution implies that its mean has variance reduced by the factor $1/n$.

To evaluate the significance level for a verification score, we generate the same number of “climatological” forecasts as actual forecasts, score the climatological forecasts, and repeat the process 10 000 times. The resulting distribution of scores is the distribution of the score under the null hypothesis. Forecast skill is deemed significant at a specified level when it exceeds (for positively oriented skill measures) the corresponding percentile of the null hypothesis distribution. In this paper we use the 95% significance level.

3. Results

a. General features: Phasing and bias

The observed Niño-3.4 SST index anomaly and the multimodel mean anomaly forecast at lead times of 1, 3, and 9 months are shown in Fig. 2 separately for all models, dynamical models only, and statistical models only. Lead time is defined as the number of months between the latest observations and the beginning of the middle month of the 3-month period being forecast. For example, a forecast for January–March (JFM) using data through December is defined as a 1-month lead. Shading indicates El Niño and La Niña conditions. Here El Niño and La Niña conditions are defined according to the IRI-defined thresholds as shown in Table 2, with no requirements for duration. These thresholds are designed to
classify El Niño (La Niña) conditions as those anomalies that fall in the seasonally defined upper (lower) quartile and are also used in the IRI probabilistic ENSO category forecast product. These definitions differ from those used in the National Oceanic and Atmospheric Administration (NOAA)/Climate Prediction Center (CPC) oceanic Niño index (ONI).

The observations indicate moderate El Niños in 2002/03 and 2009/10, weaker El Niños in 2004/05 and 2006/07, moderate La Niñas in 2007/08 and 2010/11, and weak La Niñas in 2005/06 and 2008/09. Brief warm and cold episodes observed in the early parts of 2005, 2006, and 2009 were considered neutral according to the ONI. The multimodel mean forecasts capture the observed anomalies, but with lower amplitude and somewhat delayed with respect to the observations, both effects increasing with increasing lead time. In particular, the longest lead forecasts are “slow” to capture the transition into and out of ENSO events. In this short period, there is the indication that the late transition from active ENSO conditions to

**TABLE 2.** La Niña and El Niño thresholds. ASO indicates August–October, and OND is for October–December.

<table>
<thead>
<tr>
<th></th>
<th>DJF</th>
<th>JFM</th>
<th>FMA</th>
<th>MAM</th>
<th>AMJ</th>
<th>MJJ</th>
<th>JJA</th>
<th>JAS</th>
<th>ASO</th>
<th>SON</th>
<th>OND</th>
<th>NDJ</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.65</td>
<td>0.50</td>
<td>0.40</td>
<td>0.40</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.45</td>
<td>0.50</td>
<td>0.75</td>
<td>0.75</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>−0.65</td>
<td>−0.55</td>
<td>−0.45</td>
<td>−0.45</td>
<td>−0.50</td>
<td>−0.50</td>
<td>−0.50</td>
<td>−0.55</td>
<td>−0.75</td>
<td>−0.75</td>
<td>−0.75</td>
<td>−0.70</td>
</tr>
</tbody>
</table>
neutral conditions is somewhat more pronounced during warm events than during cold events. The lagged feature raises the question of the extent to which the models are a persistence of the latest observations.

The degree to which the forecasts lag observations is seen more clearly in Fig. 3 where the observed and forecast anomalies at the same verification time are plotted as colors, and the horizontal and vertical axes are verification time and lead, respectively. Ideally, if forecast and observed anomalies occur at the same verification time, there would not be the systematic “tilt” seen in Fig. 3. The positive slope observed in Fig. 3 indicates that forecast anomalies correspond best with observed anomalies not at the verification time but rather at a time before the verification time; in particular, longer lead-forecast-anomaly peaks correspond best with observations at an earlier time (Fig. 2). This lag between forecast and observation is an increasing function of lead time. If the persisting initial conditions were shown in the same format as the forecasts, the tilt would have slope one, and

![Fig. 3. Observed and forecast anomalies: (a) all, (b) dynamical, and (c) statistical models. The horizontal axis is verification time and the vertical axis is lead time with observed anomalies occupying the bottom row. Vertical lines separate years and the slope of the diagonal lines is 1.](image-url)
lines with slope one are shown for reference. The tilt of the forecast anomalies is seen to be somewhat greater than one, especially at the longest leads. This point is made more precise in Fig. 4a, which shows lagged correlations of forecasts and observations. Maximum correlations occur for negative values of the lag somewhat less than the lead (i.e., for the lead-3 and lead-5 forecasts the maximum correlation occurs with a lag of $-2$ and $-3$ months, respectively). Some temporal offset is likely due to the initial delay associated with data retrieval and forecast production and issuance: the retrieval of current observed data is often archived at a 1–2-week delay, and production time uses another 1–2 weeks. Analysis (not shown) of individual models indicates that most models suffer from this problem to a lesser or greater degree in the study period, with the European Centre for Medium-Range Weather Forecasts (ECMWF) model being one of the least affected.

The lag between forecasts and observations also manifests itself as a conditional bias where forecasts starting during warm events (Northern Hemisphere winter) have a warm bias and forecasts starting before warm events (Northern Hemisphere spring–summer) have a cold bias. The opposite biases are seen in forecasts before and during cold events. To the extent that this conditional bias is symmetric, one would expect the associated unconditional bias to be zero. Averaging the forecasts of many models is expected to allow various unconditional and conditional model biases to at least partially cancel one another out in the multimodel mean prediction. However, in the period considered, slightly more warm events than cool events occurred (observed mean anomaly is $0.11^\circ$), and as seen in Fig. 5, both dynamical and statistical models exhibit significant negative forecast bias for forecasts made in spring–summer and significant positive bias for forecasts made in winter, consistent with the conditional bias associated with warm events. Significant biases are those that are greater in magnitude than 95% of the climatological forecasts.

This systematic discrepancy between forecasts and observations suggests the potential benefit of applying a statistical postprocessing to the model output. Typically, statistical postprocessing of SST forecasts involves forecasts with a single lead time, verifying at the same time as the observations (Tippett et al. 2005). Conceptually, Fig. 4a suggests that, for instance, lead-3 forecasts could be used to form lead-1 forecasts, lead-5 to form lead-2 forecasts, and so on. More generally, a multiple linear regression could be developed between the verification values of Niño-3.4 and the forecast values, that is, having the form

$$\hat{o}_i(k) = \sum_{j=1}^{9} a_{ij} f_j(k), \quad i = 1, 2, \ldots, 9,$$

where $\hat{o}_i(k)$ is the corrected estimate of the lead-$i$ forecast with start time $t_k$, $f_j(k)$ is the lead-$j$ forecast with start time $t_k$, and $a_{ij}$ are the regression coefficients. This form allows the corrected forecast for a given start and lead time to depend on all the leads for that start. Direct estimation of the regression coefficients is difficult given the number of predictors, sample size, and colinearity of the predictors. We use principal component analysis (PCA) to reduce the number of predictors and deal with the problem of colinearity, retaining as predictors the first two PCs, which explain 99% of the variance. We expect the coefficients to have an annual cycle dependence, and to include it without unduly reducing sample size, we use a local linear regression, which has the effect of making the regression coefficients depend on start time $t_k$ (Hastie et al. 2009, p. 194). Specifically, the data at time $t_{k'}$ used to estimate $a_{ij}(k)$ are given weight:

$$W_{kk'} = (1 - r_{kk'}^3)^3, \quad t_{kk'} = \frac{1}{2}\{1 - \cos[2\pi(k - k')/12]\}.$$

The weights, a periodic generalization of the standard tricube weighting [Hastie et al. 2009, their Eq. (6.6)], are nonnegative and vary from 1 when $t_{k'}$ is the same season as $t_k$ to 0 when $t_{k'}$ and $t_k$ correspond to seasons separated by 6 months. The cross-validation procedure leaves the
6 months before and after the target date out of the training. The lag correlation of the statistically corrected forecasts is shown in Fig. 4b. The maximum correlations of the statistically corrected forecasts occur much closer to the zero lag. The zero-lag correlation (skill) of the corrected forecast is essentially the same as that of the uncorrected forecast for the one and nine leads and slightly higher for intermediate leads. This improvement in skill is particularly noticeable (not shown) for forecasts verifying in the seasons that extend from April–June (AMJ) through July–September (JAS), when the initial growth of new ENSO events is often underpredicted. This statistical postprocessing is not used in the results that follow.

**b. Deterministic skill**

Many skill measures are available to measure the level of association between forecasts and observations. The SESS (and likewise its average, the MSESS) is perhaps the simplest and most stringent since it is based simply on the mean-squared error between forecast and observation without any adjustment for systematic deficiencies such as errors in the mean or amplitude. Unlike skill scores such as correlation, which requires averaging over forecasts, the SESS allows the skill of individual forecasts to be compared with that of a climatological forecast. Figure 6 shows results of analysis of predictive skill for all forecast target periods for each of the 9 lead times
based on SESS. Blank entries in Fig. 6 indicate when SESS is statistically insignificant at the 95% level; Fig. 6d shows values of SESS required for 95% significance, which vary according to the observed anomaly, being closest to 1 (highest) during times of weak anomaly. Also shown are the squared values of the Niño-3.4 index during the period scaled to lie between 0 and 1. The results indicate significant model skill during times of observed deviation from neutral ENSO conditions. This is expected in light of the standard for significance requiring forecasts that differ from random realizations of the climatological distribution, which hover most densely near the mean state, and because here we are considering the multimodel mean forecast without regard to the intermodel spread. All of the ENSO events during the 2002–11 period were forecast with significant skill by both dynamical and statistical models, with the exception of the 2005/06 “surprise” weak La Niña, which was predicted only at a short lead time when it was dissipating in early 2006. It is noteworthy that as a group, the dynamical
models slightly outperformed the statistical models during the study period, due largely, but not only, to performance in predicting the 2007/08 La Niña. The dynamical models often show higher skill than statistical models at longer lead times. Early in 2007, the subsurface heat content became negative but Niño-3.4 SST did not cool to La Niña levels until late northern summer. Many of the dynamical models were thought to be inordinately quick in predicting negative SST anomalies earlier than the time of the average predictability barrier (near May) in response to the negative subsurface anomalies. They did in fact predict the onset too early (Figs. 2, 3) but were closer to what took place than the statistical models that waited, on average, until the later portion of the predictability barrier before beginning to predict below-normal SSTs.

Skill measures by lead time and season, averaged over the study period, are considered next with MESSS (shown in Fig. 7). The multimodel mean forecasts from December–February (DJF) through March–May (MAM) have significant skill for all nine leads. In the case of the dynamical multimodel means, this period of high skill is extended to include November–January (NDJ). On the other hand, the statistical multimodel mean forecasts of JFM–MAM
have significant skill for the first eight leads. The spring predictability barrier is manifest with forecasts of (May–July) MJJ–NDJ having no significant skill if initiated prior to April. For the dynamical multimodel mean, the earliest starting date with skill for the period beginning with JAS is March, and for the statistical multimodel mean it is June. ENSO conditions during the MJJ and June–August (JJA) seasons are the most difficult to predict with antecedence with forecasts of MJJ having no MSESS skill at any lead. The overall skill of the all-model multimodel mean, as measured by the average of the significant values of MSESS, is slightly less than that of the dynamical multimodel mean.

Skill measures by lead time and verification season, averaged over the study period, are considered next using the correlation shown in Fig. 8. Unlike MSESS, correlation is insensitive to calibration issues, and credit is awarded for systematic linear relationships between the forecast and the observation. In the case of correlation, values of 0.6 or greater are required for significance, given the sample size. The general distribution of correlation skill is similar to that of MSESS. However, correlation awards significant skill to forecasts of AMJ at all nine leads while MSESS does so only at the first three leads. An explanation for this difference is that these forecasts coincide with a period of significant bias.
(Fig. 5) as discussed before, and relatively low variability as reflected in the ENSO thresholds in Table 2. Since we believe that this unconditional bias is the small sample manifestation of a conditional bias, this diminished MSESS may not be observed in a longer period, and the two skill measures would more closely match. Although correlation, unlike MSESS, allows systematic bias, the correlation skill for forecasts of winter ENSO conditions has significant skill for about one lead less than MSESS. The overall correlation skill of the all-model multimodel mean, as measured by its average over all starts and leads, is 0.67, slightly less than that of the dynamical multimodel mean, which is 0.71.

Kendall’s tau, shown in Fig. 9, is invariant to monotonic transformations of the data and is therefore more insensitive to a wider range of forecast bias types than is correlation. The distribution by verification and lead of significant values of Kendall’s tau is similar to correlation. However, Kendall’s tau seems to suffer more from sampling variability, likely due to its nonparametric definition, and exhibits behavior such as nonmonotonic decay of skill with lead times.

The prominent feature of all three deterministic skill measures is the effect of the northern spring predictability barrier. Skills are greatest near the beginning and end of the calendar year when most ENSO episodes are well-established and mature, having evolved since the middle of the first calendar year, and are lowest during transition times when episodes often dissipate or begin. SST conditions during the MJJ and JJA seasons are the
most difficult to predict with antecedence (e.g., Jin et al. 2008), with forecasts of MJJ having no significant MSESS skill at any lead. We note also that the performance of dynamical models is superior to that of statistical models during these two seasons, though the difference is unlikely to be statistically significant given the small sample size. The statistical model mean has no skill in predicting these two seasons at any lead by any skill measure, while the dynamical model mean has skill up to lead 3, depending on the skill measure. A consequence of the difficulty in predicting northern spring SST is that forecasts passing through this period have little skill.

c. Probabilistic skill

Thus far, our forecast verification has focused on deterministic skill measures applied to the multimodel ensemble mean of the ENSO plume. However, by its very nature, the ENSO plume emphasizes the uncertainty of ENSO predictions, and we now consider the skill of probabilistic forecasts based on the plume. We emphasize that the plume-based probabilistic forecasts presented and analyzed here are not the same as the probabilistic IRI ENSO forecast, which is informed by expert opinion as well as model outputs. However, the plume-based probability forecasts considered here may well correspond to how users may interpret and act upon the information in the plume forecast.

Two types of probabilistic forecasts are possible: categorical and continuous. Here we consider categorical forecasts where probabilities are issued for the occurrence of La Niña, neutral, or El Niño conditions. We define the occurrence of these categorical events using the IRI thresholds given in Table 2. The most direct way of computing ENSO probabilities from the prediction plume is to use the relative frequency obtained by counting the number of models in each category. For this analysis, we group dynamical and statistical models together. Figure 10 shows the dominant forecast probabilities for La Niña and El Niño conditions when the dominant forecast probabilities are for one of those nonneutral categories; the horizontal and vertical axes are verification and lead time, respectively. The ENSO events of the period are forecast for the most part with increasing probability as lead time decreases. Longer lead forecasts fail to capture the initiation and termination of events and show the same phasing problem that was noted with the deterministic forecasts.

Like the SESS, the RPSS can be computed for individual forecasts, and Fig. 11a shows the RPSS of individual forecasts as a function of verification time and lead. Many of the features are similar to those of SESS in Fig. 6, with RPSS generally becoming insignificant at shorter lead times than does SESS. Some differences are that RPSS shows more skill in 2004 than does SESS and the reverse appears in 2005 with SESS showing more skill than RPSS. Figure 11b shows the RPSS of individual forecasts where the categorical probabilities are computed using a Gaussian distribution with mean given by the ensemble mean and variance that is a function of the verification season and lead only, not of the year; this variance is computed by averaging the multimodel ensemble variance with respect to year. We expect the Gaussian model to suffer less from sampling error (Tippett et al. 2007). The skills of the two methods are similar. This suggests that for the purpose of categorical probabilities, informative changes in plume spread are primarily due to seasonal and lead-based differences rather than year-to-year differences for a given lead and season. We explore this point in more detail in the next subsection.
Figure 12 shows the RPSS as a function of verification season and lead. For the most part, the distribution of skill by verification season and lead is similar to that seen for the deterministic skill measures. One notable difference is the lack of skill in predicting the FMA ENSO category. Closer inspection indicates that the lack of FMA RPSS skill is due in part to the discrete nature of the categories. In FMA of 2005 and 2007, the forecast probabilities pointed toward warm conditions that did occur but failed to meet the El Niño threshold. These two forecasts are primarily responsible for the lack of skill in FMA since in other years the forecasts were "good." These two years are also responsible for the lack of skill of the ensemble frequency for MAM at leads 2 and 3 seen in Fig. 12a. Figure 12b shows the RPSS as a function of verification season and lead where the categorical probabilities are computed using as forecast distribution the Gaussian distribution as described previously. Again skill levels are comparable.
Reliability diagrams are shown in Fig. 13 for forecast leads 1–3. To deal with the small sample size, bins with a probability width of 0.2 are used. Error bars are calculated using a binomial approximation. Figures 13a,b show the reliability of the categorical probabilities from the plume frequency and the Gaussian approximation, respectively. They are similar, though the average RPSS of the Gaussian probabilities is slightly higher. In an average sense the probabilities are fairly reliable, though the sampling variability, as quantified by the error bars, is large.

d. The relation between spread and uncertainty

An important question related to reliability is the extent to which spread in the forecast plume reflects actual forecast uncertainty. Strictly speaking, the forecast plume is not designed to represent forecast uncertainty in the usual manner that ensemble forecasts from a single model...
do. In that case, ensemble spread is due to using a variety of initial conditions. Here, the individual dynamical model forecasts are already ensemble averages, and statistical models also represent the distribution mean. Thus, the uncertainty expressed in the plume is primarily due to model differences, although individual models also use different initial conditions. We have already noted that replacing start-dependent ensemble variance with Gaussian distributions whose variance is a function only of the starting calendar month and lead results in only modest changes in probabilistic skill. Therefore, we focus our attention on similarly averaged quantities in our consideration of ensemble spread and forecast error.

The climatological values of the plume variance and the squared error between the plume mean and observation as a function of verification season and lead are shown in Figs. 14a,b, respectively. The squared error is a reflection of both the seasonally varying variance and seasonally varying predictability, and the seasonal variation in squared error does not coincide with that of predictability as measured by MSESS or correlation. Ideally the plume spread would capture the uncertainty of the forecasts, and the average squared error should be equal to the average plume variance (Johnson and Bowler 2009). The dependence of plume variance on verification season and lead does capture much of the structure of the squared error; the correlation between the climatological values of the plume variance and the squared error is 0.92. However, the magnitude of the plume variance is generally too small relative to the squared error, especially at the longest leads, as noted by the different color scales. Figures 14c,d show scatterplots of the dynamical and statistical plume variances, respectively, against the climatological squared error along with least squares line fits; the slopes of the least squares line fits are 0.42 and 0.25, and the intercepts are 0.11 and 0.072, respectively. As was noted with the variance of the complete plume, there is a high degree of association between the squared error and both the dynamical and statistical plume variances; the correlations are 0.90 and 0.93, respectively. However, the values of the regression coefficients confirm that, in an overall sense, the dynamical plume variance, and to a greater extent the statistical plume variance, are smaller than the average squared error.

e. Comparative properties of dynamical and statistical forecasts

Past ENSO prediction model verification studies found approximately equivalent skills for dynamical and statistical models, with slightly (and statistically insignificantly) higher skills for statistical models (Barnston et al. 1999). The results of this study are different in that during the period studied, the dynamical models slightly (and again statistically insignificantly) outperform the statistical models, as indicated above by some of the verification measures (e.g., MSESS, correlation, and Kendall’s tau). The difference between dynamical and statistical skill is seen most clearly during the part of the ENSO cycle most difficult to forecast—the period during and immediately following the northern spring predictability barrier. However, these differences are not statistically significant, and may well be specific to this particular period. Analysis (not shown) of individual model performance suggests that the higher skill of dynamical models during this period is coherent across models.

A difference between the two model types that may be related to the skill difference is the use of subsurface sea temperature by all of the dynamical models in their initial conditions, but by only about half of the statistical models. The most common reasons for omitting subsurface predictors is the brevity of the period of observations that began in 1980 and the use of an ocean-model-based analysis system. Omission of predictor information known to be integral to the dynamics of ocean physics, and thus to SST, may impose limits on the skills of the statistical models in question and consequently the mean skill of the statistical models. However, examination of individual statistical models (not shown) reveals that inclusion of subsurface predictors does not consistently give superior skill.

The relationship between statistical and dynamical multimodel mean forecasts is shown in Fig. 15a in the form of a scatterplot. Statistical and dynamical forecasts are seen to be strongly correlated—the linear correlation is 0.89. However, a least squares linear fit between statistical and dynamical multimodel mean forecasts has a slope of 0.75 and intercept of 0.02, indicating that on average statistical forecasts are somewhat weaker in amplitude. The difference between the coefficient and 1 is significant; in testing for significance, we reduce the nominal number of degrees of freedom (number of forecasts) by the factor of 36, the square of the roughly 6-month decorrelation time noted in the lag correlations. The lesser amplitude of the statistical forecasts is possibly related to the lesser variance of the statistical model plume. A scatterplot between statistical multimodel mean forecast errors and dynamical multimodel mean forecast errors is shown in Fig. 15b and shows a high degree of association; the correlation is 0.9. The slope of the linear fit between statistical multimodel mean forecast errors and dynamical multimodel mean forecast errors is 1.06, indicating that overall statistical errors are only modestly larger than dynamical ones on average. The 95% confidence intervals for the coefficient include 1, so we cannot conclude that dynamical and statistical forecasts have unconditionally differing skills. However, there is some
indication in Fig. 15b of a greater difference between dynamical and statistical models for cold conditions.

4. Summary

The performance of the IRI “ENSO forecast plume” is evaluated during the 2002–11 period. The plume presents forecasts of the Niño-3.4 index for the nine subsequent overlapping 3-month periods from a number of dynamical and statistical models of varying complexity. Deterministic skill measures were used to measure the association between the multimodel mean of the plume and corresponding observations. The ENSO forecast plume by its nature lends itself to a probabilistic interpretation, and we construct categorical probabilities for the occurrence of El Niño, La Niña, and neutral

Fig. 14. Climatology of the (a) plume variance and (b) squared error, each as a function of verification season (horizontal axis) and lead time (vertical axis). Scatterplots of (c) dynamical and (d) statistical plume variance against squared error along with a least squares fit line (solid) and 45° line (dashed) for reference.
conditions based on relative plume frequency and on a Gaussian model. Consistent with previous results, skills decline with increasing lead time, and are highest for forecasts made after the northern spring predictability barrier for seasons occurring prior to the subsequent such barrier, such as forecasts for DJF made in July or later.

A notable property of both the dynamical and statistical forecasts is that they tend to verify better against observations occurring earlier than the intended forecast targets, and this effect becomes greater with increasing lead time. This property is observed, to a lesser extent, in skill assessments based on hindcasts. There is no indication that this problem of phasing is related to any mishandling of the real-time data. Since the phasing problem is so systemic during the period of study, a statistical correlation is demonstrated to alleviate it to some degree, increasing the correlation between forecasts and their verifying observations.

In the period of study, the multimodel mean forecasts of dynamical models appear to slightly (and statistically insignificantly) outperform those of statistical models, representing a subtle shift from earlier studies. The difficulty of showing statistical significance is due to the fact that the greatest differences between dynamical and statistical model skill are evident mainly during certain seasons, greatly limiting the sample size. When the complete set of forecasts is compared, we find that dynamical models produce stronger (more strongly deviating from climatological) forecasts than statistical models. A linear analysis finds dynamical and statistical forecast errors to be highly correlated and to have no significant difference in magnitude. There is some hint of a greater difference between the two types of models for cold conditions.

Probabilistic skill scores show that much of the useful variation in forecast uncertainty can be captured in a model where the spread depends only on verification season and lead, at least in the short period studied. A Gaussian model where only the mean is interannually varying has similar probabilistic skill to that of the plume. Examination of the spread versus skill connection shows that the spreads of both the dynamical and statistical models capture the climatological dependence of forecast error on verification season and lead, but generally underestimate the magnitude of forecast uncertainty. This underestimate is more severe for the statistical models. However, it is important to remember that the plume is not optimally designed to capture uncertainty; dynamical ensemble means rather than ensemble members are displayed, and statistical forecasts of the mean conditions are shown.

Acknowledgments. The author is supported by a grant/cooperative agreement from the National Oceanic and Atmospheric Administration (NA10OAR4310210). The views expressed herein are those of the authors and do not necessarily reflect the views of NOAA or any of its subagencies. The authors acknowledge the useful comments of two anonymous reviewers.

REFERENCES


Fig. 15. Scatterplots of (a) statistical and dynamical multimodel mean forecasts, and (b) the errors of each. The solid line is a least squares fit to the data and the dashed line is a 45° line.


