Toward an Improved Multimodel ENSO Prediction

ANTHONY G. BARNSTON

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

MICHAEL K. TIPPETT

International Research Institute for Climate and Society, Columbia University, Palisades, New York, and Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia

HUUG M. VAN DEN DOOL AND DAVID A. UNGER

NOAA/Climate Prediction Center, Camp Springs, Maryland

(Manuscript received 25 July 2014, in final form 25 March 2015)

ABSTRACT

Since 2002, the International Research Institute for Climate and Society, later in partnership with the Climate Prediction Center, has issued an ENSO prediction product informally called the ENSO prediction plume. Here, measures to improve the reliability and usability of this product are investigated, including bias and amplitude corrections, the multimodel ensembling method, formulation of a probability distribution, and the format of the issued product. Analyses using a subset of the current set of plume models demonstrate the necessity to correct individual models for mean bias and, less urgent, also for amplitude bias, before combining their predictions. The individual ensemble members of all models are weighted equally in combining them to form a multimodel ensemble mean forecast, because apparent model skill differences, when not extreme, are indistinguishable from sampling error when based on a sample of 30 cases or less. This option results in models with larger ensemble numbers being weighted relatively more heavily. Last, a decision is made to use the historical hindcast skill to determine the forecast uncertainty distribution rather than the models' ensemble spreads, as the spreads may not always reproduce the skill-based uncertainty closely enough to create a probabilistically reliable uncertainty distribution. Thus, the individual model ensemble members are used only for forming the models' ensemble means and the multimodel forecast mean. In other situations, the multimodel member spread may be used directly. The study also leads to some new formats in which to more effectively show both the mean ENSO prediction and its probability distribution.

1. Introduction

Since early 2002, the International Research Institute for Climate and Society (IRI) has issued, each month, a collection of the forecasts from a large number of ENSO forecasting institutions, in the form of an ENSO prediction plume (Fig. 1). The forecasts predict the Niño-3.4 index in the tropical Pacific Ocean [SST averaged over 5°N–5°S, 120°–170°W; Barnston et al. (1997)]. The original idea behind the plume was to collect all of the current ENSO forecasts and show them on the same chart, in hopes of gleaning a sense of their collective forecast—both in central tendency and intermodel variation.

In late 2011 the forecast plume became a product of both IRI and the NOAA/Climate Prediction Center (CPC). Although the product has been popular and frequently viewed on the Internet, it has had several significant problems that have justifiably provoked criticism. A very simple problem is that the forecast producers do not unfailingly form their anomalies with respect to the same 30-yr base periods as encouraged, and IRI–CPC does not correct for such (usually minor) deviations. A more substantial problem is that model biases, evident upon examination of hindcasts, are not

Denotes Open Access content.

Corresponding author address: Anthony G. Barnston, IRI, Monell Bldg., 61 Rte. 9W, Palisades, NY 10964.
E-mail: tonyb@iri.columbia.edu

DOI: 10.1175/JAMC-D-14-0188.1

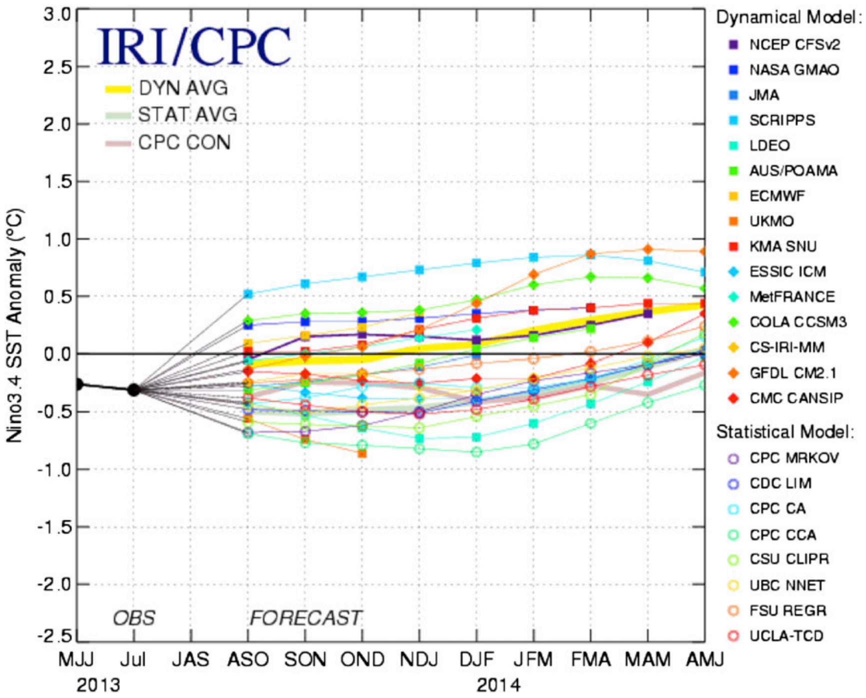

FIG. 1. ENSO prediction plume issued by IRI and NOAA/CPC in mid-August 2013, for the periods of August–October 2013 through April–June 2014. Recent observed SST anomalies in the Niño-3.4 region are shown by the black line on the left side.

corrected, and one or two of the forecasts are from models that lack hindcasts over an adequate historical period. Another problem is that the forecast spread within any individual model (the ensemble spread in dynamical models, or the standard error of estimate in statistical models), indicative of model uncertainty, is ignored and only the ensemble means of the forecasts of each model are shown. A final problem, affecting users, is that no attempt is made to provide a final forecast probability distribution, and users see the spread among the ensemble mean model forecasts by eye and are left to quantify the uncertainty on their own.

Of the problems listed above, the failure to adjust for mean biases appears most serious, because some of the dynamical models are found to have substantial (>0.5°C) biases. The sizes and signs of the biases vary widely among models, and often across forecast target times and lead times within an individual model. Hence, some of the spread in the model forecasts shown in Fig. 1, even at short lead times, may be due to differing model biases. The ENSO forecast plumes posted on the CPC website from the North American Multimodel Ensemble (NMME) project [Kirtman et al. (2014); shown online at http://www.cpc.ncep.noaa.gov/products/NMME/current/plume.html] have undergone hindcast-based bias correction by target month and lead time, and the resulting plume is noticeably less wide than the IRI–CPC plume at short leads. One option for the NMME plume also shows all ensemble members of all models, forming a set of very many lines on the plot.

The current work attempts to develop a protocol for selecting and processing the incoming forecasts for the IRI–CPC plume so as to reduce or eliminate the problems identified above, toward a more probabilistically reliable[1] and useful ENSO prediction product. Because bias correction requires a multidecadal hindcast history to evaluate bias, forecasts from models lacking an adequate hindcast history will not qualify for a higher quality version of the plume.

2. Data and methods

a. Data

This work uses as a test case a set of six models from the NMME project, because those models all have global 1982–2010 (29 years) hindcast data conveniently available in common format. The six models include

[1] Probabilistic reliability (Murphy 1973; Wilks 2006) refers to the condition that for a sufficiently large set of all forecasts of a given probability for an event (such as a 40% likelihood for an El Niño), the corresponding relative frequency of later observed occurrence of the event matches that probability.

TABLE 1. The models whose hindcasts are used in the prototype research on consolidation toward an improved ENSO prediction plume.

Model	Expanded model name	No. of members/max lead (months)
CMC1-CanCM3	Canadian coupled model 1	10/12
CMC2-CanCM4	Canadian coupled model 2	10/12
COLA-RSMAS-CCSM3	COLA–University of Miami–NCAR coupled model	6/12
GFDL-CM2pl-aer04	Modified version of the GFDL coupled model	10/12
NASA-GMAO-062012	Modified version of the NASA coupled model	12/9
NCEP-CFSv2	NOAA/NCEP coupled model	24/10

1) COLA–RSMAS–CCSM3, 2) NCEP CFSv2, 3) CMC1-CanCM3, 4) CMC2-CanCM4, 5) the GFDL-CM2pl-aer04 model, and 6) the NASA-GMAO-062012 model. All of the model data are placed onto a 1° grid at the originating center. Among the six models, the number of ensemble members varies from 6 to 24, and the maximum lead time varies from 9 to 12 months (see Table 1). Besides looking at the forecast characteristics of each model, those of the combined forecast (the multimodel ensemble or MME) are studied. The MME is formed by combining the individual ensemble members of all of the models. More will be said about this methodological choice in section 2b.

Here, we forecast the 1-month-mean Niño-3.4 SST index rather than seasonal mean SST as is done in the IRI–CPC plume and assume there is no loss of generality in the forthcoming findings. For observed verifying data, the Reynolds et al. (2002) OIv2 data on a 1° grid are used, matching the resolution of the NMME model forecast data.

b. *Methods*

1) FORMATION OF MME FORECASTS

Because the MME is formed by combining the individual ensemble members of all of the models, and some models have more members than others, the number of members acts as an effective weighting factor; for example, the NCEP CFSv2 has 4 times as many ensemble members as the COLA–RSMAS–CCSM3, so it exerts 4 times the weight of CCSM3 in forming the MME forecast. We chose this method of forming the MME because, while the difference may not be large, assigning as much weight to a model with relatively few members as to a model with a large number of members is expected to diminish the skill of the MME if the model forecasts have similar average skill, because it diminishes the effective number of independent realizations.

Another choice to be made in combining the forecasts of individual models into an MME is whether to weight each model's forecast in accordance with its hindcast skill. Past research has demonstrated that when there are no more than moderate apparent skill differences among models, and only 30 or fewer years of hindcast data are available, use of variable model weighting based on model hindcast skill has not been shown to result in higher cross-validated MME skill than equal weighting (Tippett and Barnston 2008; Peña and van den Dool 2008; Barnston et al. 2012; DelSole et al. 2013). The reason given in these studies is that when the model skills vary by the amounts seen here for the NMME models—not drastically—the skill differences do not exceed differences that are explainable by sampling variability and hence may not reflect true model quality differences. When weighting differences come about largely because of sampling variability, they often produce worse forecast results when applied to independent forecasts than when equal weighting is used. When a model shows skill much lower than that of most of the other models, that model may be removed entirely from the model set by subjective decision. Such action was not considered in our case, as the model showing the lowest average skill over all months/leads still contributes to the multimodel forecast skill during some of the months and leads.

2) VERIFICATION AND ADDITIONAL DIAGNOSTICS

For assessing the quality of forecasts in the analyses, we select the basic verification measures of temporal correlation and root-mean-square error (RMSE) and its skill score (RMSSS). The correlation measures discrimination in that it computes the extent to which the temporal phases of the variability in the observations are represented in the forecasts. RMSE is more a measure of final accuracy, as it summarizes the differences between the forecasts and the observations. Here, the actual physical differences are standardized using the standard deviation (SD) for the respective month, given that the interannual variability has a marked seasonal cycle, being lowest in northern spring and highest in late fall. Even with excellent discrimination, as measured by the correlation, RMSE may indicate large differences between forecasts and corresponding observations, as for example in the case of a large mean bias.

In addition to basic verification, each of the six individual models is analyzed for mean bias and for the ratio of the interannual standard deviation of the model's ensemble mean forecasts to that of the corresponding observations. Both analyses are done for purposes of possible correction of each type of systematic error. The standard deviation ratio, for example, for a model's ensemble mean, should be less than 1 in proportion to the historical correlation skill of the model. (The standard deviation ratio for individual ensemble members, however, should be approximately 1.) Both of the above analyses are done for each forecast target month and lead time, making for a target month/lead matrix of results.

Another analysis is carried out by examining the relationship between the individual models' ensemble spreads and the uncertainty implied by their hindcast skill. This relationship has implications for the value of the individual members of model ensembles in establishing a forecast probability distribution. We look at indicators in favor of, versus against, using the member spreads, as opposed to ignoring them and instead using only the multimodel ensemble mean forecast and generating the probability distribution based on the hindcast skill of the MME system.

In this study, cross validation is not used in assessing the hindcast skills of various methodological configurations. Reasons for this lack of use of cross validation are we are looking to assess 1) the *relative* skills of one configuration versus another and 2) the skill in prediction of the ENSO state is often at least moderately high, and the difference in skill between cross-validated and not-cross-validated skill is small at these higher skill levels. A somewhat similar reasoning was given in Becker et al. (2014). For low-skill forecasts, cross validation results in larger decreases in skill and can even lead to strongly negative skills as a result of a methodological degeneracy (Barnston and van den Dool 1993).

3. Results

Two basic diagnostic analyses done for each of the six models are related to 1) mean bias and 2) the forecast amplitude, as represented by the ratio of the interannual standard deviation of the forecasts to that of the corresponding observations.

a. Mean bias

An important finding is the presence of significant mean model biases that vary widely among models and often across forecast target months within an individual model. The large variation in model bias is illustrated in Fig. 2, showing the mean biases for each of the NMME models as a function of target month and lead time. For example, the CCSM3 model has a moderate negative bias for moderate-lead forecasts for northern late spring through summer, and this negative bias becomes severe for long-lead forecasts. The CMC1, on the other hand, has a slight positive bias for short-lead forecasts for targets in the first half of the calendar year. The GFDL has a strong negative bias for forecasts for northern autumn made at most nonshort leads. It is clear that individual model biases can be large and diverse, varying considerably as a function of target month and to a lesser extent as a function of lead time. Combining the ensemble mean forecasts of the multiple models can partially cancel the differing individual model biases in the MME, as shown in the last panel of Fig. 2. However, substantial bias may remain in the MME for some target and lead times, such as a negative bias for autumn target months at many leads. Correction of biases in the individual models would ensure a more bias-free MME forecast. The effect of varying uncorrected biases is an increase in the spread of the model ensemble means, as is often apparent in the existing IRI–CPC ENSO prediction plume (but not in the NMME plume). For example, in the IRI–CPC plume issued in August 2013 (Fig. 1), it is noted that the model disagreements at short lead times, such as 1–3 months, are larger than they should be considering the high skill levels at those leads. Correction of model biases by forecast target and lead times is considered a minimal adjustment toward an improved plume product.

Mean bias is corrected for an individual model by subtracting the mean of the difference between the model forecast and the observation over the 29-yr period of the hindcasts (1982–2010) for the forecasts of each target month for each lead time. The bias correction here is applied only to the ensemble mean. For purposes that require use of the individual members, a bias correction may be applied independently to each individual member. Although the corrections differ slightly from member to member for the same target month and lead time, the average of the member corrections is equal to the correction when applied just once to the ensemble mean forecast.

Mean biases do not affect a model's temporal correlation skill but do degrade the RMSE. Here, the RMSE-based skill score is calculated as

$$\mathrm{RMSSS} = 1 - \frac{\mathrm{RMSE}_{\mathrm{fct}}}{\mathrm{RMSE}_{\mathrm{cli}}}, \qquad (1)$$

where the numerator of the second term on the right-hand side is the RMSE of the forecasts and the denominator is the RMSE of perpetual forecasts of the observed climate mean. The first two panels in Fig. 3 show

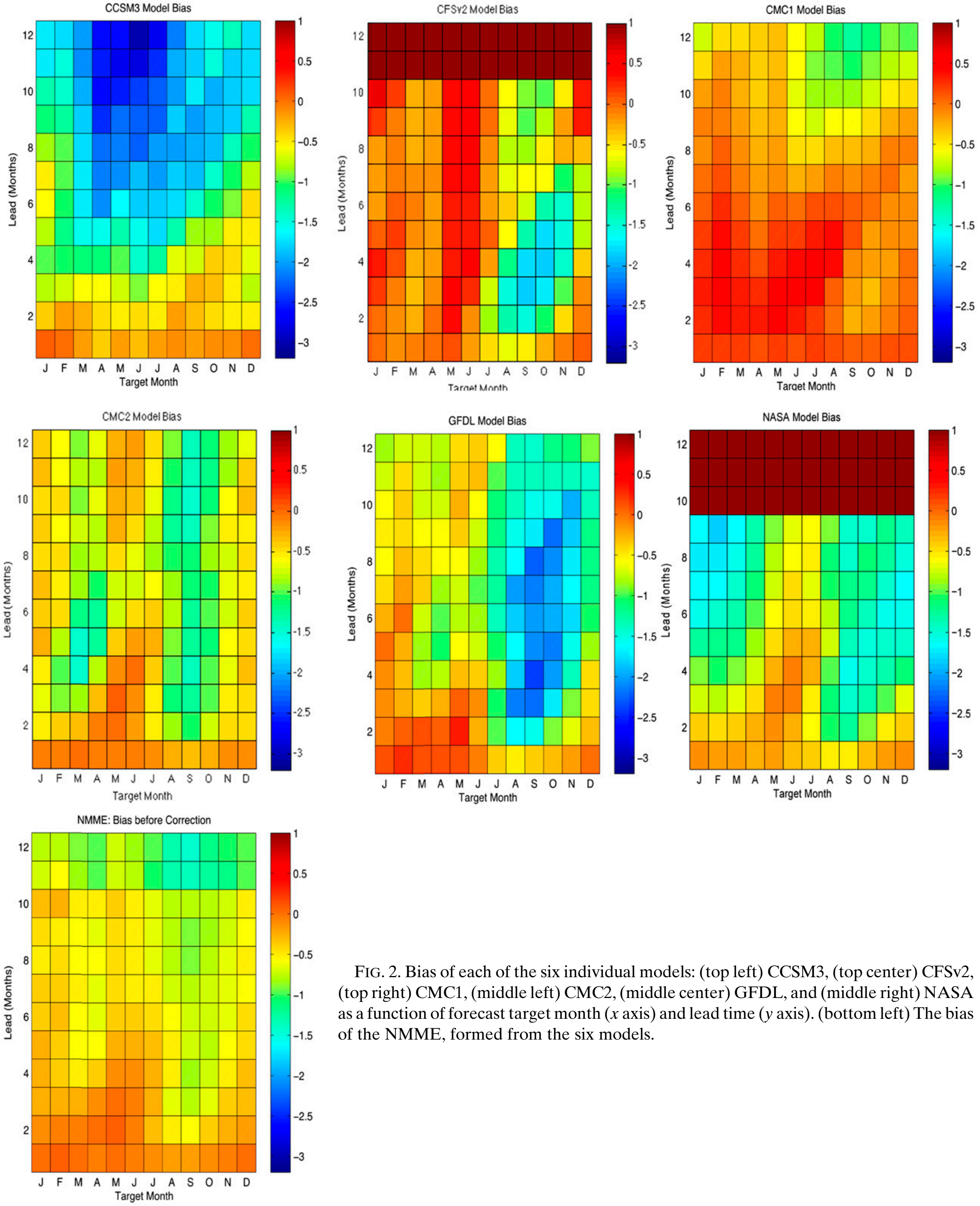

FIG. 2. Bias of each of the six individual models: (top left) CCSM3, (top center) CFSv2, (top right) CMC1, (middle left) CMC2, (middle center) GFDL, and (middle right) NASA as a function of forecast target month (x axis) and lead time (y axis). (bottom left) The bias of the NMME, formed from the six models.

the RMSSS of the MME ENSO forecasts first without any mean bias corrections for individual models, then with bias corrections. The mean bias correction makes for a substantial improvement in MME skill.

b. Forecast amplitude bias

Forecast amplitude is represented by the temporal standard deviation of the ensemble mean forecasts

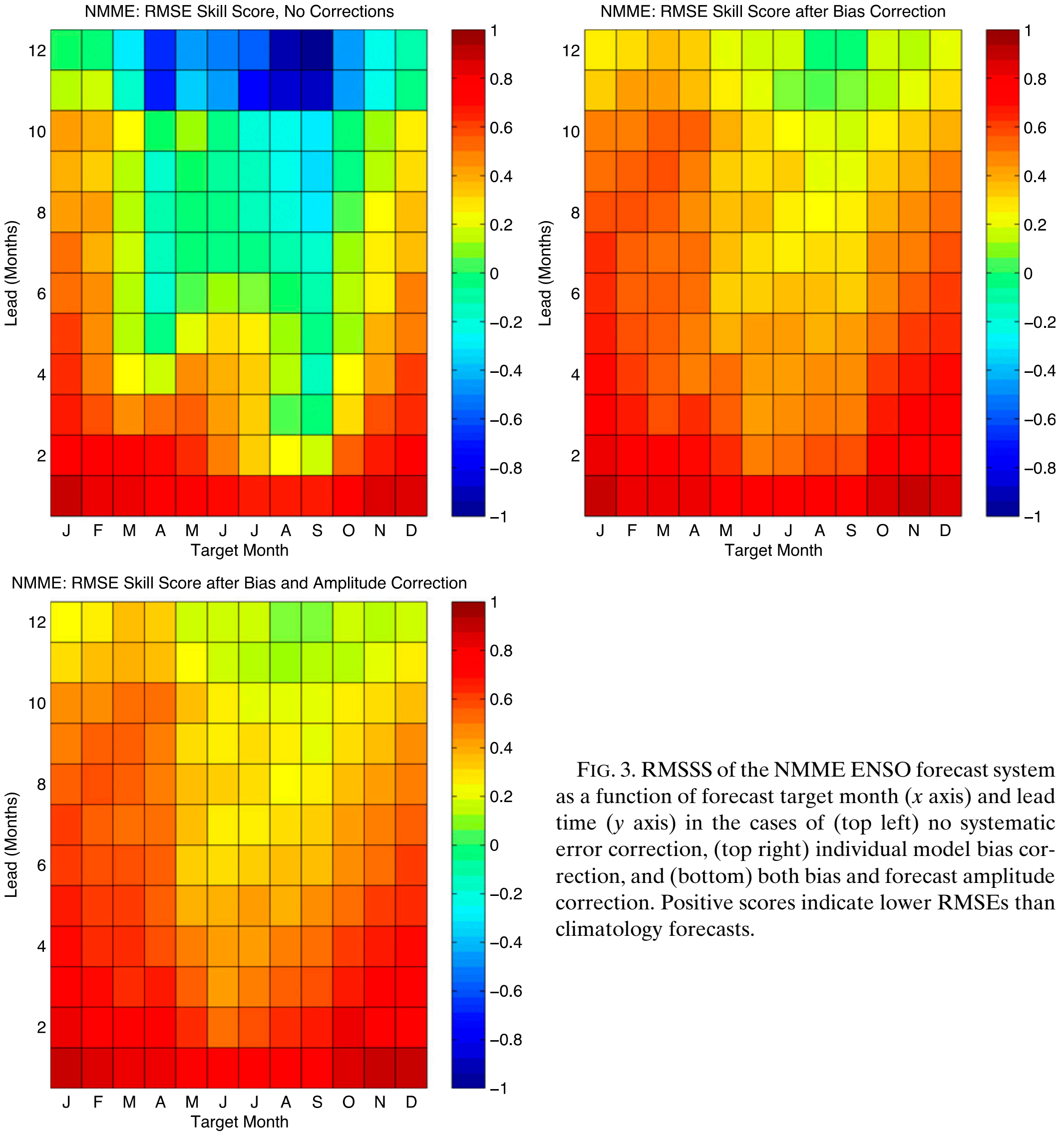

FIG. 3. RMSSS of the NMME ENSO forecast system as a function of forecast target month (x axis) and lead time (y axis) in the cases of (top left) no systematic error correction, (top right) individual model bias correction, and (bottom) both bias and forecast amplitude correction. Positive scores indicate lower RMSEs than climatology forecasts.

from a given individual model (or from the MME) for a given target month and lead time.[2] This amplitude may be compared with the amplitude of the corresponding observations. The set of forecasts of a single ensemble member forecast is expected to have amplitude approximately equal to that of the observations. On the other hand, unless there is perfect predictive skill, the set of ensemble mean forecasts (here, treated on an individual model basis) should have lower amplitude. From the perspective of a linear regression model, with a goal of minimizing the RMSE, the forecast amplitude of the ensemble mean should equal that of the observations multiplied by the correlation between the forecasts (represented by hindcasts) and the observations:

$$\text{amp}_{\text{correct}} = \text{SD}_{\text{obs}}\text{cor}_{\text{fct,obs}}\,. \tag{2}$$

The reduction in amplitude may not be fully realized with the typically used ensemble sizes of 20 members or less. This fact is easiest to see in an example of no predictive skill (cor = 0).[3] In this case, although (2) implies there should be no forecast amplitude (i.e., all forecasts equaling the climatological mean), there is still a small

[2] The amplitude of individual models or the MME is not affected by bias corrections.

[3] When the correlation between ensemble mean forecasts and observations is negative, a zero correlation is assumed.

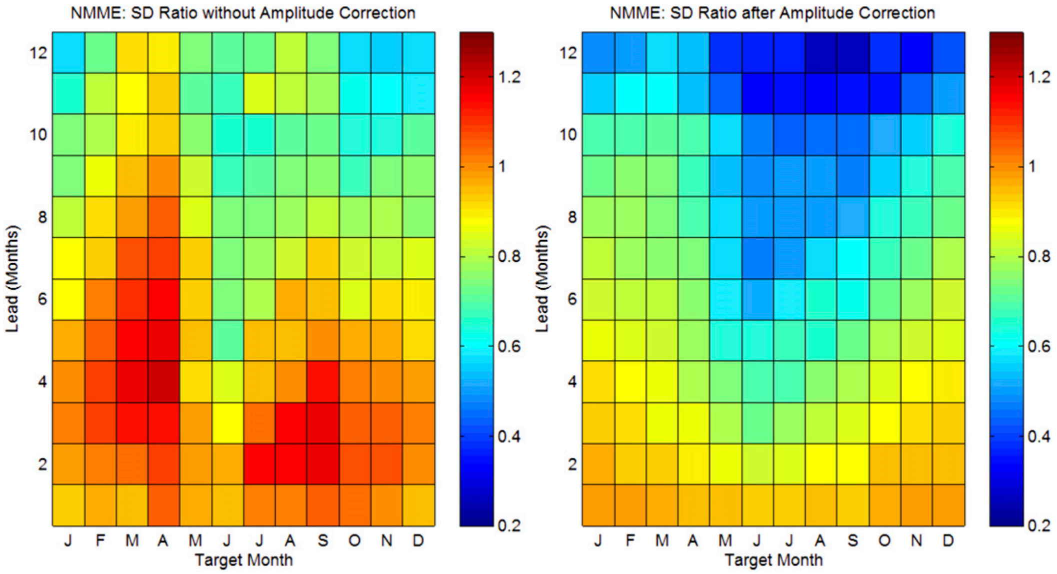

FIG. 4. Ratio of the amplitude of the MME forecasts to that of observations (left) before and (right) after correction of the forecast amplitudes of the individual models, as a function of forecast target month (*x* axis) and lead time (*y* axis). See text for details.

expected residual amplitude due to uncanceled noise, and

$$\mathrm{SD}_{\mathrm{ensmean}} = \frac{\mathrm{SD}_{\mathrm{obs}}}{\sqrt{n}}, \qquad (3)$$

where n is the number of ensemble members. A variation of this relationship is that between the SD of a single ensemble member of an individual model and that of the ensemble mean:

$$\mathrm{SD}_{\mathrm{ensmean}} = \mathrm{SD}_{\mathrm{member}}\sqrt{\frac{1}{n} + \mathrm{cor}\frac{n-1}{n}},$$

where cor is the correlation among the ensemble members associated with common model signals, even when not confirmed by real-world signals (Becker et al. 2014). When predictability (in the model) is high, this correlation is high and the reduction of $\mathrm{SD}_{\mathrm{ensmean}}$ with ensemble size is slower.

Including the contribution of the expected residual component of the noise amplitude due to the finite ensemble size, the amplitude error (or bias) can be expressed as the ratio between the ratio of the standard deviation of the ensemble mean forecasts to that of the observations, and $\mathrm{amp}_{\mathrm{correct}}$ as given in (2) above:

$$\mathrm{error}_{\mathrm{amp}} = \frac{\mathrm{SD}_{\mathrm{ensmeanfct}}}{\mathrm{SD}_{\mathrm{obs}}} \bigg/ \mathrm{amp}_{\mathrm{correct}}. \qquad (4)$$

The amplitude correction factor that would make $\mathrm{error}_{\mathrm{amp}}$ equal to 1 is exactly the linear regression coefficient. The correction process is therefore a correction of each individual model's ensemble mean using linear regression. Just as bias can be corrected by adding a constant to all forecasts, systematic errors in amplitude can be corrected by multiplying all forecast anomalies by the factor that makes their standard deviation equal to $\mathrm{amp}_{\mathrm{correct}}$. Although correcting the forecast amplitude does not change the correlation skill of an individual model, it improves RMSSS and can improve both skill measures in the MME. Amplitude correction and the resulting proper specification of the forecast signal are also relevant for probabilistic measures of forecast quality such as the ranked probability skill score and reliability (Tippett et al. 2014).

Figure 4 shows the ratio of the amplitude of MME forecasts to that of observations before and after correction of the amplitudes of the individual models. Reductions in the original amplitude are pervasive from the correction and are greatest at long leads, especially in forecasting the second half of the calendar year from very early in the year—when expected skills are lowest (Barnston et al. 2012) as a result of the northern spring ENSO predictability barrier (Jin et al. 2008).

Correction of amplitude bias does not have a dramatic effect on the RMSSS for the MME for the six models used here. The second and third panels in Fig. 3 show the RMSSS before and after amplitude correction. The

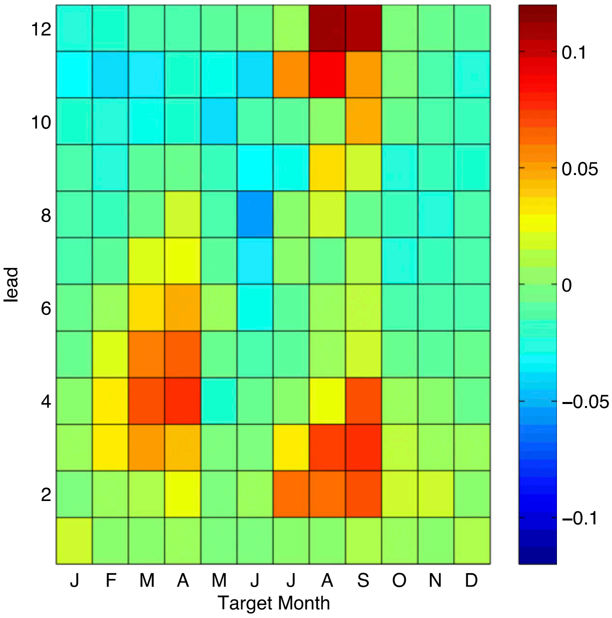

FIG. 5. Difference in RMSSS between amplitude-corrected and non-amplitude-corrected MME forecasts (cf. top-right and bottom panels of Fig. 3) as a function of forecast target month and lead time.

difference between the two RMSSS results is shown in Fig. 5,[4] which shows improvements much smaller and more specific to target/lead time than those associated with the correction of mean bias.

c. Ensemble spread: Useful or not?

Ideally, for any given single forecast, the standard deviation of the forecasts of the members of an individual model should indicate the uncertainty in the forecast. Much research has been done exploring the degree to which ensemble spreads provide realistic estimates of forecast uncertainty (e.g., Hamill et al. 2004; Schubert et al. 2008), and while results have been somewhat inconclusive, they have suggested at most a weak relationship, particularly for weather-aggregative (e.g., weekly to seasonal) time scales.

Here, we assess the relationship between the ensemble spreads of individual model forecasts and the uncertainty as measured by an alternative approach based on the hindcast correlation skill of the model. In the alternative approach, we compute the standard error of estimate (SEE), which is a function of the hindcast temporal correlation skill (cor) as

$$\mathrm{SEE} = \mathrm{SD}y\sqrt{1 - \mathrm{cor}_{xy}^2}, \qquad (5)$$

where SD_y is the standard deviation of the observations for the target month in question.[5] The SEE implies an uncertainty that remains fixed for a given start and lead time; that is, it does not change from year to year. Recently, Kumar and Hu (2014) demonstrated that ensemble spreads of individual models are in fact approximately constant from year to year for the same season and lead time. This constant spread, changing only slightly year to year because of the expected sampling variability for the given ensemble size, exists because of the constant underlying signal-to-noise ratio, leading to an unchanging uncertainty regardless of the signal strength for any particular year.[6] Changes in spread from year to year are attributed mainly to this sampling variability, which decreases with increasing ensemble number. Spread changes are also attributed to a likely lesser, but still debated, extent to varying true uncertainties related to differing physical situations from one year to another.

In our analyses, we find that the spreads of the individual models do remain approximately constant from year to year for given target month/lead time combinations. More revealing, however, is that different models have differing relationships between their mean ensemble spreads and the uncertainties implied by their respective average hindcast-based skills. Specifically, some models have spreads that are usually too small considering the uncertainty associated with their expected skill, while other models have spreads more appropriate for their expected skill. Figure 6 shows the ratio of ensemble spread to the skill-based SEE for each of the NMME models. The CFSv2 model generally has an approximate equivalence between ensemble spread and SEE. The CMC2 and GFDL models also have this favorable correspondence between spread and SEE, but only for some of the target months and lead times. Otherwise, in the case of the other three models (CCSM3, CMC1, and NASA), the spreads are more pervasively smaller than the hindcast skill-based SEE.

The forecast amplitude correction, discussed in section 3b above, should be applied to the ensemble mean

[4] A few target–lead combinations show degradation of RMSSS because of the sampling issue associated with the finite and differing ensemble sizes of the individual models, which are corrected on an individual basis before combining.

[5] In a normally distributed forecast set, the SEE should equal the RMSE of the verified forecasts, and, additionally, with a very large set of ensemble members the average ensemble spread ideally should equal both of these.

[6] A related finding appears for the spread among the ensemble means of the model forecasts on the ENSO prediction plume: year-to-year differences for a fixed lead time and target season are at most weakly related to true year-to-year differences in uncertainty or skill (Tippett et al. 2012).

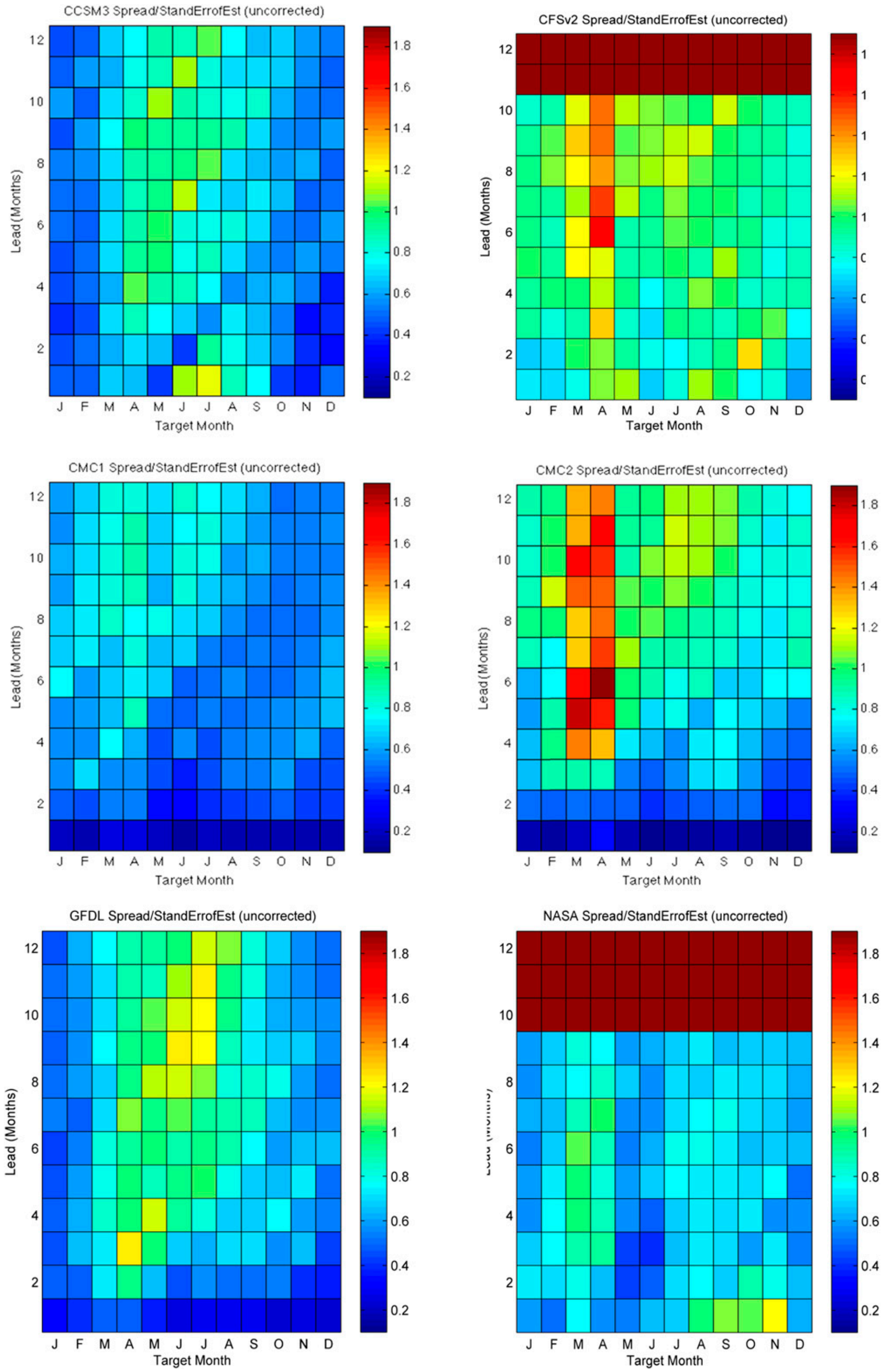

FIG. 6. Ratio of ensemble spread (as standard deviation of ensemble members about the ensemble mean) to the SEE based on the hindcast correlation skill, as a function of forecast target month (*x* axis) and lead time (*y* axis). Results are shown for the (top left) CCSM3, (top right) CFSv2, (middle left) CMC1, (middle right) CMC2, (bottom left) GFDL, and (bottom right) NASA models.

forecasts of an individual model, and not to the individual ensemble members whose amplitudes are expected to be larger. Because amplitude adjustments to ensemble means are usually decreases rather than increases, when applied to individual members they make an often already too small spread (e.g., Fig. 6) even smaller (not shown). A correction for the individual members should just translate the members' deviations from the original ensemble mean forecast to the amplitude-corrected mean by adding the difference between the latter and the former means. The preserved deviations about the corrected ensemble mean could then be adjusted by multiplying by the factor making the members' interannual standard deviation equal SEE—a factor that would most often exceed unity.[7]

Figure 7 shows the ratio of the spread of all members of all models about the multimodel ensemble mean to the SEE of the NMME over the hindcast period, under conditions of 1) no corrections, 2) bias corrections for all members of each of the individual models, and 3) bias and amplitude corrections for all members of each of the individual models.[8] The ratio for the uncorrected forecasts shows too large a spread, due to significant uncorrected biases that differ among the models. A plume showing such a collection of uncorrected forecasts typically features relatively tightly clustered members within individual models, but comparatively widely differing central locations of the forecasts by one model versus another. This situation gives rise to a plume of ensemble means that vary considerably even during times of high expected ENSO forecast skill, such as that illustrated in Fig. 1, and implies larger forecast uncertainty than actually exists.

Following individual model bias corrections the spread/SEE ratio for the NMME decreases toward unity, although the spread of forecasts for boreal spring target months remains at higher than ideal levels, likely because of inflated ensemble mean amplitudes during the northern spring predictability barrier, a time of longstanding difficulty in ENSO prediction (e.g., Barnston et al. 1999; Jin et al. 2008) but not fully acknowledged by the models as a time of low signal-to-noise ratio. The inflated amplitudes of ensemble mean forecasts of some of the models cause inflated forecast spread for this challenging target season because the ensemble mean forecasts of some models diverge from one another by large amounts, a situation similar to that caused by differing uncorrected model biases.

When the forecasts of individual models are corrected for both mean bias and amplitude bias, applying the ensemble mean amplitude correction to the individual members of each as well (a practice not recommended for an individual model considered alone, as discussed above), ratios of spread to SEE decrease further and average slightly lower than unity as a result of some leads and target months having too small a spread (e.g., forecasts for June for the second lead time have a ratio of 0.68, and some of the very longest lead forecasts have ratios below 0.5). This result shows many seasons/leads with acceptable ratios, despite the spreads about the ensemble means of individual models tending to be too small (Fig. 6), and usually even smaller following amplitude bias correction. The NMME ratios may average close to unity, with some seasons/leads deviating somewhat in either direction, because of remaining differences in the ensemble mean forecasts of individual models, even after bias and the amplitude corrections that are usually damping. These differences may represent legitimate forecast "differences in opinion" among models, which may remain substantial because of the different physical representations (representing uncertainty) among the models. In other words, the final result may be acceptable because of the offsetting influences of two (possibly related, as both may be based on the signal-to-noise ratio) model flaws: 1) too small a spread in most of the individual models, particularly after applying an amplitude correction to individual members that is intended for the ensemble mean, and 2) too much "difference in opinion" among bias-corrected model ensemble means due to the unique details of the model physics leading to unique errors. The final ratios would likely be larger if the individual ensemble members were not subjected to the amplitude correction using (2), but rather by adding a constant to the members' forecasts equaling the corrected ensemble mean minus the original ensemble mean for the given model.

When spread/SEE ratios do not approximate unity, a viable alternative to estimating forecast uncertainty by using the ensemble member spreads directly (following the above-mentioned adjustments) is to use the uncertainty distribution that is statistically derived from the NMME hindcast skill, using the standard error of estimate (SEE) as in (5). In this case the only error in the uncertainty estimation would come from estimating

[7] A practical alternative for establishing the forecast probability distribution is to disregard the individual members, and use SEE to construct a Gaussian probability density straddling the corrected ensemble mean forecast. This alternative ignores details (gaps, clusters, and asymmetries) within the distribution of the ensemble members—details that may not be believed to be meaningful in seasonal prediction.

[8] These amplitude corrections are those applicable to the ensemble means of the respective individual models and, thus, would usually contract each model's spread.

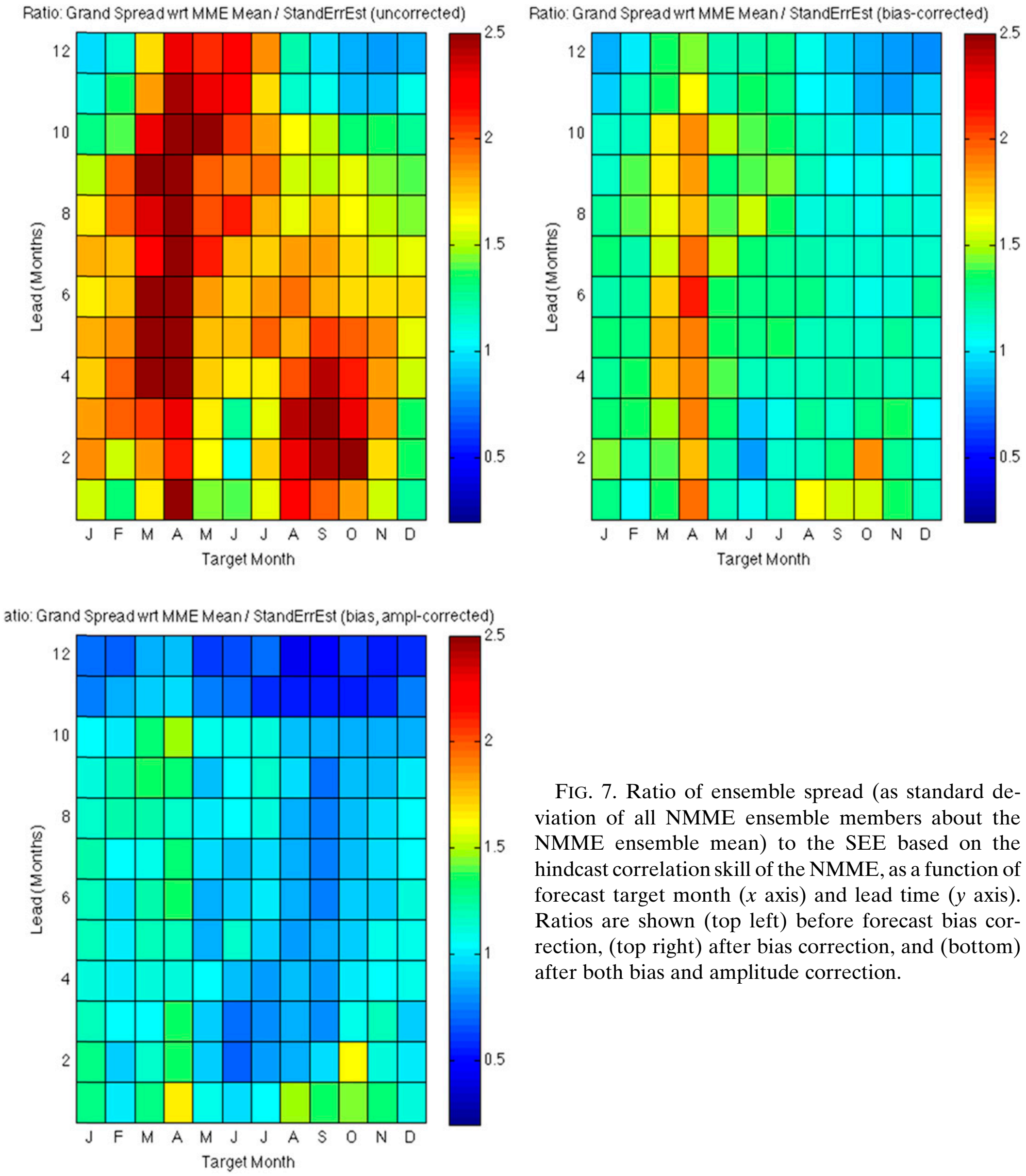

FIG. 7. Ratio of ensemble spread (as standard deviation of all NMME ensemble members about the NMME ensemble mean) to the SEE based on the hindcast correlation skill of the NMME, as a function of forecast target month (x axis) and lead time (y axis). Ratios are shown (top left) before forecast bias correction, (top right) after bias correction, and (bottom) after both bias and amplitude correction.

hindcast skill on the basis of the finite sample of cases (29 years here). Let us briefly consider the pros and cons to selecting this alternative approach.

First, to weigh the acceptability and desirability of using SEE instead of the NMME's ensemble spread to estimate forecast uncertainty in the case of the current ENSO forecasts, we ask whether the deviations from unity in the ratios shown in Fig. 7 are statistically significant, or if they could have arisen because of sampling variability even if the true underlying ratio is unity. If we cannot reject the hypothesis that the true ratio is unity, then we lack evidence that the forecast system is not acceptably calibrated.[9] For any individual cell in Fig. 7, the F distribution can be used to determine a confidence interval straddling unity for the spread/SEE ratio, given the sample sizes for the number of ensemble members in the NMME spread ($n = 72$ for lead times of 9 months or less) and the number of years contributing to SEE

[9] If the power of the statistical test is low, because of inadequate sample sizes of the forecast ensemble members and/or years of forecasts, then a true excursion from unity of the spread/SEE ratio may not be detected as being statistically significant.

($n = 29$). The resulting 90% confidence interval for the ratio is 0.76 to 1.32, indicating a moderately wide range of possibilities when the true ratio is unity. Within the first 9 months of lead time, 14 cells have a ratio outside of this interval, where 11 would be expected on average by chance. An analytic assessment of collective statistical significance (field significance) across all 108 cells requires knowledge of the number of degrees of freedom provided in the combined target season and lead time dimensions; this can also be done empirically using a Monte Carlo approach. An assessment of degrees of freedom for ENSO forecasts by the CFSv1 and CFSv2 models was conducted previously (Barnston and Tippett 2013), where the set of 12 target seasons and the first 9 lead times (108 cells) were estimated to provide only 2.9 times as many degrees of freedom as a single cell in Fig. 7.[10] Multiplying the number of years (29) by 2.9 and assessing the number of excursions outside of the 90% confidence interval among the 108 individual cell ratios in Fig. 7, we cannot reject the hypothesis that the true underlying ratio of spread to SEE is unity in the NMME forecasts derived from the six models used here. We therefore conclude that the ratios here do not differ statistically from unity.

These significance diagnostics imply that the NMME ensemble spreads may be used directly to form the forecast uncertainty estimates for our multimodel forecasts. However, the ratio of spread to SEE might have been statistically different from unity if a different set of individual models had been used, such as a set that omits CFSv2 and either one of CMC2 or GFDL. (In fact, the ratios appear to be substantially less than unity for leads greater than 10 months, where CFSv2 is not included.) The addition of more models could also alter the resulting NMME ratios of spread to SEE. When spread/SEE ratios are not in the neighborhood of unity, either overall or for specific target seasons or lead times, the spreads may not be able to be taken literally to best estimate the uncertainty of their forecasts. There are examples showing that some of today's leading models still have calibration needs (e.g., Goddard et al. 2013), not just for forecast mean and amplitude but also for ensemble spread, and the use of SEE to represent forecast uncertainty ensures such calibration.

When SEE rather than the spread is used to determine the uncertainty envelope, the individual member forecasts still have value in forming the individual model ensemble means, following bias correction, as they contribute skill to the MME mean forecast, and the more contributing members, the better the ensemble mean reflects the forecast signal. The multimodel ensemble mean has been shown to yield greater average deterministic and probabilistic skill, leading to better value, than the most skillful individual model (e.g., Tippett and Barnston 2008; Kirtman et al. 2014). Probabilistic reliability is one aspect of forecast value, as the probability distribution is often just as important to user needs as the best-guess single deterministic forecast. Toward providing the most reliable and useful probability distribution, it is safe practice (assuming an adequate sample of hindcasts, e.g., at least 30 years) to generate the uncertainty distribution on the basis of the historical hindcast skill of the multimodel system. Instances of a lack of realism of individual model ensemble spread may be a result of an unrealistic model signal-to-noise ratio (e.g., Becker et al. 2014). The use of SEE in describing the uncertainty distribution is commonly embedded in regression-based statistical prediction methodologies (e.g., Shongwe et al. 2006); and a symmetric, Gaussian distribution is often assumed. The Gaussian is a reasonable approximation for the distributions of tropical Pacific Ocean SST in the Niño-3.4 region (Chiodi and Harrison 2009), but not for SSTs farther east (which have positive skewness) or farther west (negative skewness).

At present, we intend to use SEE for the uncertainty distribution for this improved ENSO prediction plume, unless we find clearly favorable spread/SEE ratios in the larger (currently unknown) set of models to be used. However, depending on the individual case, either approach may be justified. We expect that in the future, dynamical model physics and engineering issues associated with model forecast runs (e.g., initialization, numerics, computer power) will be improved to the point where model spreads will routinely be relied upon as direct indicators of forecast uncertainty. It is desirable to use the ensemble member distribution because the individual members contain the physics, and so ideally they could be used to identify the physical sources of uncertainty and could be trusted in cases of unique forecast probability distributions such as bimodal or highly skewed ones as seen occasionally even in forecasts of a 3-month mean climate or ENSO state.

4. Most useful graphical format for ENSO prediction

In addition to determining the methodologies that deliver best predictive skill, another goal is to produce an ENSO prediction plume that shows probabilistic

[10] This seemingly poor benefit to degrees of freedom results from the high autocorrelation between model forecasts of a given target season at varying lead times, as well as between forecasts for varying seasons at a fixed lead time.

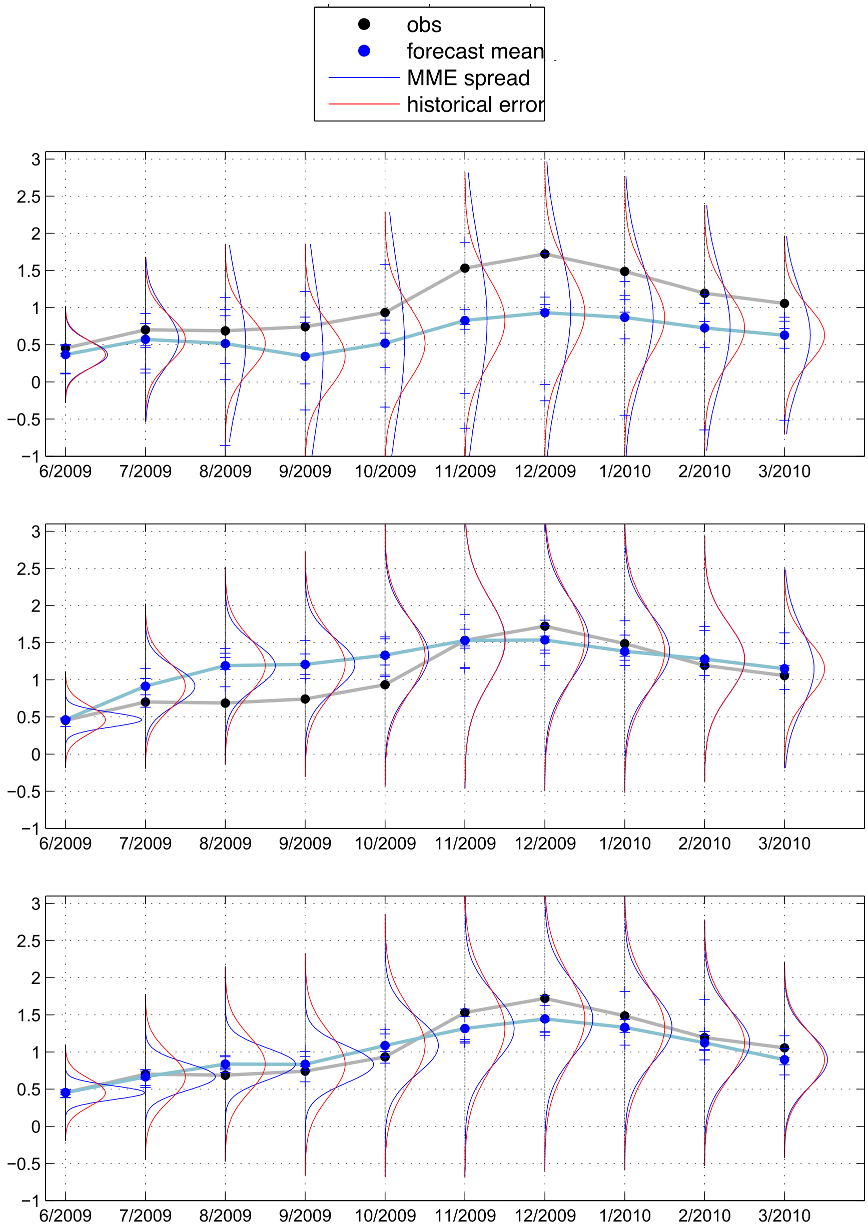

FIG. 8. MME forecasts from June 2009 for the period of the 2009/10 El Niño event. Forecasts (top) without any corrections, (middle) after bias correction, and (bottom) after bias and amplitude correction. The blue line and solid dots show the MME mean forecasts; the black line and dots show the observations. The horizontal ticks on the vertical line for each month show individual model ensemble mean forecasts. The thin blue vertical Gaussian distribution curves show forecast uncertainty based on the MME spread, and the thin red vertical distribution curves show uncertainty based on the hindcast skill-based SEE.

predictions more clearly than the raw "spaghetti" plumes of the first decade of the 2000s, to be of greater utility to decision makers. Toward that end we experiment with various choices of graphical formats. Figure 8 shows a format in which the "best guess" single forecast is shown, while the uncertainty distribution about that forecast is shown using vertically oriented bell-shaped curves. Besides illustrating the format, Fig. 8 also shows the effects of mean bias correction and amplitude corrections on the hindcast for the 2009–10 El Niño made in June 2009, when the event was just about to begin. The panels in Fig. 8 show the forecast, along with its uncertainty as represented directly by the MME ensemble spread as well as by the SEE that reflects the historical hindcast skill. We regard the SEE-based uncertainty distribution as the most realistic one, as it leads to a probabilistically reliable probability distribution when the Gaussian assumption is adequately satisfied.[11] In the top panel no bias corrections are done on individual models, and the MME ensemble spread is larger than the SEE-determined spread because the differing individual model mean biases artificially inflate the former. Correction of the mean biases leads to an improved MME forecast (middle panel) and more realistic widths of the uncertainty distributions. Correction for the amplitude as well as the bias results in underestimation of the uncertainty at short leads and better forecasts during the event's amplification phase, but slightly worsened underestimation of the peak strength of the event. While the amplitude correction may affect the forecasts for individual El Niño or La Niña events differently, it is expected to slightly improve the RMSE of the forecasts on average over all years.

Informal feedback regarding the above forecast format indicates that the vertically oriented Gaussian curves are often ineffective in communicating the uncertainty, as many nonclimate specialists do not easily understand the meaning of the probability densities implied by the vertical curves. A format more similar to that of the existing plume, but with probability-indicating lines or interval bands, is believed to be more understandable to the average user.

Another, more fundamental, problem with the methodology and format shown in Fig. 8 is that there is no information about the joint distribution of forecasts at different leads. For instance, even given the parameters of the probability distribution functions, it is not possible to evaluate correctly the probability of Niño-3.4 SST exceeding a given threshold for multiple consecutive months.

It is possible to generate a plume of equally likely scenarios for a prediction of the ENSO state from a given starting month with the correct joint distribution among the different leads using a Gaussian random number generator, by employing the MME mean forecast in combination with the historical covariance of the errors over the hindcast period. This error covariance contains the error distributions for each lead time (i.e., the SEE) and, thereby, accounts for the correlation skills for each lead. The main idea behind this formulation is that the forecast scenarios are not entirely "reset" with each increment in lead time; rather, errors at one lead time tend to persist to the next lead time because of a positive correlation of errors between two lead times. For the forecasts from June 2009, for example, for the case of mean bias correction but no amplitude correction (top-right panel in Fig. 3), the mean forecasts for the months of June–March 2010 are 0.44, 0.64, 0.83, 0.82, 1.05, 1.27, 1.43, 1.37, 1.19, and 0.94, respectively; and there is a matrix of error covariances for each lead time with each other lead time for the June start time computed over all years in the hindcast period.

Figure 9 shows a number of options for expressing the forecast from June 2009 along with its uncertainty distribution. In the top-left panel of Fig. 9, the thick line in the middle shows the mean forecast, among a family of lines showing various percentiles within the forecast distribution: 1, 5, 15, 25, 50, 75, 85, 95, and 99. This format provides a choice of intervals that may matter most to various users. A similar format is shown in the top-right panel, except the lines are represented by smooth curves rather than line segments, and the more likely intervals are shaded with increasingly dark color. The two bottom panels in Fig. 9 both show the forecast mean, the 15th and 85th percentiles, and numerous randomly generated lines showing equally likely individual scenarios (100 in the left panel, 200 in the right panel). Although each line is equally likely, their density (able to be seen by eye) is greatest near the forecast mean, and lowest far from the mean, and this density difference expresses the relative likelihood that the observation will occur in any of the regions on the plot.

The bottom two panels in Fig. 9, using the historical error covariance, most resemble the existing IRI–CPC plume, showing individual model predictions, in that a dearth of specific probability levels is indicated and the user is left to surmise the probabilities largely by visual inspection. However, the large number of lines better describes the probability distribution than the ~25 lines (one for each model's ensemble mean) shown in today's existing plume. A main reason for the better description, besides the larger sample of lines, is that the new plots are equivalents to the forecasts of individual model ensemble members, while the lines on the existing plume plot are the ensemble means of the various models, whose differences are likely due to differing model biases as well as true uncertainty. The observation behaves as a single ensemble member, not like an ensemble mean, so that the simulated lines on the plots shown here are the more appropriate reproduction of the possible scenarios. (The current IRI–CPC plume format for this six-model example would consist of six lines connecting the tick marks—one for each of the models—on the vertical axis of the middle panel in Fig. 8.)

[11] The correlation-based SEE, used to define the width of the uncertainty distribution, produces a probabilistically reliable forecast distribution as shown using linear regression theory (Tippett et al. 2014).

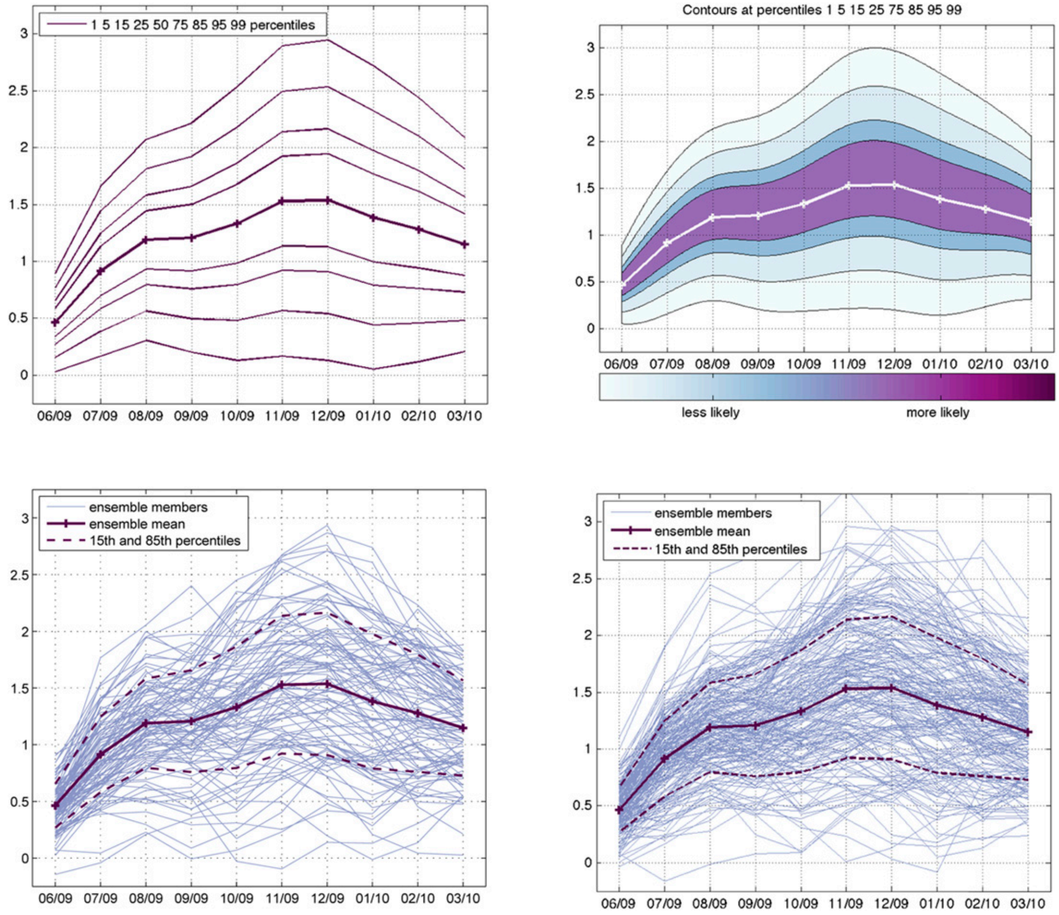

FIG. 9. Four possible formats for an improved ENSO prediction plume. All show the forecast made in June 2009 for the 2009–10 El Niño episode where the models are corrected for mean bias but not amplitude bias (corresponding respectively to the top-right and middle panels in Figs. 3 and 8). In the top-right panel, ''less likely'' and ''more likely'' refer to lower and higher probability densities, respectively. See the text for details.

A user's preference among the four options shown here is expected to vary in accordance with how readily they find that each communicates the uncertainty distribution, as well their particular decision needs. We can imagine that there are sophisticated users who can use the forecast probability distribution in more general and complex ways than just looking at pictures. For such users we can provide the complete forecast distribution (i.e., the forecast mean and covariance). With this information users can evaluate more complex questions that depend on the joint probability. The best way to present the plume is likely an evolving process.

5. Summary

In forming multimodel forecasts for the ENSO-related Niño-3.4 SST, experiments led us to the decision to correct all individual models for mean bias, and preferably also for amplitude bias, before combining their predictions. A second decision is to weight the individual ensemble members of all models equally in consolidating their ensemble means to form a multimodel ensemble mean forecast, regardless of estimated hindcast skill, because skill differences, when not extreme, are indistinguishable from sampling error when based on a sample of approximately 30 cases. Thus, models with larger ensemble numbers are effectively weighted proportionally more heavily.

A final decision is made to use the historical hindcast skill—by tradition, correlating the ensemble mean forecasts with their corresponding observations—to determine the uncertainty distribution rather than using the models' ensemble spreads, as this ensures probabilistic reliability in the forecast uncertainty distribution. The ensemble spreads of most of the individual models are found to be too small, implying less uncertainty than exists in reality. This underestimation of uncertainty is seen to be much reduced or eliminated in the multimodel ensemble for many target seasons and lead times, owing to sizeable differences in the mean forecasts among constituent models. In fact, the ratio of the multimodel member spread to the skill-based standard error of the estimate is found to be statistically indistinguishable from unity. Nonetheless, the decision to use the standard error is believed to be safer, in view of the expected addition of more individual models having unknown ensemble spread (signal to noise) characteristics. Using model hindcast skill to form the forecast uncertainty distribution implies that the individual model ensemble members are not used explicitly, but instead for their role in forming the individual models' ensemble means and, in turn, the mean of the multimodel forecast distribution.

Using the multimodel ensemble's historical standard error rather than its spread is not necessarily a superior option in all cases of multimodel ensemble ENSO or climate prediction. When the set of models can be demonstrated to provide a well-calibrated uncertainty distribution, then the spread of its members may serve as an equally valid, if not superior,[12] indicator of forecast uncertainty.

In the near future, we expect to establish the most usable and actionable formats for the ENSO prediction forecast graphic, based on additional extensive feedback from users. Users will be approached with various candidate formats and asked to rank them for usability, provide reasons for their ranking decisions, and provide ideas for formats yet better than those offered. It is likely that multiple plume formats will be adopted, and that the accompanying forecast data will be provided for those wishing to create their own graphics. What users have become used to previously also plays a role. The popularity of the current IRI–CPC plume, despite its problems, seems to favor the two bottom panels of Fig. 9. A variation of this scheme is found on the NMME site where a real-time plume based on approximately 100 individual ensemble members is shown. These are actual model realizations, whereas the bottom panels of Fig. 9 are recreated model traces generated based on the properties of the system.

Another expectation is to expand the set of model inputs to the multimodel ensemble beyond the six NMME models used for the experimentation here. Additional models would minimally need to have comparable 30-yr hindcast records from which to correct the mean bias and amplitude bias. At least 10–15 models are expected to qualify, with hopes for still greater numbers.

Acknowledgments. We thank Dr. Klaus Wolter for his thorough review of this paper and his set of constructive suggestions, and also the anonymous reviewers who provided helpful comments and criticism. This study was supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections Program Award NA12OAR4310082. Most of the authors participated in NOAA's Climate Prediction Task Force.

REFERENCES

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977, doi:10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

——, and M. K. Tippett, 2013: Predictions of Nino3.4 SST in CFSv1 and CFSv2: A diagnostic comparison. *Climate Dyn.,* **41,** 1615–1633, doi:10.1007/s00382-013-1845-2.

——, M. Chellia, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.–Ocean,* **35,** 367–383, doi:10.1080/07055900.1997.9649597.

——, M. H. Glantz, and Y. He, 1999: Predictive skill of statistical and dynamical climate models in SST forecasts during the 1997/98 El Niño episode and the 1998 La Niña onset. *Bull. Amer. Meteor. Soc.,* **80,** 217–243, doi:10.1175/1520-0477(1999)080<0217:PSOSAD>2.0.CO;2.

——, M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions during 2002–11: Is our capability increasing? *Bull. Amer. Meteor. Soc.,* **93,** 631–651, doi:10.1175/BAMS-D-11-00111.1.

Becker, E., H. van den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate,* **27,** 5891–5906, doi:10.1175/JCLI-D-13-00597.1.

Chiodi, A. M., and D. E. Harrison, 2009: Characterizing warm-ENSO variability in the equatorial Pacific: An OLR perspective. *J. Climate,* **23,** 22 428–22 439, doi:10.1175/2009JCLI3030.1.

DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.,* **139,** 176–183, doi:10.1002/qj.1961.

Goddard, L., and Coauthors, 2013: A verification framework for interannual-to-decadal predictions experiments. *Climate Dyn.,* **40,** 245–272, doi:10.1007/s00382-012-1481-2.

Hamill, T. M., J. S. Whitaker, and X. Wei, 2004: Ensemble reforecasting: Improving medium-range forecast skill using retrospective forecasts. *Mon. Wea. Rev.,* **132,** 1434–1447, doi:10.1175/1520-0493(2004)132<1434:ERIMFS>2.0.CO;2.

Jin, E. K., and Coauthors, 2008: Current status of ENSO prediction skill in coupled ocean–atmosphere models. *Climate Dyn.,* **31,** 647–664, doi:10.1007/s00382-008-0397-3.

Kirtman, B. P., and Coauthors, 2014: The North American Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction; phase-2 toward developing intra-seasonal prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585–601, doi:10.1175/BAMS-D-12-00050.1.

[12] If the details of the distribution of the multimodel forecast members can be trusted to reflect the specific physical circumstances in an individual forecast, then direct use of the members provides additional information about the uncertainty distribution over that provided by the skill-based standard error.

Kumar, A., and Z.-Z. Hu, 2014: How variable is the uncertainty in ENSO sea surface temperature prediction? *J. Climate,* **27,** 2779–2788, doi:10.1175/JCLI-D-13-00576.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.,* **12,** 595–600, doi:10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Peña, M., and H. van den Dool, 2008: Consolidation of multimodel forecasts by ridge regression: Application to Pacific sea surface temperature. *J. Climate,* **21,** 6521–6538, doi:10.1175/2008JCLI2226.1.

Reynolds, R. W., N. A. Rayner, T. M. Smith, D. C. Stokes, and W. Wang, 2002: An improved in situ and satellite SST analysis for climate. *J. Climate,* **15,** 1609–1625, doi:10.1175/1520-0442(2002)015<1609:AIISAS>2.0.CO;2.

Schubert, S. D., M. J. Suarez, P. J. Pegion, R. D. Koster, and J. T. Bacmeister, 2008: Potential predictability of long-term drought and pluvial conditions in the U.S. Great Plains. *J. Climate,* **21,** 802–816, doi:10.1175/2007JCLI1741.1.

Shongwe, M. E., W. A. Landman, and S. J. Mason, 2006: Performance of recalibration systems of GCM forecasts for southern Africa. *Int. J. Climatol.,* **26,** 1567–1585, doi:10.1002/joc.1319.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev.,* **136,** 3933–3946, doi:10.1175/2008MWR2431.1.

——, ——, and S. Li, 2012: Performance of recent multimodel ENSO forecasts. *J. Appl. Meteor. Climatol.,* **51,** 637–654, doi:10.1175/JAMC-D-11-093.1.

——, T. DelSole, and A. G. Barnston, 2014: Reliability of regression-corrected climate forecast. *J. Climate,* **27,** 3393–3404, doi:10.1175/JJCLI-D-13-00565.1.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences.* 2nd ed. Academic Press, 648 pp.