Strength Outlooks for the El Niño–Southern Oscillation

MICHELLE L. L'HEUREUX,[a] MICHAEL K. TIPPETT,[b] KEN TAKAHASHI,[c] ANTHONY G. BARNSTON,[d] EMILY J. BECKER,[a,e] GERALD D. BELL,[a] TOM E. DI LIBERTO,[f,g] JON GOTTSCHALCK,[a] MICHAEL S. HALPERT,[a] ZENG-ZHEN HU,[a] NATHANIEL C. JOHNSON,[h,i] YAN XUE,[a] AND WANQIU WANG[a]

[a] *NOAA/NWS/NCEP/Climate Prediction Center, College Park, Maryland*
[b] *Department of Applied Physics and Applied Mathematics, Columbia University, New York, New York*
[c] *Servicio Nacional de Meteorología e Hidrología del Perú, Instituto Geofísico del Perú, Lima, Peru*
[d] *International Research Institute for Climate and Society, Columbia University, New York, New York*
[e] *Innovim, College Park, Maryland*
[f] *NOAA/Climate Program Office, Silver Spring, Maryland*
[g] *CollabraLink Technologies, Silver Spring, Maryland*
[h] *NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey*
[i] *Atmospheric and Oceanic Sciences Program, Princeton University, Princeton, New Jersey*

(Manuscript received 27 July 2018, in final form 7 November 2018)

ABSTRACT

Three strategies for creating probabilistic forecast outlooks for El Niño–Southern Oscillation (ENSO) are compared. One is subjective and is currently used by the NOAA/Climate Prediction Center (CPC) to produce official ENSO outlooks. A second is purely objective and is based on the North American Multimodel Ensemble (NMME). A new third strategy is proposed in which the forecaster only provides the expected value of the Niño-3.4 index, and then categorical probabilities are objectively determined based on past skill. The new strategy results in more confident probabilities compared to the subjective approach and higher verification scores, while avoiding the significant forecast busts that sometimes afflict the NMME-based objective approach. The higher verification scores of the new strategy appear to result from the added value that forecasters provide in predicting the mean, combined with more reliable representations of uncertainty, which is difficult to represent because forecasters often assume less confidence than is justified. Moreover, the new approach can produce higher-resolution probabilistic forecasts that include ENSO strength information and that are difficult, if not impossible, for forecasters to produce. To illustrate, a nine-category ENSO outlook based on the new strategy is assessed and found to be skillful. The new approach can be applied to other outlooks where users desire higher-resolution probabilistic forecasts, including the extremes.

1. Introduction

Aside from the contribution of climate change, the state of El Niño–Southern Oscillation (ENSO) is the leading source of skill in subseasonal-to-seasonal climate outlooks out to a year (Barnston et al. 2010; Peng et al. 2012). Hence, many national and international climate services provide outlooks for three categories: warm (El Niño), neutral, and cool (La Niña) phases of ENSO. While the indices and thresholds for ENSO vary among the services, outlooks are presented probabilistically, with percentages assigned to the three categories for the coming months or seasons. Each month at NOAA's Climate Prediction Center (CPC), a team of forecasters (who are included as authors on this paper) considers the latest predictions of the Niño-3.4 sea surface temperature (SST) index (area-averaged SST anomaly over 120°–170°W, 5°S–5°N) and issues probabilities for overlapping seasons that cover the next 9 months. Forecast guidance comes in many different forms: often model displays are deterministic (Barnston et al. 2012, 2017), as well as probabilistic (Unger et al. 2009). While probabilities can be developed using individual models, forecasters typically consider those from

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-18-0126.s1.

Corresponding author: Michelle L'Heureux, michelle.lheureux@noaa.gov

DOI: 10.1175/WAF-D-18-0126.1

various multimodel combinations (Tippett et al. 2012; Barnston et al. 2015; Becker and van den Dool 2016). The probabilities from each forecaster are collected and then averaged to form the official NOAA/CPC ENSO probabilistic outlook.

As forecasts started to point to the arrival of the major 2015–16 El Niño (e.g., L'Heureux et al. 2017), there was clear user interest in a "strength outlook" that would not only provide the probability for El Niño occurrence, but also give the likelihood that El Niño would exceed specified thresholds. Such outlooks are critical for differentiating between weaker and more extreme events. Guidance tools are capable of providing strength information, and official updates have sometimes provided qualitative predictions of forecast strength. However, there are challenges inherent in expanding the number of categories in the probabilistic outlooks.

One issue is related to asking human forecasters to expand the number of categories they predict. Even with three-category probabilities, forecasters sometimes struggle to produce a probability distribution that is statistically consistent with what they expect to occur. For instance, one can derive the expected Niño-3.4 value associated with a forecaster's probability forecast because three-category probabilities are sufficient to derive a Gaussian probability distribution function (PDF; see the appendix for details on converting a three-category probability forecast into a Gaussian PDF). However, when the probability of El Niño (or, conversely, La Niña) is near 100%, as in the midst of the strong 2015–16 El Niño, the Gaussian PDF is highly sensitive to the precise probabilities assigned to the three categories. That is, whether the probability of El Niño is 99.9% or 99.99% has a significant impact on the mean and variance of the implied PDF. The problem with this situation is that even the experienced forecaster is unlikely to appreciate all the implications of these probabilities.

Figure 1 illustrates this difficulty, showing two hypothetical distributions with close to 100% chances for El Niño (using NOAA's +0.5°C threshold). In this example, the chances of Niño-3.4 values in excess of +2.0°C (as was observed in 2015–16) can vary anywhere from 86% (area under the blue curve) to 97% (area under the green curve). Also, the implied expected value (the mean) is either +2.9° or +3.8°C. Thus, the choice of a categorical probability to the 10th or 100th of the 99th percentile can have substantial consequences on the probabilities of an extreme and the expected mean value.

Given that it is difficult for forecasters to provide probabilistic information for the extremes, perhaps objective model guidance alone could be used to form the

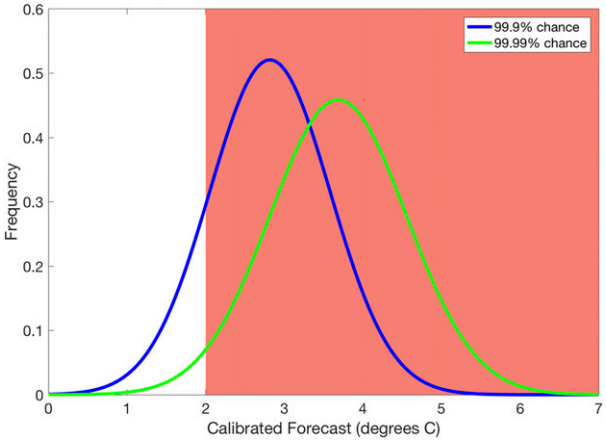

FIG. 1. Gaussian probability distribution functions based on a 99.9% chance of El Niño (blue curve) vs a 99.99% chance (green curve). El Niño is defined by Niño-3.4 index values (shown along the *x* axis) that are greater than or equal to +0.5°C. Red shading under the curves indicates the probability that the Niño-3.4 forecast is in excess of +2.0°C, indicating a major El Niño.

probabilities for ENSO strength outlooks. However, it quickly becomes controversial which model or set of models should be relied upon, and even then, selecting a method to develop the probabilities from the selected model is also not trivial.[1] Putting this issue aside, addressing the question of whether objective forecasts should be prioritized requires comparing them against the performance of subjective forecasts. In addition, we would argue that measures of average skill alone are not adequate because large forecast busts, as well as average performance, are an important consideration for consumers of forecast products. To date, few, if any, verification assessments have gone beyond presenting average skill, and we will go beyond that here.

To navigate these forecasting challenges, we develop a new scheme for probabilistic ENSO outlooks that is based on limited subjective forecaster inputs. Instead of relying on the forecaster to develop probabilities, we propose that the forecaster only be responsible for predicting mean values, and that probabilities be

[1] One might hope that someday the scientific community settles on a single, all-encompassing forecast strategy, but until such a day, forecasters will desire the flexibility to consider multiple forecast inputs. Model developers and statisticians enable this approach by continually creating updated model(s) and new forecast tools. At any given time, there are several model suites and combinations that are favored by forecasters over others, but these evolve and get supplanted by other methods over time. So, the role of forecasters is often akin to the role of jurors, facing multiple lines of evidence and a requirement to come up with the most defensible judgment possible.

derived from the statistics of the variable of interest (in the case of ENSO, the Niño-3.4 SST index) and the estimated forecast skill. Application of this scheme to ENSO forecasts also serves as a proof of concept for other probabilistic outlooks that depend on forecaster inputs, which might be the mean, median, or most likely value (mode). This hybrid approach is arguably intuitive to the forecaster, as most of their model guidance is presented as ensemble means, therefore relieving them of the obligation to characterize the forecast uncertainty. And, for the purposes of developing an ENSO strength outlook, it has the benefit of generating probabilities of exceedance for any specified threshold, including the extremes. While we test these ideas on a record of limited length (41 forecasts), the results of our proposed strategy shows consistency across a number of different events and provide a future pathway for extending the current outlook beyond three categories.

The data and methods are discussed in section 2, forecast verifications are presented in section 3, and section 4 provides a discussion and ideas for future work.

2. Data and methods

a. Forecast strategy

Observed monthly values of the Niño-3.4 index are nearly Gaussian distributed, passing the Kolmogorov–Smirnov test at the 5% level (see Fig. S1 in the online supplemental material). The mean and variance are required to specify a Gaussian distribution. In the proposed approach, the forecasters are responsible for supplying the mean Niño-3.4 index value for each forecast lead based on their assessment of available dynamical and statistical models. The forecast standard deviation σ is taken to be

$$\sigma = \sigma_{\text{climo}} \times \sqrt{(1 - r^2)}, \tag{1}$$

where r is the Pearson's correlation coefficient between the predictions and their corresponding observations, and σ_{climo} is the observed climatological standard deviation. So far, research suggests that most of the variation in uncertainty is across leads and starts/targets, and not from one year to another (e.g., Kumar and Hu 2014). Therefore, it is reasonable to assume that forecasts and observations for a given target and lead are joint-Gaussian distributed. Thus, the forecast correlation skill can be used to adjust the observed (or climatological) standard deviation to match the implied forecast standard deviation. This variance adjustment reflects the fact that the forecast uncertainty tends to increase with longer lead times and that there are months when Niño-3.4 prediction is more difficult (e.g., associated with the spring predictability barrier).

In this instance, the climatological, seasonally varying standard deviation of the Niño-3.4 index is computed over the 1982–2010 period. The adjusted standard deviation is computed for each start month and across forecast lead times. This adjustment has the effect of narrowing the probability distribution function proportionately to forecast skill (or confidence). For starts and leads where r is zero or close to zero, the forecast distribution will be close to the climatological distribution. Though not encountered in this analysis, if the forecast skill were negative, the distribution would typically be set to climatology. Here, the forecast skill is estimated using the 1982–2010 hindcasts of the Niño-3.4 index from the North American Multimodel Ensemble (NMME; see more details below). Therefore, with only the forecaster input of the mean Niño-3.4 index, probabilities can be obtained for any number of categories or for exceeding any specified threshold. Within this paper, this proposed forecast strategy will be referred to as CPCcalib.

Often, when calibrating a model forecast, the mean of the distribution is also adjusted by regressing the ensemble mean forecast against the observations. In principle, forecasts with near-zero skill should also have a forecast mean (signal) near zero. Within this paper, the NMME ensemble mean anomalies are used (see discussion below), but regression adjustment for the mean has not been employed in the interest of a more direct comparison between the spread calibration applied to the forecaster mean and the spread calibration applied to NMME. To help justify this simplification, Barnston et al. (2017; see their Fig. 11) demonstrated that the NMME is mostly free of amplitude biases, so a regression adjustment would likely result in only small changes.

b. Model and observational data

The North American Multimodel Ensemble comprises eight coupled ocean–atmosphere models: GFDL-CM2p1-aer04, NASA-GMAO-062012, COLA-RSMAS-CCSM4, GFDL-CM2p5-FLOR-A06, GFDL-CM2p5-FLOR-B01, CMC1-CanCM3, CMC2-CanCM4, and NCEP CFSv2 (Kirtman et al. 2014). Most of these models are initialized either toward the end of each month or near the beginning of the following month and are run forward up to 12 months. The one exception is the CFSv2 hindcast, which initializes four members every 5 days, for a total of 24 forecasts over the span of the month. The real-time runs of CFSv2 (after 2011), however, are initialized within the first week of each month. Herein, the 0-lead forecast refers to the first season following the

initialization: so for example, for runs made in early January, 0 lead is the average of January–March.

Observed Niño-3.4 index values are based on version 5 of the NOAA Extended Reconstructed Sea Surface Temperature dataset (ERSST; Huang et al. (2017)). This dataset is the basis for NOAA's official ENSO index, the oceanic Niño index, which is the 3-month running average of Niño-3.4 index values (Kousky and Higgins 2007). Because there are trends in the Niño-3.4 index (L'Heureux et al. 2013), departures are computed based on rolling 30-yr monthly average periods that are updated every 5 years.

The correlation coefficient between the observed Niño-3.4 index and the NMME-predicted index is determined using the hindcast (1982–2010) ensemble mean from ~100 equally weighted members. Figure 2 shows that skill varies as a function of lead time and start month. For most models the seasonal cycle and mean biases are removed by subtracting out the lead-dependent, monthly mean climatology from 1982 to 2010. However, because of a discontinuity in the ocean initial conditions around 1999 (Xue et al. 2011), the CFSv2 and CCSM4 models use a 1999–2010 base period (Tippett et al. 2017; Barnston et al. 2017). Both the model data and observations are averaged across overlapping 3-month seasons, so for January–March, February–April, March–May, etc.

c. *Forecast verification*

Two probabilistic verification measures are used to test the quality of the forecasts: the ranked probability skill score (RPSS) and logarithmic skill score (LSS). The RPSS evaluates the sum-squared error of cumulative forecast probabilities, while LSS is the log of the probability for the verifying category. RPSS evaluates the entire probability distribution function (inclusive of all categories), whereas LSS is a local score, meaning the probabilities given to categories other than the verifying one do not affect the score.

Both RPSS and LSS are computed relative to a reference forecast, which here is the climatological probabilities computed over 1982–2017. Positive scores indicate skill that is better than a climatology forecast, scores at zero have the same skill as a climatology forecast, and negative scores mean the forecast scored worse than the climatology forecast. Higher scores are achieved if the forecaster assigns higher probabilities to the category that verifies, especially if the climatological chance of occurrence is very low. LSS and RPSS are "strictly proper," which indicates forecasters maximize their skill score by issuing forecasts that match their true beliefs, rather than by "playing the system" in any way (Murphy and Winkler 1971).

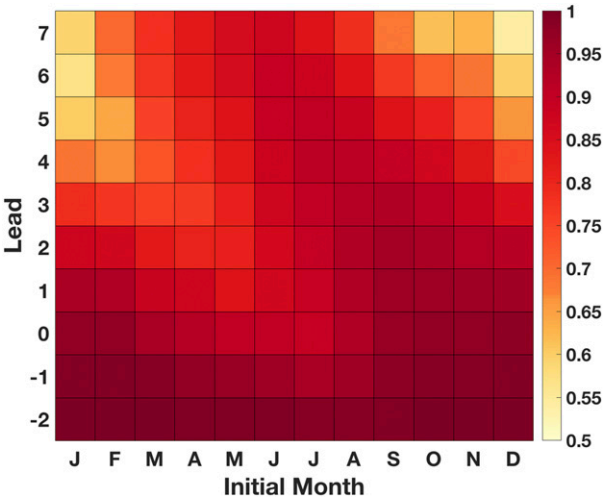

FIG. 2. Anomaly correlation between the ensemble mean NMME seasonal predictions (based on the 1982–2010 hindcasts) and the observed (ERSSTv5) seasonal Niño-3.4 index. The *x* axis shows the month the forecasts are initialized, and the *y* axis is the forecast lead time (for overlapping seasons). For an initial month of January, −2 lead corresponds to the November–January average, where November and December are computed from observations and January is a forecast. Lead-zero and beyond predictions are based entirely on forecasts and, for a January start, 0 lead is the January–March average.

One of the characteristics of LSS is that it is a measure of information and increases with the number of categories that are skillfully forecast, such as forecasting for various ENSO strengths (Tippett et al. 2017). In other words, skillful predictions made for nine forecast categories will achieve higher LSSs than forecasts of comparable skill made for three categories. RPSS is the weighted average of the Brier skill scores of the cumulative probabilities (Bradley and Schwartz 2011), so that skillful forecasting, even with an increasing number of forecast categories, does not significantly change RPSS.

At NOAA/CPC, the forecasters have been providing three-category ENSO probabilities since January 2012. For internal, test purposes, they have been providing the mean values of Niño-3.4 (°C) since July 2015. For comparison between these two methods and, also, the NMME, only the period from June 2015 to September 2018 is evaluated, for a total of 41 forecasts. Figures in the online supplemental material contain a comparison between available methods from January 2012 to the present.

3. Results

a. *Three-category outlooks*

The probabilities produced by the forecasters (CPCoff) are compared with the proposed forecast strategy

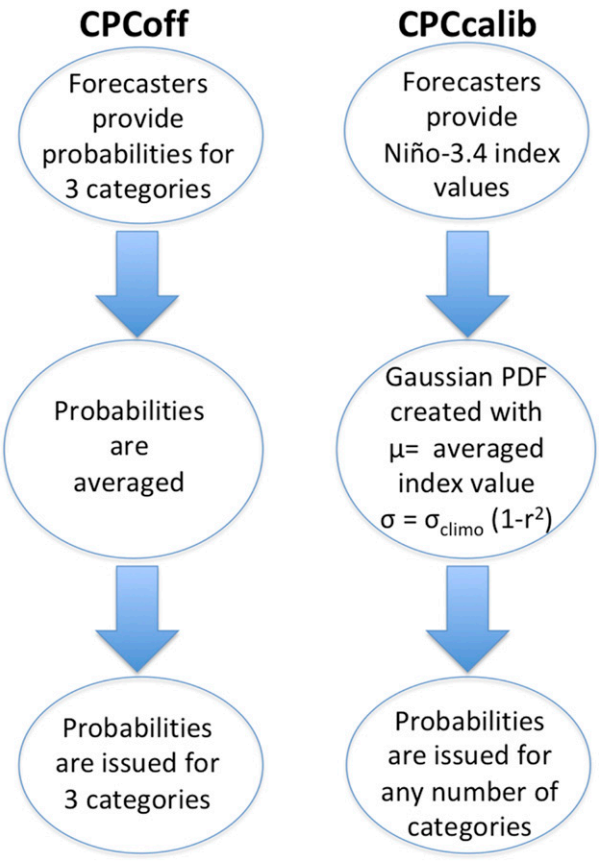

FIG. 3. Schematic of the CPCoff and CPCcalib forecast strategies.

(CPCcalib), using −0.5°/+0.5°C as thresholds in Niño-3.4 (see schematic in Fig. 3). Probabilities are also presented for the calibrated NMME (NMMEcalib), which is computed exactly as in CPCcalib, but using the multimodel's ensemble mean instead of the forecaster mean. Using these three different forecast strategies, Fig. 4 shows the probabilities for La Niña (Niño-3.4 ≤ −0.5°C), neutral (−0.5°C < Niño-3.4 < +0.5°C), and El Niño (Niño-3.4 ≥ +0.5°C) conditions for the −2-lead, 0-lead, 3-lead, and 7-lead forecasts. At all forecast lead times, probabilities from CPCoff (green line) generally have lower variance than those from NMMEcalib (orange line) or CPCcalib (black line). Most of the time the NMMEcalib or CPCcalib approaches have more extreme probabilities than CPCoff. In other words, the probabilities for NMMEcalib or CPCcalib deviate more from their climatological frequencies, while probabilities for CPCoff are more muted and tend to be more conservative than those from the other methods. However, there are times when CPCoff has more aggressive probabilities that are closer to 0% or 100%, but these times appear to occur only when one or both of the other two strategies are similarly extreme.

Figures 5 and 6 display the RPSS (left panels) and LSS (right panels) verifications from June 2015 to September 2018 of the three-category ENSO outlooks. From June 2015 to June 2016, a strong El Niño was conducive to very skillful forecasts among all forecast strategies (Fig. 5), which correctly gave chances close to 100% (Fig. 4). With the decay of El Niño and the return to ENSO-neutral conditions during the summer of 2016, the skill scores of all three approaches were negative across all lead times. However, the NMMEcalib struggled noticeably and had the most negative skill scores compared to the other methods (Fig. 5). This negative skill is in large part due to NMMEcalib assigning high probabilities for the rapid onset of La Niña during summer 2016 even at short lead times (see 0- and 3-lead panels in Fig. 4). The RPSS and LSS measures are especially unforgiving of high forecast probabilities for an incorrect category, and their values are negatively skewed.

Once again in the late spring and summer of 2017, the NMMEcalib had significantly negative skill scores (Fig. 5). This time, NMMEcalib was favoring higher chances, relative to the other approaches, for El Niño in mid-2017, yet ENSO remained neutral during this period. Prediction errors arose across all three strategies, but were especially severe for NMMEcalib for long-lead forecasts made for targets toward the last quarter of 2017 when La Niña returned and yet probabilities for La Niña remained low (see 3- and 7-lead panels in Fig. 4).

Due to these strongly negative LSS and RPSS values, the mean NMMEcalib scores are mostly negative and generally lower than those of CPCcalib and CPCoff, presented by lead in Fig. 6 (top row). However, upon closer inspection, NMMEcalib frequently outscores CPCoff (more positive RPSS and LSS) in part because the NMMEcalib probabilities are more confident overall. This feature of NMMEcalib skill is therefore more apparent in the median score than the mean score (Fig. 6, second row). Generally, CPCoff has lower median scores, which is likely due to the more conservative probabilities assigned by the forecaster. However, the higher probabilities associated with NMMEcalib clearly have a downside when they are assigned to the incorrect verifying category, as demonstrated by the lower minimum RPSS and LSS values (Fig. 6, third row).

Analogous to the Sharpe ratio (Sharpe 1994), the bottom row in Fig. 6 shows the ratio of the mean skill score over the standard deviation. The Sharpe ratio, or reward-to-variability ratio, is often used to track the performance of stocks and mutual funds. For portfolios with a similar level of financial return, the portfolio with higher volatility will have a lower ratio. Therefore, in

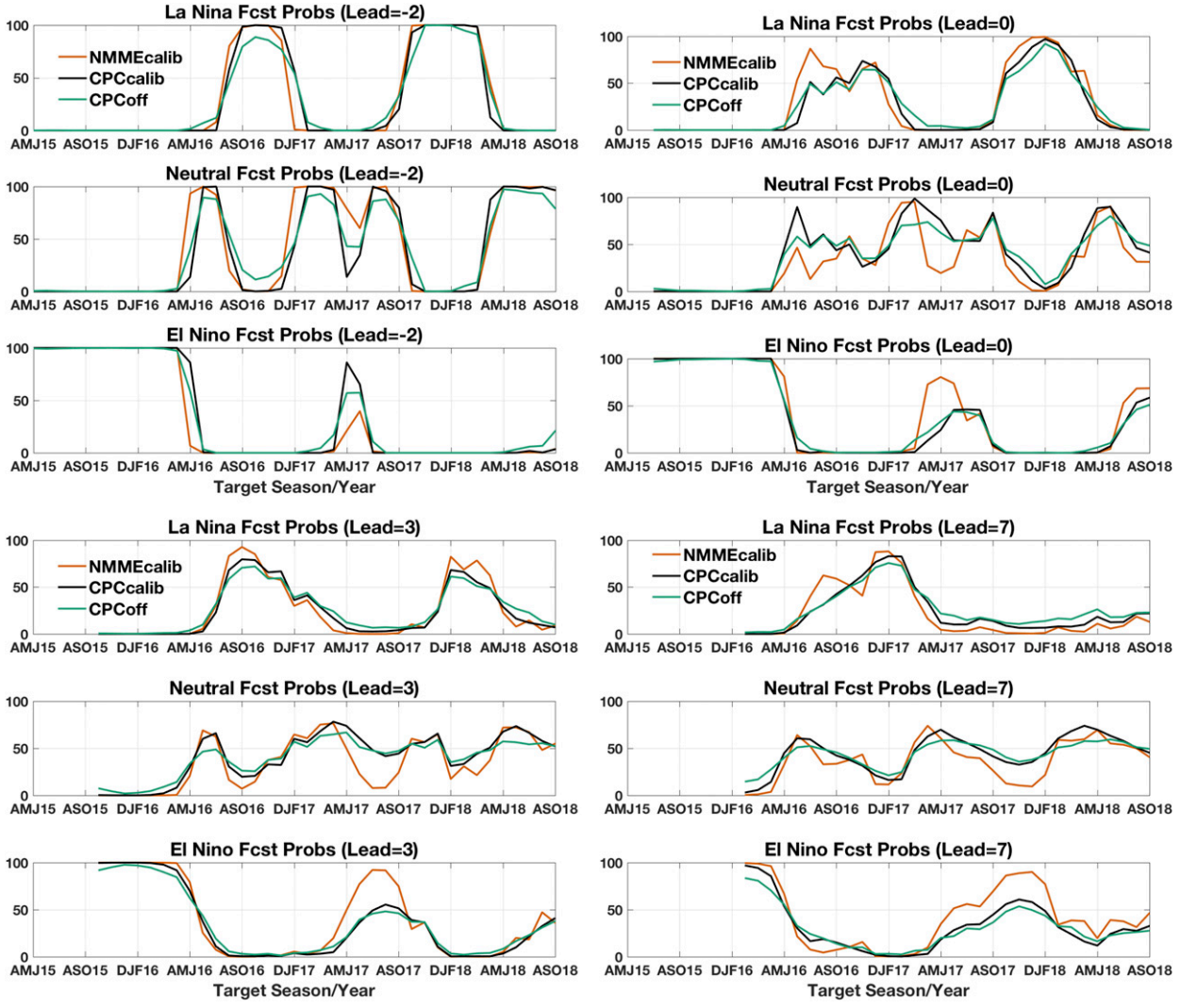

FIG. 4. Predicted probabilities for the three-category ENSO outlooks using the NMMEcalib (orange), CPCcalib (black), and CPCoff (green) forecasting strategies. Within each set of panels, (top) La Niña (Niño-3.4 ≤ -0.5°C), (middle) ENSO neutral (-0.5°C $<$ Niño-3.4 $< +0.5$°C), and (bottom) El Niño (Niño-3.4 $\geq +0.5$°C) conditions are shown. The target season of the forecast is displayed by the centered month along the x axis. The top-left set of panels displays the -2-lead forecast, the top-right set shows the 0-lead forecast, the bottom-left set displays the 3-lead forecast, and the bottom-right set shows the 7-lead forecast. Forecasts are made each month from June 2015 to September 2018.

order to increase returns, a high-variability portfolio needs to compensate with a higher rate of return. In the current application, a user may desire higher skill scores overall, but if that skill is also accompanied with a higher risk of large errors, then such a strategy may not be beneficial. Given the higher variability in the LSS and RPSS scores for NMMEcalib (Fig. 5), the ratio for NMMEcalib tends to be lower than CPCcalib and CPCoff across nearly all forecast leads (except for -2 lead).

Overall, CPCcalib has several attractive attributes that subsume the best qualities of the other two approaches. It allows forecasters to avoid the large forecast busts of NMMEcalib, yet have better overall skill than CPCoff. In other words, CPCcalib simultaneously provides the forecaster more courage in the form of higher probabilities in situations where it is appropriate to do so, yet guards against foolishly high probabilities in scenarios that deserve more caution. CPCcalib provides a higher ratio with a reduction in negative RPSS/LSS scores, along with higher mean and median scores that compare well against the most skillful approaches. Further, CPCcalib removes the burden of estimating uncertainty, which is a challenge for forecasters considering multiple models

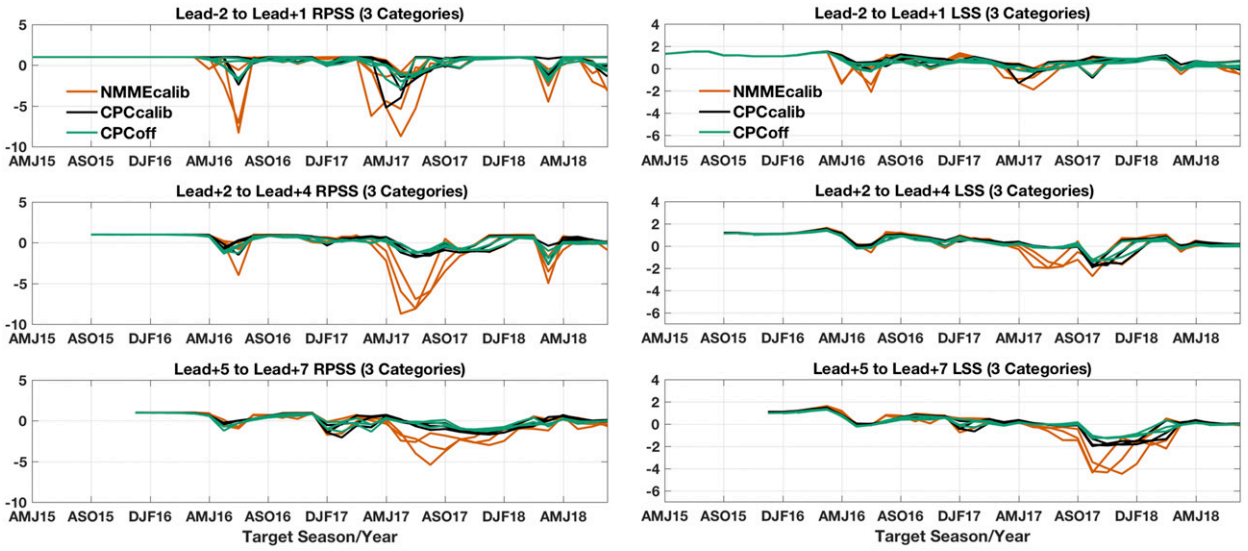

FIG. 5. (left) RPSSs and (right) LSSs for three-category ENSO outlooks presented as a function of target season (centered month displayed along the *x* axis) for the NMMEcalib (orange), CPCcalib (black), and CPCoff (green) forecasting strategies. Each line shows forecast leads (top) from −2 to +1, (middle) from +2 to +4, and (bottom) from +5 to +7. Forecasts are made each month from June 2015 to September 2018.

and model combinations, all with varying skill levels and characteristics.

Finally, in a longer record going back to January 2012 (CPCcalib is not available prior to June 2015), the skill characteristics are generally reproduced (Figs. S2 and S3), with the exception that the median RPSS and LSS values are higher for CPCoff than NMMEcalib for lead 4 and beyond.

b. Nine-category outlooks

We present verifications for nine categories in Figs. 7 and 8 to evaluate whether the forecast process is skillful at finer resolutions. These categories are based on 0.5°C increments in the Niño-3.4 index (±2.0°, ±1.5°, ±1.0°, ±0.5°C), which correspond to informal strength thresholds of ENSO. Because the forecaster does not create probabilities at such fine resolution, only NMMEcalib and CPCcalib are compared. The median RPSS values for both strategies are very similar to the three-category outlook, with hardly any discernible change at all. The only noticeable change in RPSS going from three to nine categories is the reduction in the mean and minimum values associated with NMMEcalib (Figs. 5 and 6). For a forecast with the same skill, a slight decline in RPSS can occur with the addition of more categories as shown by the joint-Gaussian model of Tippett et al. (2017). In contrast, CPCcalib is very similar to its three-category result, so the addition of categories does not noticeably change the RPSS.

For LSS, however, a more promising result emerges, as shown by a slight increase with the addition of more categories, which is particularly noticeable at shorter leads. The mean, median, and maximum LSSs mostly increase for CPCcalib and NMMEcalib compared to three categories. The one conspicuous exception is the miss in the −2-lead CPCcalib forecast that occurs for the August–October 2015 target (forecast made in early October 2015), which was in excess of +2.0°C and was not anticipated. Overall, as Tippett et al. (2017) demonstrated using NMME, the LSS measure indicates that ENSO forecasts for narrowly defined categories, which include the extremes, have more skill than wider categories for some leads and targets.

4. Discussion

The three strategies are compared only over a fairly short duration, but contain the major 2015–16 El Niño, two subsequent La Niñas in 2016–17 and 2017–18, and intervening periods of ENSO neutral conditions. We expect that this period is representative of forecaster tendencies over a longer record, but to increase confidence, we plan to continue to track these strategies over an extended period and revisit these results. Yet it is encouraging that the proposed forecast strategy is already paying dividends over such a short period. An attractive aspect of CPCcalib is that it can provide probabilistic predictions for any threshold or extreme, and not just for the Niño-3.4 index, but for any variable.

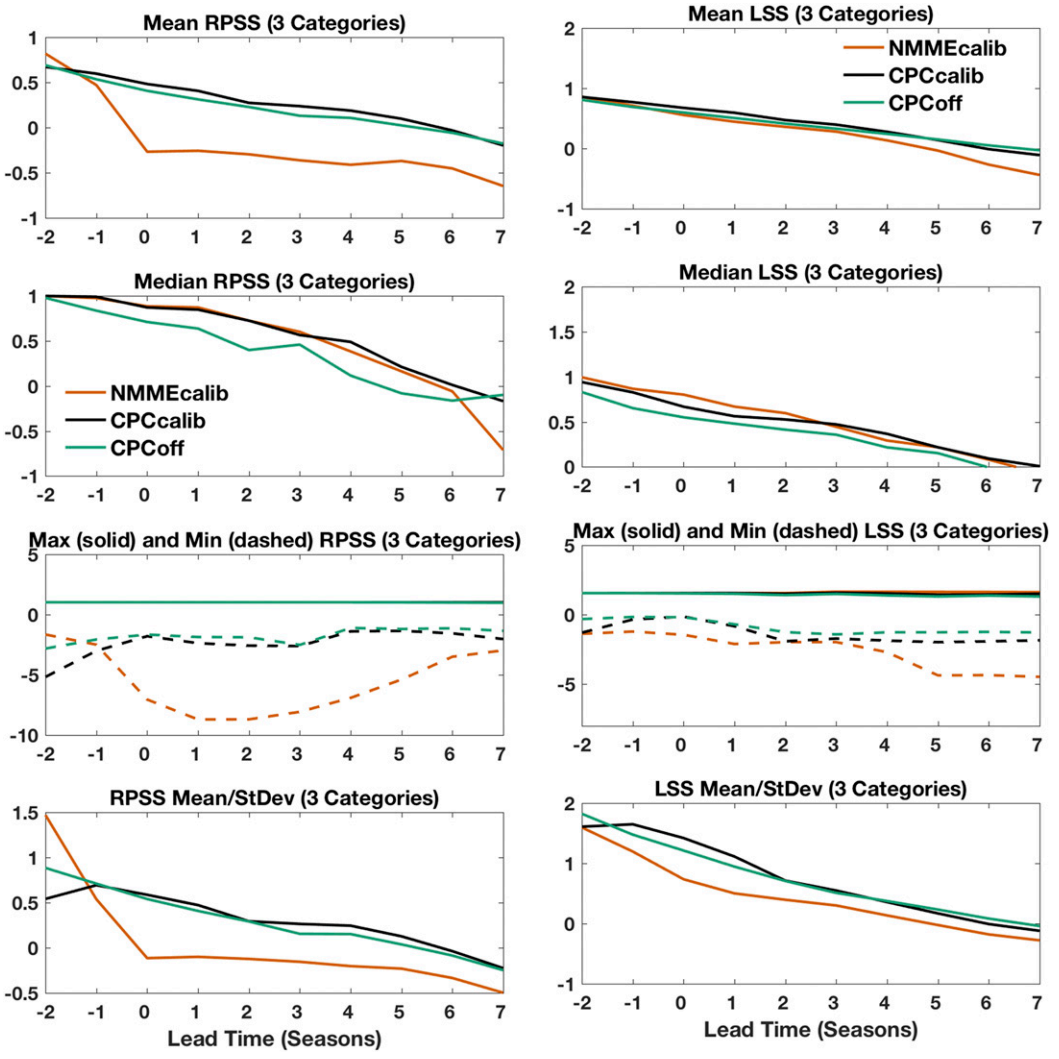

FIG. 6. As in Fig. 5, but RPSSs and LSSs are displayed as a function of lead time, from −2 to 7 lead. (top) Mean skill scores, (second row) median skill scores, (third row) the maximum scores (solid lines) and minimum scores (dashed lines), and (bottom) the ratio of the mean skill score over the standard deviation.

For variables that are non-Gaussian (e.g., Niño-1+2 index, precipitation), a different distribution or a transformation can be used to generate the probabilities. Other strategies to arrive at the forecast probabilities from forecaster inputs can also be tested and employed, such as using Bayes's theorem. For a Gaussian distribution, such as that of Niño-3.4, the application of Bayes's theorem is equivalent to linear regression, so this approach has already been implicitly accounted for in this study.

Overall, it appears that when forecasters are relied upon to develop probabilities from model inputs (CPCoff), they tend to err toward more conservative probabilities. This is not necessarily a bad strategy, as it reduces the frequency of large errors, but it comes at the cost of higher skill scores in the median. In other words, forecasters provide less confident probabilities, which is suitable in situations where there should be less confidence (e.g., long-lead forecasts made early in the year). However, the less confident probabilities more often result in lower RPSSs and LSSs. Perhaps this is not altogether surprising because of ''loss aversion'' biases in which people prefer avoiding losses more than acquiring comparable gains (Kahneman and Tversky 1979).

However, on the flip side of the coin, we demonstrate that using the NMMEcalib strategy, which does not rely on forecaster adjustment, can result in probabilities that are too high, even in situations where it may be more prudent to scale them back. Interestingly, the overconfidence of NMMEcalib is not due to underestimating uncertainty but to overestimating signal (shifts in the

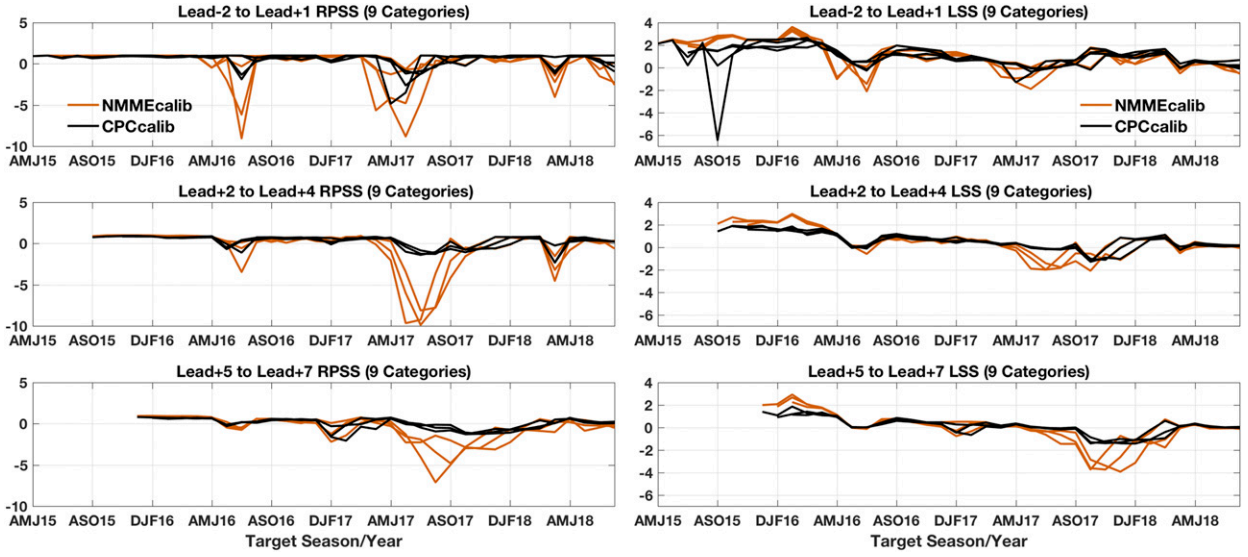

FIG. 7. As in Fig. 5, but for nine-category ENSO outlooks. CPCoff was not generated for nine categories, so only NMMEcalib (orange) and CPCcalib (black) are displayed.

mean). But, our three-category results also show that NMMEcalib scores higher overall than CPCoff in the median (notably, not in the mean). No doubt this in part due to more frequent assignment of higher probabilities in the verifying category, which result in higher RPSS and LSS values. This may be a desirable quality for certain users, but others may be troubled by the higher volatility (and lower mean to standard deviation ratio) in the skill scores associated with NMMEcalib.

Therefore, an ideal forecast strategy would be to guard against large forecasts errors (as in CPCoff), while increasing the probabilities (NMMEcalib), especially when the forecaster should be more confident. Our results suggest the CPCcalib forecast strategy is a step toward this ideal. Why is this the case? We can only speculate given the subjectivity of human involvement, but the model guidance that the forecasters consider is typically presented as ensemble means. Even when individual members are available, it is not as intuitive to optimally combine the data to arrive at the forecast probabilities. Several sets of probabilistic guidance are also considered each month, but again, it is not clear to the forecaster how to optimally synthesize all of this information. Regardless, it appears forecasters are quite capable of assimilating large amounts of forecast guidance to provide an appropriate value for the mean, but where they struggle is in representing the uncertainty.

This study carries with it the implication that forecasters are capable of providing confident forecast signals, but should not be relied upon to predict the spread and come up with probabilities. The fact that the CPCcalib has higher median/mean LSSs and RPSSs than CPCoff is an indicator that forecasters tend to imagine the spread is wider than it really is. By adopting the CPCcalib strategy, the forecaster is also relieved of the responsibility to properly adhere to the probability distribution, and does not have to generate forecast probabilities for a specific cutoff or number of categories. It is difficult to envision a forecaster reliably creating probabilities for nine ENSO categories, so it is encouraging that the CPCcalib approach is capable of skillfully providing strength information.

Acknowledgments. ML, EB, TD, and NJ are grateful for support from the NOAA Climate Programs Office (CPO)/Climate.gov for the ENSO Blog. KT is thankful for the support of UCAR's Cooperative Programs for the Advancement of Earth System Science (CPAESS) and the NOAA/National Centers for Environmental Prediction (NCEP) Short-term Research Collaboration Program.

APPENDIX

Converting a Three-Category Probability Forecast into a Gaussian PDF

For any three-category probability forecast, there is a Gaussian distribution with matching probabilities. Suppose that P_b is the forecast probability of being below the threshold x_b, and P_A is the forecast probability of exceeding the threshold x_a. The Gaussian forecast distribution with mean μ_f and variance σ_f^2 has matching probabilities if

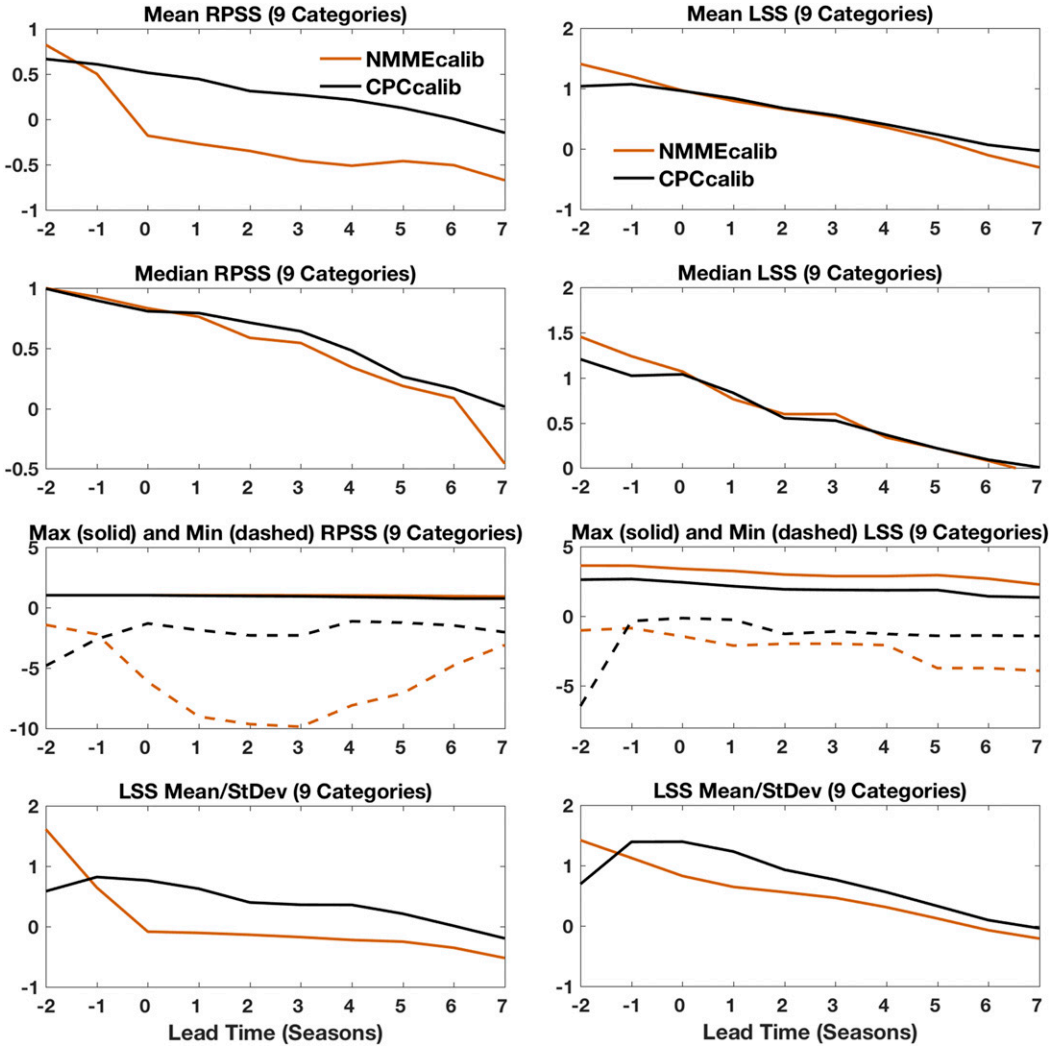

FIG. 8. As in Fig. 6, but for nine-category ENSO outlooks.

$$P_b = \Phi\left(\frac{x_b - \mu_f}{\sigma_f}\right), \qquad P_a = 1 - \Phi\left(\frac{x_a - \mu_f}{\sigma_f}\right), \tag{A1}$$

where Φ is the cumulative distribution of a Gaussian with mean zero and unit variance. Equivalently,

$$\Phi^{-1}(P_b) = \frac{x_b - \mu_f}{\sigma_f}, \qquad \Phi^{-1}(1 - P_a) = \frac{x_a - \mu_f}{\sigma_f}, \tag{A2}$$

which are two linear equations for μ_f and σ_f, whose solutions are

$$\mu_f = \frac{\Phi^{-1}(1 - P_a)x_b - \Phi^{-1}(P_b)x_a}{\Phi^{-1}(1 - P_a) - \Phi^{-1}(P_b)}, \tag{A3}$$

and

$$\sigma_f = \frac{x_a - x_b}{\Phi^{-1}(1 - P_a) - \Phi^{-1}(P_b)}. \tag{A4}$$

REFERENCES

Barnston, A. G., S. Li, S. J. Mason, D. G. DeWitt, L. Goddard, and X. Gong, 2010: Verification of the first 11 years of IRI's seasonal climate forecasts. *J. Appl. Meteor. Climatol.*, **49**, 493–520, https://doi.org/10.1175/2009JAMC2325.1.

——, 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, https://doi.org/10.1175/BAMS-D-11-00111.1.

——, ——, H. M. van den Dool, and D. A. Unger, 2015: Toward an improved multimodel ENSO prediction. *J. Appl. Meteor. Climatol.*, **54**, 1579–1595, https://doi.org/10.1175/JAMC-D-14-0188.1.

——, ——, M. Ranganathan, and M. L. L'Heureux, 2017: Deterministic skill of ENSO predictions from the North American Multimodel Ensemble. *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3603-3, in press.

Becker, E., and H. van den Dool, 2016: Probabilistic seasonal forecasts in the North American Multimodel Ensemble: A baseline skill assessment. *J. Climate*, **29**, 3015–3026, https://doi.org/10.1175/JCLI-D-14-00862.1.

Bradley, A. A., and S. S. Schwartz, 2011: Summary verification measures and their interpretation for ensemble forecasts. *Mon. Wea. Rev.*, **139**, 3075–3089, https://doi.org/10.1175/2010MWR3305.1.

Huang, B., and Coauthors, 2017: Extended Reconstructed Sea Surface Temperature, version 5 (ERSSTv5): Upgrades, validations, and intercomparisons. *J. Climate*, **30**, 8179–8205, https://doi.org/10.1175/JCLI-D-16-0836.1.

Kahneman, D., and A. Tversky, 1979: Prospect theory: An analysis of decision under risk. *Econometrica*, **47**, 263–291, https://doi.org/10.2307/1914185.

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

Kousky, V. E., and R. W. Higgins, 2007: An alert classification system for monitoring and assessing the ENSO cycle. *Wea. Forecasting*, **22**, 353–371, https://doi.org/10.1175/WAF987.1.

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

L'Heureux, M. L., D. C. Collins, and Z.-Z. Hu, 2013: Linear trends in sea surface temperature of the tropical Pacific Ocean and implications for the El Niño–Southern Oscillation. *Climate Dyn.*, **40**, 1223–1236, https://doi.org/10.1007/s00382-012-1331-2.

——, and Coauthors, 2017: Observing and predicting the 2015/16 El Niño. *Bull. Amer. Meteor. Soc.*, **98**, 1363–1382, https://doi.org/10.1175/BAMS-D-16-0009.1.

Murphy, A. H., and R. L. Winkler, 1971: Forecasters and probability forecasts: Some current problems. *Bull. Amer. Meteor. Soc.*, **52**, 239–248, https://doi.org/10.1175/1520-0477(1971)052<0239:FAPFSC>2.0.CO;2.

Peng, P., A. Kumar, M. S. Halpert, and A. G. Barnston, 2012: An analysis of CPC's operational 0.5-month lead seasonal outlooks. *Wea. Forecasting*, **27**, 898–917, https://doi.org/10.1175/WAF-D-11-00143.1.

Sharpe, W. F., 1994: The Sharpe ratio. *J. Portfolio Manage.*, **21**, 49–58, https://doi.org/10.3905/jpm.1994.409501.

Tippett, M. K., A. G. Barnston, and S. Li, 2012: Performance of recent multimodel ENSO forecasts. *J. Appl. Meteor. Climatol.*, **51**, 637–654, https://doi.org/10.1175/JAMC-D-11-093.1.

——, M. Ranganathan, M. L'Heureux, A. G. Barnston, and T. DelSole, 2017: Assessing probabilistic predictions of ENSO phase and intensity from the North American Multimodel Ensemble. *Climate Dyn.*, https://doi.org/10.1007/s00382-017-3721-y, in press.

Unger, D. A., H. van den Dool, E. O'Lenic, and D. Collins, 2009: Ensemble regression. *Mon. Wea. Rev.*, **137**, 2365–2379, https://doi.org/10.1175/2008MWR2605.1.

Xue, Y., B. Huang, Z.-Z. Hu, A. Kumar, C. Wen, D. Behringer, and S. Nadiga, 2011: An assessment of oceanic variability in the NCEP Climate Forecast System Reanalysis. *Climate Dyn.*, **37**, 2511–2539, https://doi.org/10.1007/s00382-010-0954-4.