Final Report for Grant: NA12OAR4310082

Developing an Optimum Multi-model ENSO Prediction

Anthony Barnston[1], Michael Tippett[1], Huug van den Dool[2], and David Unger[2]

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

[2]Climate Prediction Center, National Centers for Environmental Prediction, National Weather Service, National Oceanic and Atmospheric Administration, College Park, Maryland

Period of Activity: August 1, 2012 to July 31, 2016

Table of Contents

1. Purpose……..2

2. Highlights of Accomplishments……..2

3. Publications and Reports……..8

3.1 Publications by Principal Investigators……..8

3.2 Other Relevant Publications……..8

4. PI Contact Information……..9

5. Slides……..10

Purpose

The purpose of the project is to investigate ways to improve the reliability and usability of the ENSO prediction plume product that began being issued at the International Research Institute for Climate and Society (IRI) in early 2002 (Barnston et al. 2012). Improvements are expected through corrections of individual model mean bias and amplitude, and through refinement of the multimodel combining method. Finally, improvements will come through the formulation of a more realistically defined and easily understood probability distribution, including the user-friendliness of the format of the issued product(s).

Results and Accomplishments

Research Phase

The initial task in the project was obtaining hindcast data for a starting set of ENSO prediction models, including individual ensemble members for the dynamical models. It was decided to use the ensemble data from the models participating in the North American Multimodel Ensemble project (see Table 1). The hindcast data for the Nino3.4 index were created from these models and used for experiments targeting the above-mentioned improvements.

Model	Expanded model name	No. ensemble members
1. CMC1-CanCM3	Canadian coupled model #1	10
2. CMC2-CanCM4	Canadian coupled model #2	10
3. COLA-RSMAS-CCSM3	COLA/Univ. Miami/NCAR coupled model	6
4. GFDL-CM2pl-aer04	Modified version of GFDL coupled model	10
5. NASA-GMAO-062012	Modified version of NASA coupled model	12
6. NCEP-CFSv2	NOAA/NCEP coupled model	24

Table 1. List of models whose hindcasts are used in the initial research toward an improved ENSO prediction plume.

A set of candidate methodological frameworks, and verification measures, were established so that experiments varying these would provide quality comparisons needed to make the set of decisions.

A few of the basic design decisions for which the hindcast skill experiments were aimed, are:

- Use of the NMME hindcast period of January 1982 to December 2010
- Use of the 1˚ observed Reynolds and Smith OI-SST data set for verification

- Use of 1-month averaging periods.
- For verification, cross-validation (c-v) design will leave out 1 year; non-c-v also will be used where appropriate
- Verification scores will include temporal correlation and RMSE for deterministic forecasts, and RPSS if probabilistic forecasts are verified.

The following list contains the specific methodologies for testing relative skills:

- Value of weighting models by historical hindcast skill (correlation)
- Choice of weighting by skill regardless of model-to-model forecast correlation, versus using partial correlations (as in multiple regression)
- Whether (and how) to bias-correct individual models prior to consolidation
- Role of ensemble number in weighting of individual models for forecast mean; should ensemble size be used as a weighting factor?
- Method to establish uncertainty: hindcast skill (and its standard error) versus ensemble member spread for the forecast probability distribution

In addition to determining the methodologies that deliver best predictive skill, another goal is to produce an ENSO prediction plume that most effectively shows the probabilistic predictions. Toward that end, we are experimenting with various choices of graphical formats.

Initial experiments indicate, as also determined earlier by Tippett and Barnston (2008) and more recently by Delsole et al. (2013), that hindcast skill-based weighting of models usually does not help cross-validated skill when 30 or fewer years of hindcasts are available. The reason is that when the model skills vary by the amounts seen here for the NMME models—not sharply—skill differences are most likely related to sampling variability rather than to true model quality differences.

The next, and perhaps most important, finding is the presence of significant mean model biases that vary widely among models, and often across forecast start times and lead times within an individual model. Some model producers do not remove the model's own climatology from the forecasts to form anomalies, but may remove the observed climatology, leaving a substantial mean model bias. The effect of these varying biases is an increase in the spread of the model ensemble means, as often shown in the existing ENSO prediction plume. Figure 1 shows the plume issued in July 2013. We see that the model disagreements at short lead times, such as 1-3 months, appear larger than expected, considering the skill levels at those leads. Figure 2 shows root-mean squared error (RMSE) skill score for the NMME ENSO forecast first without any mean bias corrections for individual models, then with bias correction, and finally with both bias and amplitude correction. The mean bias correction achieves a very noticeable improvement in the skill. Forecast amplitude is represented by the temporal standard deviation of the forecasts over the years, for a fixed start time and lead time. Correction of amplitude so that it equals the interannual standard deviation of observations multiplied by the cross-validated

hindcast correlation with observations, results in a slight further improvement in the RMSE skill score.

Another finding of the hindcast experiments was that the ensemble spreads of the individual models are approximately constant from year to year for the same season and lead time (confirming the results of Kumar and Hu 2014), but have differing relationships to their respective expected skill, which is based on hindcast verifications. Thus, some models have spreads that are too small considering the uncertainty associated with their expected skill, while other models have spreads more appropriate for their expected skill. This finding suggests that model ensemble spreads may not be able to be taken literally in best estimating the uncertainty of the multi-model ENSO forecast. The fact that individual model ensemble means do contribute some skill, and the multimodel ensemble mean usually has still greater skill and value than the most skillful individual model, has been documented in Kirtman et al. (2014) and earlier studies. Probabilistic reliability is valuable, as the probability distribution is just as important as the best guess single deterministic forecast. The ensemble mean is expected, on statistical grounds, to be a more skillful forecast when the ensemble sizes of the individual models are allowed to act as weighting factors. Further, toward providing the most useful possible probability distribution, it is believed safer to generate the uncertainty distribution on the basis of the historical cross-validated hindcast skill of the multi-model system than on the basis of the probabilistically less realistic spread of the individual ensemble members of the models. A common method for using the historical skill to determine the optimum uncertainty distribution is to assume a Gaussian distribution. The Gaussian is a reasonable approximation for the distributions of tropical Pacific SST in the Nino4 and Nino3.4 regions, but not for SSTs farther east or in the far western Pacific or Indonesia. The optimum spread is then given by the standard error of estimate (SEE), which is a function of the hindcast correlation skill (cor):

$$SEE = SD_y \sqrt{1 - cor_{xy}^2}$$

Using this formula, the spread of the forecasts of an individual model equals that of the standard deviation of the observations for the given season/lead when there is no forecast skill (i.e., the correlation=0), and becomes smaller as the individual model's historical skill increases. The spread is always symmetric about the ensemble mean and remains the same, year to year, for a given forecast start time and lead time. As noted in the right panel of Fig. 2, amplitude correction is not as important as mean bias correction.

On another research issue, analyses were conducted to determine the best method to formulate the optimum multimodel ensemble mean, regarding amplitude (deviation from average), as well as the width of the forecast probability distribution. Figure 3 shows the ratio of the skill-based standard error of estimate,

shown to the left of the "divide" sign, to the standard deviation of the members of the multi-model ensemble of the ENSO forecast (to right of "divide" sign):

$$SD_y\sqrt{1-cor_{xy}^2} \;/\; SD_{members}$$

Here SD_y is the standard deviation of the observed Nino3.4 index and cor_{xy} is the correlation between the hindcasts and the observations.

For naturally good model calibration, this ratio should be close to 1. The first two panels of Fig. 3 show that the ratio substantially deviates from 1, particularly without bias corrections (left panel), but also for certain target periods near the northern spring ENSO predictability barrier after bias corrections but without amplitude corrections (middle panel). When both types of bias are corrected (right panel), ratios are more uniform across seasons and lead times, but average somewhat lower than 1, particularly for very long lead times (when the CFSv2 model does not have forecasts) and seasons near the middle of the year that are affected by the predictability barrier. These results suggest that using the multi-model ensemble spreads directly lead to less favorable probabilistic reliability than forecasts whose probability distribution are statistically derived from the hindcast skills, consistent with Goddard et al. (2013). Hence, a decision is made to use the hindcast performance-based standard error rather than the spread of the members to develop the forecast probability distribution. This decision is compatible with the finding in Kumar and Hu (2014) that model spread is quasi-constant from forecast to forecast for the same target season, lead time and model—i.e., the noise part of the signal-to-noise ratio remains about the same year to year in seasonal forecasts.

Finding a user-friendly format for the ENSO predictions is an important goal of the project. To produce a forecast diagram that shows the probabilistic predictions most reliably and understandably, we experimented with various choices of graphical formats. Figure 4 shows a possible 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. Beyond the format, the figure shows the effects of mean bias correction and amplitude corrections on the hindcast for the 2009-10 El Nino made in June 2009, when the event was just about to begin. The panels 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. 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 a much improved MME forecast (middle panel) and more realistic widths of the uncertainty distributions. Correction for the amplitude as well as the bias results in slight underestimation of the strength of the event, and underestimation of the amount of uncertainty at short leads. While the amplitude correction may have slightly degraded in this particular forecast, it would be expected to improve the RMSE of the forecasts, on average, in general.

Feedback on the above forecast format indicated that the vertically oriented Gaussian curves are often ineffective in communicating the uncertainty, as many non-climate specialists do not easily understand the meaning of the relative probability density implied by the curves. A format more similar to that of the existing plume, with horizontal lines or interval bands, was found to be more understandable to users.

It is possible to generate a plume of equally likely scenarios for a prediction of the ENSO state from a given starting month, using a Gaussian random number generator, using the MME mean forecast in combination with the historical covariance of the errors over the hindcast period. The 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 (there is a positive correlation of errors between consecutive lead times). Hence, there is a matrix of error covariances for each lead time with each other lead time for any forecast start time, computed using all years in the hindcast period. This ability to generate equally likely realistic forecast scenarios can be used to create "spaghetti diagrams".

Figure 5 shows a number of options for expressing the forecast from June 2009, including its uncertainty distribution. In the upper left panel, the thick line showing the mean forecast is in the middle, among a family of lines showing various percentiles within the forecast distribution: 1, 5, 15, 25, 50, 75, 85, 95, and 99. This provides a wide choice of intervals that may matter most to various users. A similar product is shown in the upper right panel, except the more likely intervals are shaded with increasingly dark color. The two bottom panels both show the forecast mean, the 15th and 85th percentiles, and numerous randomly generated lines showing equally likely individual scenarios (100 on left panel, 200 on right panel). These equally likely forecasts have greatest density near the forecast mean, and lowest far from the mean. The bottom two panels most resemble today's 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. A reason for the better forecast description here, is that the new plots are equivalents to the forecasts of individual model ensemble members, while the lines on today's plume plot are the ensemble means of the various models. The observation will be like a single ensemble member rather than like an ensemble mean.

Analysis of the preferences of users, mostly non-climate scientists, based on informal interrogation, narrows the set of favored forecast formats to the upper right and lower left panels in Fig 5. In the lower left panel 100 lines are randomly generated under the constraints of the given skill level and error covariance, showing 100 equally likely individual scenarios. On that plot, user can get a sense of the probabilities largely by visual inspection, while in the upper right panel there is more explicit guidance on the forecast probabilities.

A paper describing the technical aspects of the optimized prediction plume (Barnston et al. 2015) was published in Journal of Applied Meteorology and

Climatology. The main results evolving from the research of the current grant involved the importance of model mean bias and amplitude corrections, as well as the importance of a user-friendly product format was also introduced.

The final choice of the format of the ENSO prediction forecast, based on the accumulated feedback from various users over at least two years, includes three plots, each showing aspects of the forecast distribution that partially differ from one another.

The first choice (Fig. 6) is a graph showing the ensemble mean forecast of each model, and a thicker line showing the multi-model ensemble forecast. The latter is a straight average (i.e., using equal weighting) of the models. This graph, representing the most basic rendering of the forecast, was understandable to all potential users. It is also essentially the same as the existing plume diagram that has been popular.

The second plot (Fig. 7) summarizes the overall forecast distribution, showing a set of percentile values (1, 5, 15, 25, 50, 75, 85, 95 and 99 percentiles) with descriptive color coding. The 50 percentile matches the multi-model mean forecast shown in Fig. 6. This forecast format choice was valued most highly by users with a college degree having some level of understanding of the probabilistic aspect of the forecast. Many have advanced degrees in their field (e.g. hydrology, agriculture, climate, economics).

The third forecast format (Fig. 8), sometimes called a spaghetti diagram, shows the forecast mean and the 15^{th} and 85^{th} percentiles, along with 100 lines randomly generated under the constraints derived from the multi-model mean forecast and the given skill level and error statistics of the NMME forecast. The error covariance governs the degree of persistence of the deviation of the scenario from the multi-model mean forecast, making the scenarios statistically realistic. The lines represent 100 equally likely forecast scenarios. This forecast format is desired mainly by users who think in terms of possible individual concrete scenarios, sometimes merely for a visual interpretation, and sometimes for purposes of assessing the likelihoods of various outcomes in their application (e.g., water management or agriculture) based on their historical data. The individual scenarios may be regarded as analogues by this set of users, even in the absence of this many actually observed analogues in the data archives. The choice of 100 scenarios is arbitrary, and could have been any number. Using many more than 100 creates a more crowded plot in which the lines near the middle of the distribution mainly overwrite one another and form a uniformly dark color. Figures 6, 7 and 8 are based on the forecast from July, 2016.

Operational Implementation Phase

Following the research phase, a slightly larger set of NMME models had become available for real-time implementation. Additionally, several state-of-the-art non-NMME models were added to the set of models to participate in the improved ENSO prediction plume. The final list of models is shown in Table 2.

Model	Full model name (all are ocean-atmos. Coupled models)
1. CMC1-CanCM3	Canadian coupled model #1
2. CMC2-CanCM4	Canadian coupled model #2
3. COLA-RSMAS-CCSM4	COLA/Univ. Miami/NCAR coupled model: CCSM4 from 2016
4. GFDL-CM2pl-aer04	Modified version (as of 2012) of GFDL coupled model
5. GFDL-CM2.5-FLOR-B	Higher resolution GFDL model, lower ocean resolution
6. NASA-GMAO-062012	Modified version of NASA coupled model
7. NCEP-CFSv2	NOAA/NCEP climate forecast system coupled model
8. ECMWF	European Center, System 4 coupled model
9. UKMO	United Kingdom Met. Office coupled forecast model
10. Meteo France	Meteo France coupled forecast model
11. JMA	Japan Meteorological Agency coupled forecast model
12. JAMSTEC SINTEX-F	Japan Agency for Marine-Earth Sci. & Tech., SINTEX-F model

Table 2. List of coupled comprehensive models whose real-time forecast are included in the improved ENSO prediction plume. The first seven are in the NMME.

Routine real-time issuance of the three above-mentioned plot formats commenced at IRI in June 2016, on the existing schedule on which NCEP/CPC's seasonal forecasts are released (the third Thursday of each month). The web page for accessing them is: http://iri.columbia.edu/our-expertise/climate/forecasts/enso/current/?enso_tab=enso-nmme

If NCEP desires to take on the responsibility of producing this set of monthly ENSO prediction plume products, the Matlab and Fortran software can be provided. While the new plume product is similar to the NMME ENSO plumes issued by Climate Prediction Center and the existing IRI/CPC ENSO prediction plumes, it offers a different perspective on the ENSO forecast through the introduction of the new formats. The new product may include a different set of models over the course of the coming months and years as model improvements or changes occur, or when some models are discontinued or new models are created.

3 Publications and Reports

3.1 Publications by Principal Investigators

Barnston, A. G., M. K. Tippett, H. M. van den Dool, and D. A. Unger, 2015: Toward an improved multi-model ENSO prediction. *J. Appl. Meteor. Climatol.*, **54**, 1579-1595.

3.2 Other Relevant Publications

Barnston, A. G., M. K. Tippett, M. L. L'Heureux, S. Li, and D. G. DeWitt, 2012: Skill of real-time seasonal ENSO model predictions: Is our capability increasing? *Bull. Amer. Meteor. Soc.*, **93**, 631-651.

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

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

Kirtman, B. P., et al., 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.

Kumar, A, and Z.-Z. Hu, 2014: How variable is the uncertainty in ENSO sea surface temperature prediction? *J. Climate*, **27**, 2779-2788.

Tippett, M. K., and A. G. Barnston, 2008: Skill of multimodel ENSO probability forecasts. *Mon. Wea. Rev*., **136**, 3933-3946.

4. PI Contact Information

Anthony Barnston, tonyb@iri.columbia.edu 845-680-4447, International Research Institute for Climate and Society, Columbia University, Palisades, New York 10964.

Michael Tippett, tippett@iri.columbia.edu 845-680-4420, International Research Institute for Climate and Society, Columbia University, Palisades, New York 10964.

Huug van den Dool, huug.vandendool@noaa.gov Climate Prediction Center, National Oceanic and Atmospheric Administration, College Park, Maryland

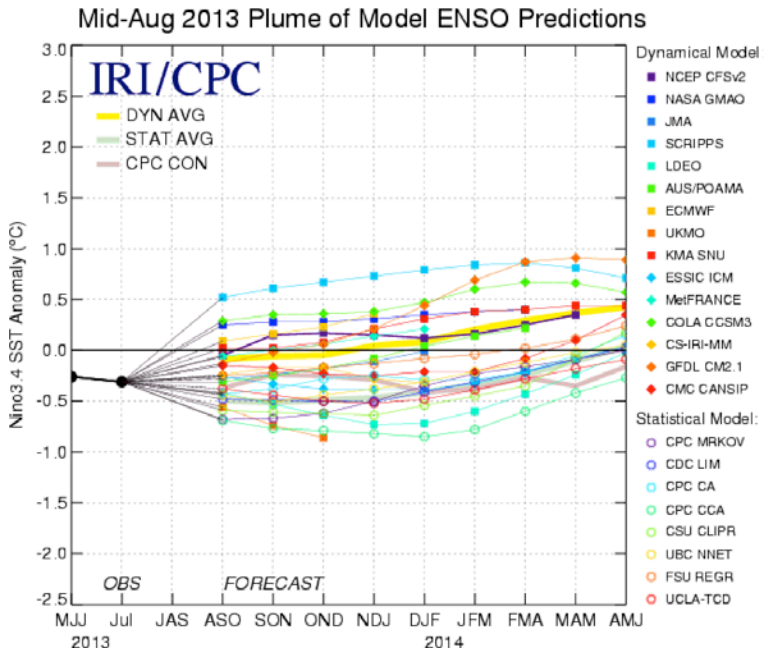

Figure 1. *ENSO prediction plume issued by IRI and NOAA/CPC in July, 2013, for the periods of August-October 2013 through April-June 2014. Recent observed SST anomalies in the Nino3.4 region are shown by the black curves on the left side.*

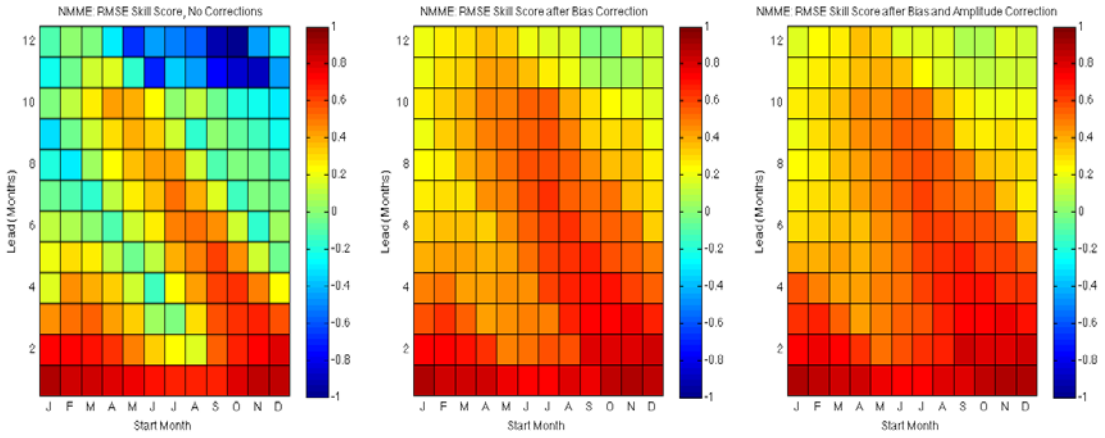

Figure 2. *Root-mean-squared error (RMSE) skill score of the NMME ENSO forecast system as a function of forecast start month (x-axis) and lead time (y-axis) in the cases of no systematic error correction (left), individual model bias correction (middle), and both bias and forecast amplitude correction (right). Positive scores have lower RMSE than climatology forecasts.*

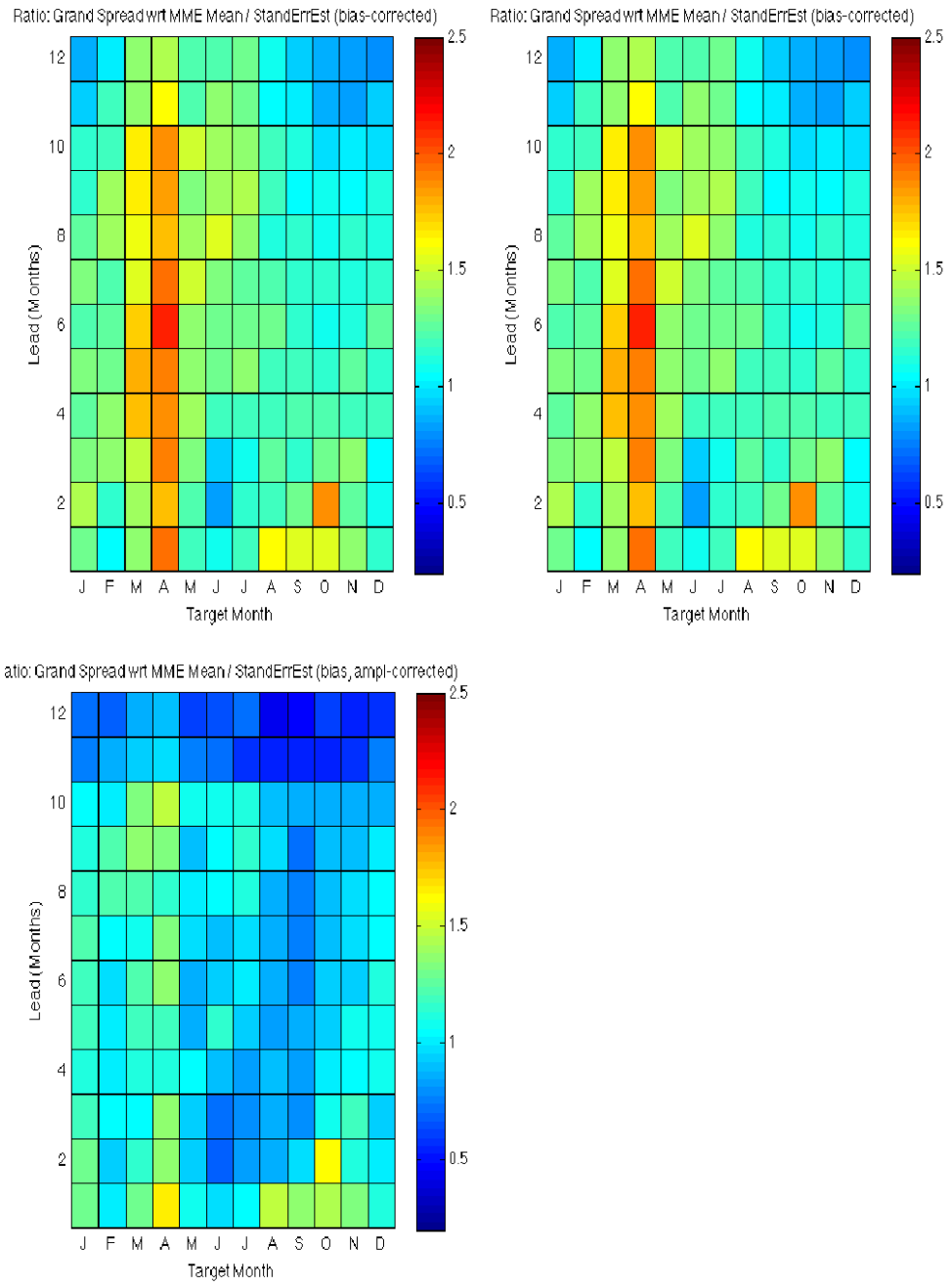

Figure 3. *Ratio of ensemble spread (as standard deviation of all NMME ensemble members about the NMME ensemble mean) to the standard error of estimate (SEE) based on the hindcast correlation skill of the NMME, as a function of forecast target month (x-axis) and lead time (y-axis). Ratios are shown before forecast bias correction (upper left), after bias correction (upper right), and after both bias and amplitude correction (bottom).*

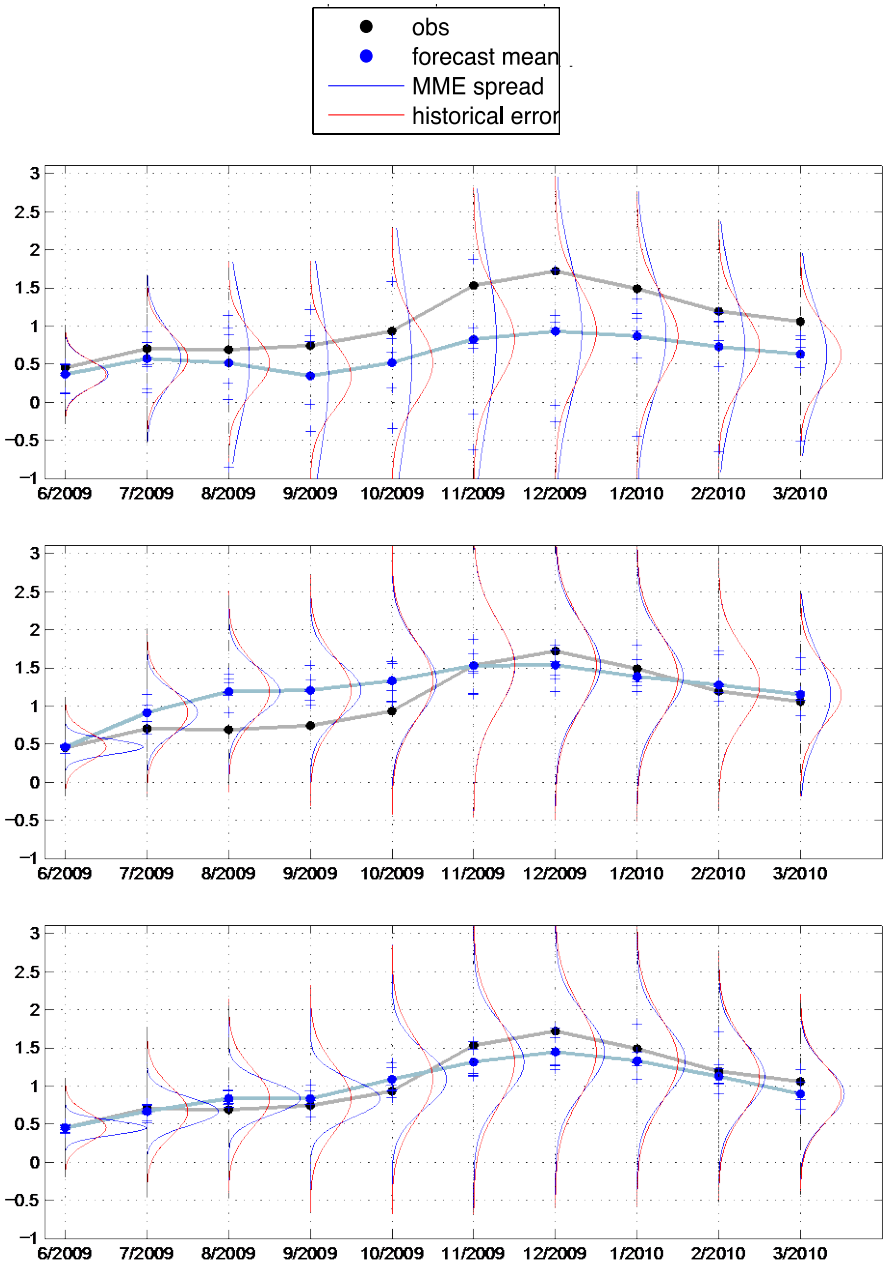

Figure 4. *MME forecasts from June 2009 for the period of the 2009/2010 El Niño event. Top panel shows forecasts without any corrections, middle panel after bias correction, and bottom panel after bias and amplitude correction. The blue line and solid dots show the MME mean forecasts; the black line 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 standard error of estimate (SEE).*

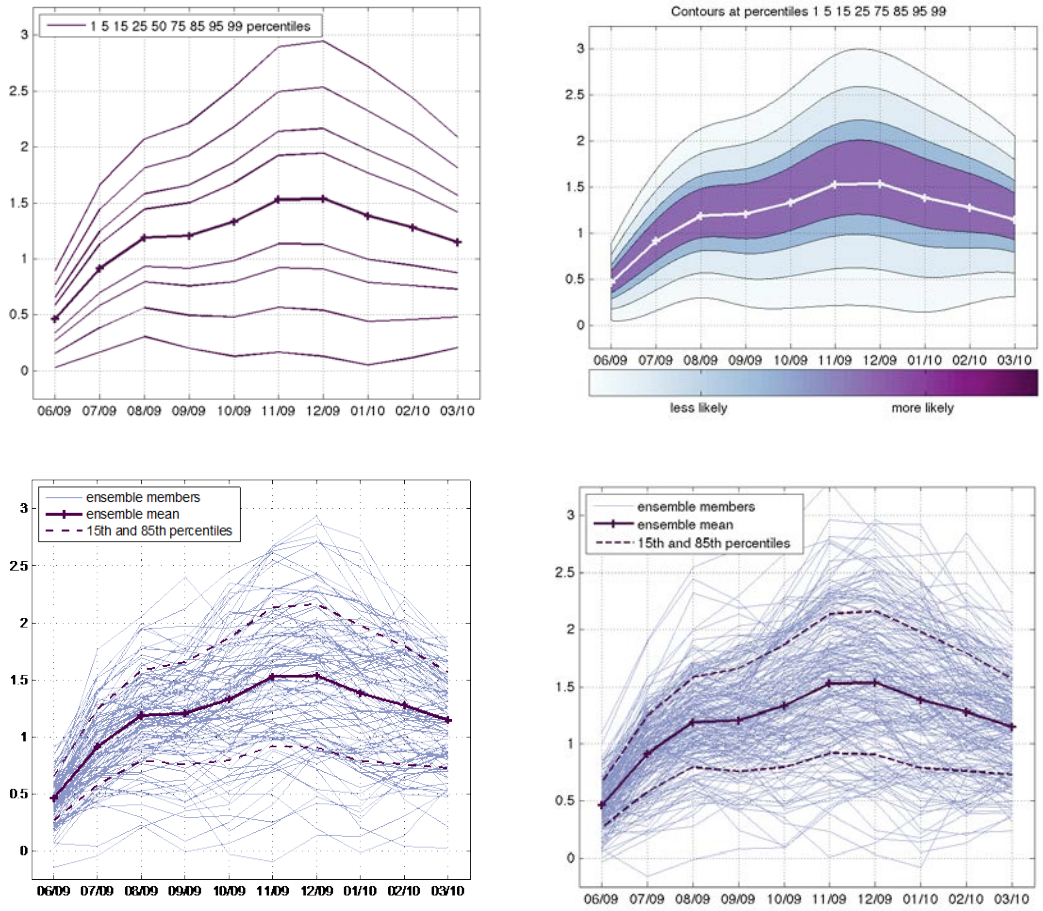

Figure 5. *Four possible formats for an improved ENSO prediction plume. All show the forecast made in June 2009 for the 2009-2010 El Nino episode where the models are corrected for mean bias but not amplitude bias (corresponding to the middle panel of Fig. 4). In the upper right figure, "less likely" and "more likely" refer to lower and higher probability density, respectively. See the text for details.*

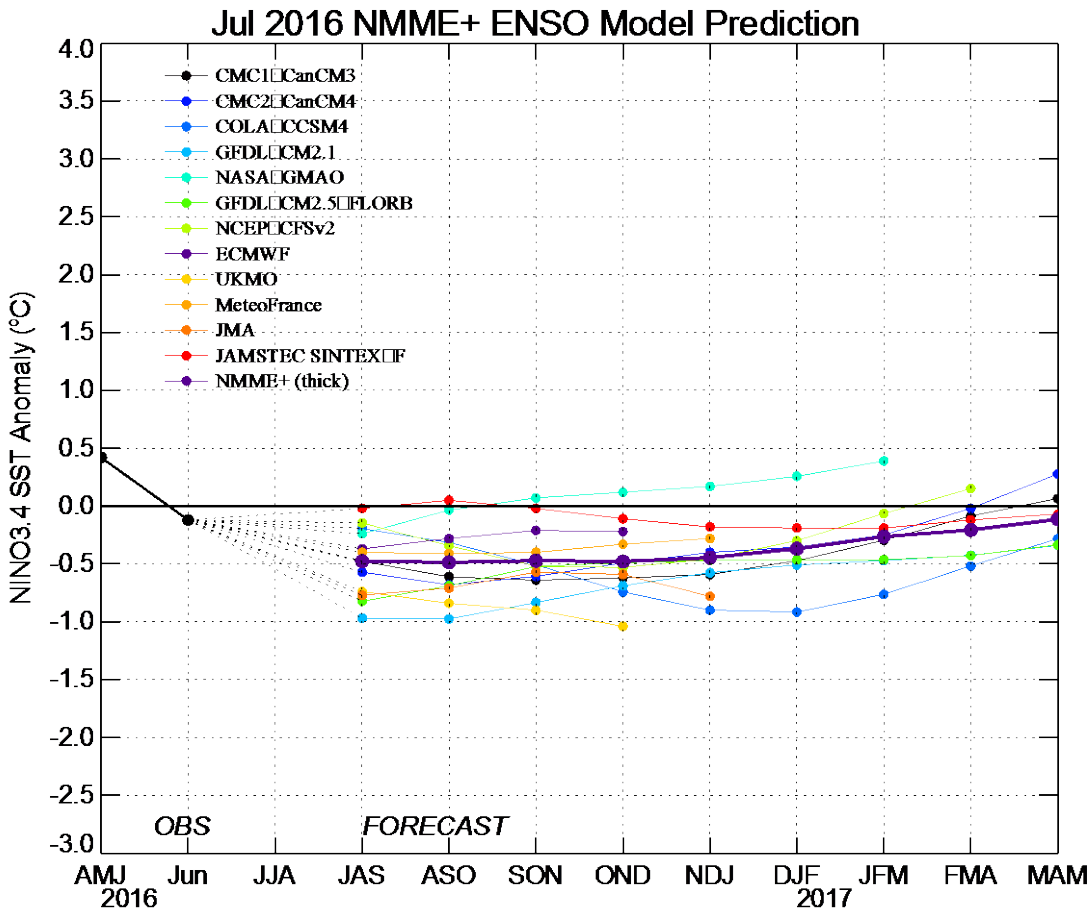

Figure 6*. Ensemble mean forecasts of each of the seven individual models making up the NMME set, and the multi-model ensemble mean forecast (thicker purple line), computed using equal weighting of the means of the individual models. Forecast was made in July 2016 as the borderline La Nina of 2016 was just beginning. This diagram is updated each month on the web page given just below Table 2 above.*

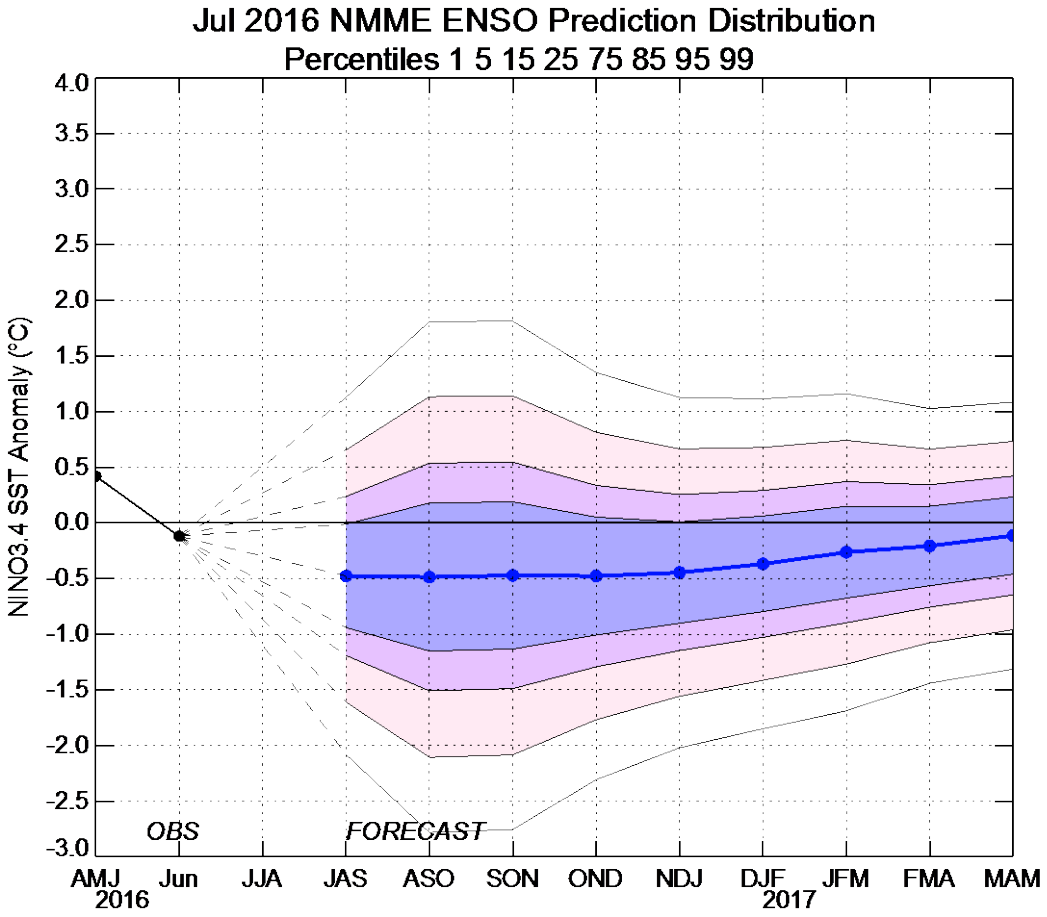

Figure 7. *Summary of the overall uncertainty distribution for the July 2016 forecast, as described by key percentile values. The multi-model ensemble mean is shown, representing the 50 percentile, as well as four percentile values above it and four below it. Colors are darker in regions of greater likelihood of occurrence. This diagram is updated each month on the web page given just below Table 2 above.*

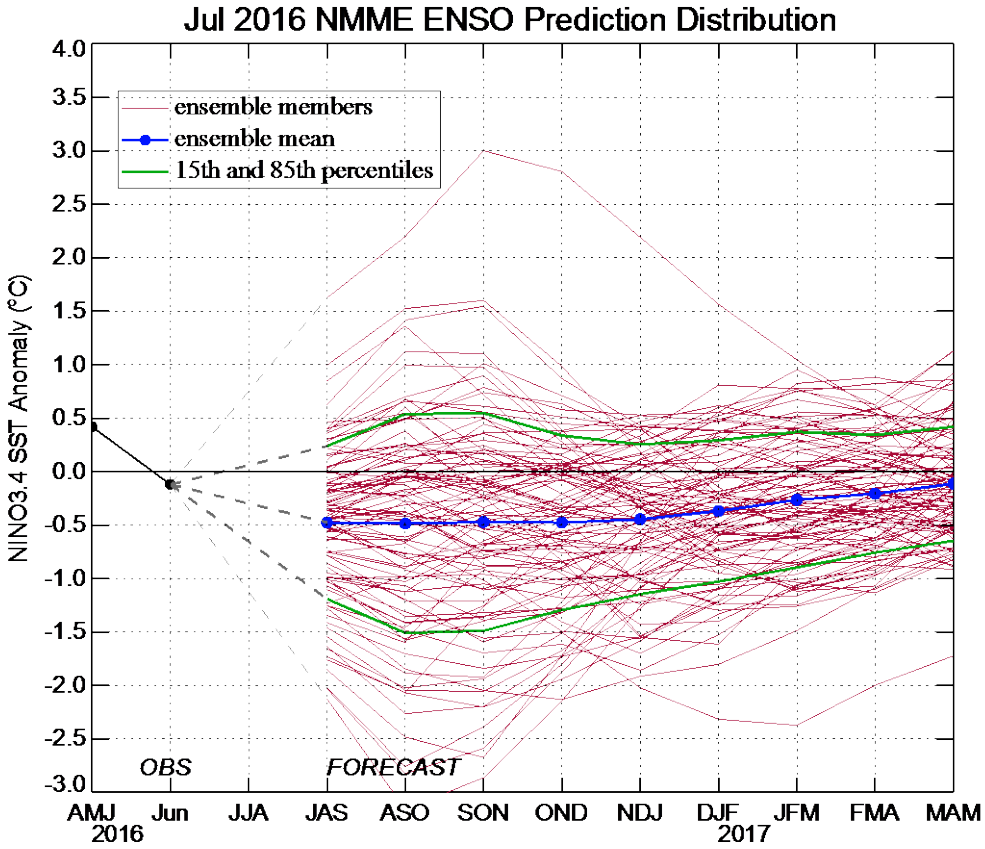

***Figure 8**. "Spaghetti plot" showing the ENSO forecast made in July 2016. The blue line shows the multi-model ensemble mean forecast, the green lines the 15th and 85th percentiles of the forecast uncertainty distribution, and the purple lines 100 equally likely scenarios statistically calibrated using the historical forecast errors and their covariance among lead times. The purple spaghetti lines can be viewed as statistically generated "ensemble members" of the forecast system. This diagram is updated each month on the web page given just below Table 2 above.*