# THE NORTH AMERICAN MULTIMODEL ENSEMBLE

## Phase-1 Seasonal-to-Interannual Prediction; Phase-2 toward Developing Intraseasonal Prediction

BY BEN P. KIRTMAN, DUGHONG MIN, JOHNNA M. INFANTI, JAMES L. KINTER III, DANIEL A. PAOLINO, QIN ZHANG, HUUG VAN DEN DOOL, SURANJANA SAHA, MALAQUIAS PENA MENDEZ, EMILY BECKER, PEITAO PENG, PATRICK TRIPP, JIN HUANG, DAVID G. DEWITT, MICHAEL K. TIPPETT, ANTHONY G. BARNSTON, SHUHUA LI, ANTHONY ROSATI, SIEGFRIED D. SCHUBERT, MICHELE RIENECKER, MAX SUAREZ, ZHAO E. LI, JELENA MARSHAK, YOUNG-KWON LIM, JOSEPH TRIBBIA, KATHLEEN PEGION, WILLIAM J. MERRYFIELD, BERTRAND DENIS, AND ERIC F. WOOD

The North American Multimodel Ensemble prediction experiment is described, and forecast quality and methods for accessing digital and graphical data from the model are discussed.

After more than three decades of research into the origins of seasonal climate predictability and the development of dynamical model-based seasonal prediction systems, the continuing relatively deliberate pace of progress has inspired two notable changes in prediction strategy, largely based on multi-institutional international collaborations. One change in strategy is the inclusion of quantitative information regarding uncertainty (i.e., probabilistic prediction) in forecasts and probabilistic measures of forecast quality in the verifications (e.g., Palmer et al. 2000; Goddard et al. 2001; Kirtman 2003; Palmer et al. 2004; DeWitt 2005; Hagedorn et al. 2005; Doblas-Reyes et al. 2005; Saha et al. 2006; among many others). The other change is the recognition that a multimodel ensemble strategy is a viable approach for adequately resolving forecast uncertainty (Palmer et al. 2004; Hagedorn et al. 2005; Doblas-Reyes et al. 2005; Palmer et al. 2008), although other techniques such as perturbed physics ensembles (currently in use at the Met Office for their operational system) or stochastic physics (e.g., Berner et al. 2008) have been developed and appear to be quite promising. The first change in prediction strategy naturally follows from the fact that climate variability includes a chaotic or irregular component, and, because of this, forecasts must include a quantitative assessment of this uncertainty. More importantly, the climate prediction community now understands that the potential utility of climate forecasts is based on end-user decision support (Palmer et al. 2000; Morse et al. 2005; Challinor et al. 2005), which requires probabilistic forecasts that include quantitative information regarding forecast uncertainty. The second change in prediction strategy follows from the first, because, given our current modeling capabilities, a multimodel strategy is a practical and relatively simple approach for quantifying forecast uncertainty due to uncertainty in model formulation, although it is likely that the uncertainty is not fully resolved.

More recently, there has been a growing interest in forecast information on time scales beyond 10 days but less than a season. For example, the National Centers for Environmental Prediction Climate Prediction Center (NCEP/CPC) in the United

States currently makes outlook-type forecasts for extended weather forecast ranges (i.e., 2 weeks) such as the NCEP/CPC Global Tropical Hazards/Benefits Assessment provides forecasts of anomalous tropical temperature and precipitation. The U.S. Hazards Assessment product, also issued by NCEP/CPC, includes outlooks of potential hazards in the United States up to 16 days. At present, such outlook-style forecast products are based on a subjective combination of various statistical and dynamical methods, although there is momentum to make the process more objective using real-time dynamic model forecasts. These developments demonstrate the demand for such dynamical forecast information.

This week 2–4 time scale is coupled to the seasonal time scale[1] and is often viewed as a source of predictability for seasonal time scales, yet the mechanisms for predictability on this time scale are less well understood (as compared to, say, ENSO). Despite this, there is substantial evidence for dynamic subseasonal predictions that are of sufficient quality to be useful (e.g., Pegion and Sardeshmukh 2011) and evidence that a multimodel approach will enhance forecast quality on this time scale [see the coordinated Intraseasonal Variability Hindcast Experiment (ISVHE); http://iprc.soest.hawaii.edu/users/jylee/clipas/].

Given the pragmatic utility of the multimodel approach, there is multiagency [National Oceanic and Atmospheric Administration (NOAA), National Science Foundation (NSF), National Aeronautics and Space Administration (NASA), and U.S. Department of Energy (DOE)] support for a North American Multimodel Ensemble (NMME) intraseasonal to seasonal to interannual (ISI) prediction experiment. This experiment leverages an NMME team that has already formed and began producing routine real-time multimodel ensemble ISI predictions since August 2011. The forecasts are provided to the NOAA CPC on an experimental basis for evaluation and consolidation as a multimodel ensemble ISI prediction system. The experimental prediction system developed by this NMME team is as an "NMME of opportunity" in that the seasonal-to-interannual prediction systems are readily available and each team member has independently developed the initialization and prediction protocol. We will refer to the NMME of opportunity as phase-1 NMME (or NMME-1). The NMME-1 focuses on seasonal-to-interannual time scales in that the data that are exchanged monthly.

The newly funded multiagency experiment will develop a more "purposeful NMME" in which the requirements for operational ISI prediction will be used to define the parameters of a rigorous reforecast experiment and evaluation regime. This will be phase-2 NMME (or NMME-2). The NMME team will design and test an operational NMME protocol that will guide future research, development, and implementation of the NMME beyond what can be achieved based on the NMME-1 project.

The NMME-2 experiment will do as follows:

i) Build on existing state-of-the-art U.S. climate prediction models and data assimilation systems that are already in use in NMME-1 (as well as upgraded versions of these forecast systems), introduce a new forecast system, and ensure interoperability so as to easily incorporate future model developments.

[1] Any dynamical seasonal prediction system (e.g., coupled atmosphere–ocean model) must pass through the subseasonal time scale.

**AFFILIATIONS:** Kirtman, Min, and Infanti—Rosenstiel School for Marine and Atmospheric Science, University of Miami, Miami, Florida; Kinter and Paolino—Center for Ocean–Land–Atmosphere Studies, Calverton, Maryland; Zhang, van den Dool, Saha, Mendez, Becker, Peng, Tripp, and Huang—NOAA/National Centers for Environmental Prediction, Camp Springs, Maryland; Tippett— International Research Institute for Climate and Society, Palisades, New York, and Center of Excellence for Climate Change Research, Department of Meteorology, King Abdulaziz University, Jeddah, Saudi Arabia; Dewitt,* Barnston, and S. Li—International Research Institute for Climate and Society, Palisades, New York; Rosati—NOAA/Geophysical Fluid Dynamics Laboratory, Princeton, New Jersey; Schubert, Rienecker, Suarez, Z. Li, Marshak, and Lim—NASA Goddard Space Flight Center, Greenbelt, Maryland; Tribbia—National Center for Atmospheric Research, Boulder, Colorado; Pegion—CIRES, University of Colorado Boulder, Boulder, Colorado; Merryfield and Denis—Environment Canada, Fredericton, New Brunswick, Canada; Wood—Princeton University, Princeton, New Jersey
***CURRENT AFFILIATION:** NOAA/National Weather Service, Washington, D.C.
**CORRESPONDING AUTHOR:** Ben Kirtman, Rosenstiel School for Marine and Atmospheric Science, 4600 Rickenbacker Causeway, Miami, FL 33149
E-mail: bkirtman@rsmas.miami.edu

*The abstract for this article can be found in this issue, following the table of contents.*
DOI:10.1175/BAMS-D-12-00050.1

In final form 6 June 2013

ii) Take into account operational forecast requirements (forecast frequency, lead time, duration, number of ensemble members, etc.) and regional/user-specific needs. A focus of this aspect of the experiment will be the hydrology of various regions in the United States and elsewhere in order to address drought and extreme event prediction. An additional focus of NMME-2 will be to develop and evaluate a protocol for intraseasonal or subseasonal multimodel prediction.
iii) Utilize the NMME system experimentally in a near-operational mode to demonstrate the feasibility and advantages of running such a system as part of NOAA's operations.
iv) Enable rapid sharing of quality-controlled reforecast data among the NMME team members and develop procedures for timely and open access to the data, including documentation of models and forecast procedures, by the broader climate research and applications community.

This paper describes the ongoing NMME-1 project, including a preliminary multimodel forecast quality assessment and our strategy for evaluating how the multimodel approach contributes to the forecast quality. We also describe how NMME-2 will evolve from NMME-1 and the coordinated research activities and data dissemination strategy envisaged.

**THE PHASE-1 NMME.** Based on two Climate Test bed (CTB) NMME workshops (18 February and 8 April 2011), a collaborative and coordinated implementation strategy for a NMME prediction system (NMME-1) was developed. The strategy included calendar year 2011 (CY2011) experimental real-time ISI forecasting (summarized below) that leveraged existing CTB partner activities.

*Hindcast and real-time experimental prediction protocol.* The CY2011 NMME experimental predictions have been made in real time since August 2011. As part of the development of the real-time capability, the NMME partners agreed on a hindcast and real-time prediction protocol. Some of the key elements of this protocol include the following:

- Real-time ISI prediction system must be identical to the system used to produce hindcasts. This necessarily includes the procedure for initializing the prediction system. The number of ensemble members per forecast, however, can be larger for the real-time system.
- Hindcast start times must include all 12 calendar months, but the specific day of the month or the ensemble generation strategy is left open to the forecast provider.
- Lead times up to 9 months are required, but longer leads are encouraged.
- The target hindcast period is 30 years (typically 1981–2010).
- The ensemble size is left open to the forecast provider, but larger ensembles are considered better.
- Data distributed must include each ensemble member (not the ensemble mean). Total fields are required [i.e., systematic error corrections to be coordinated by multimodel ensemble (MME) combination lead; NOAA/CPC]. Forecast providers are welcome to also provide bias-corrected forecasts and to develop their own MME combinations.
- Model configurations—resolution, version, physical parameterizations, initialization strategies, and ensemble generation strategies—are left open to forecast providers.
- Required output is monthly means of global grids of SST, 2-m temperature (T2m), and precipitation rate. More fields will be added based on experience and demand. It is also recognized that higher-frequency data are desirable and this will be implemented as feasible.
- Routine real-time forecast data must be available by the eighth of each month.

The NMME-1 activity began in February 2011 and became an experimental real-time system in August 2011. Specifically, on 8 August, NCEP [CPC and the Environmental Modeling Center (EMC)] collected from the respective FTP sites of the NMME partners the real-time seasonal predictions. In the months before August 2011, the hindcast data were collected and climatologies and skill assessments for each model to be applied to subsequent real-time predictions were calculated. Graphical forecast guidance based on the NMME was prepared and given to NOAA operational forecasters in time for the CPC seasonal prediction cycle. The graphical forecast guidance includes North American and global domains and T2m ($T$), precipitation ($P$), and SST fields, and the plots are for monthly and seasonal means with and without a skill mask applied. All NMME forecasts are bias corrected (making use of the hindcasts) using cross validation [see Kirtman and Min (2009) for details of how to make the bias correction].

The effort is significant because, although experimental, the NMME protocol adheres to CPC's operational schedule, so the forecasters can use the information for operational guidance. The scripts for

the data ingest and graphical outputs are intended to be robust (i.e., any number of models) with any number of ensemble members can be used. A major element of the NMME experiment is to continue this effort for the benefit of operations. Meanwhile, we have built up a live hindcast dataset of about 30 years that is open to anybody and can be used for research. Quite probably, this NMME dataset is now the most extensive multimodel seasonal prediction archive currently available that includes models that are continuing to make real-time predictions. Table 1 summarizes the NMME-1 hindcast datasets and identifies the point of contact for each prediction system.

In addition, NOAA/CPC has agreed to evaluate the hindcasts, combine the forecasts, perform verification, provide an NMME website (www.cpc.ncep.noaa.gov/products/NMME), and make the real-time NMME forecast delivery to NOAA forecasters. CPC is also maintaining an NMME newsletter. The hindcast data and real-time forecast data are also available for download or analysis at the International Research Institute for Climate and Society (IRI) (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/). The CPC site primarily serves the real-time needs of the project, and the IRI site, along with the analysis tools that are being developed at the IRI (http://iridl.ldeo.columbia.edu/home/.tippett/.NMME/.Verification/), primarily serves research needs in terms of assessing the prediction skill and predictability limits associated with NMME-1 in terms of designing the NMME-2

**Table 1. NMME partner models and forecasts.**

| Model | Hindcast period | Ensemble size | Lead times (months) | Arrangement of ensemble members | Contact and reference |
|---|---|---|---|---|---|
| CFSv1 | 1981–2009 | 15 | 0.5–8.5 | First 0000 UTC ±2 days, 21st 0000 UTC ±2 days, and 11th 0000 UTC ±2 days | Saha (Saha et al. 2006) |
| CFSv2 | 1982–2010 | 24(28) | 0.5–9.5 | Four members (0000, 0600, 1200, and 1800 UTC) every fifth day | Saha (Saha et al. 2014) |
| GFDL Climate Model, version 2.2 (GFDL CM2.2) | 1982–2010 | 10 | 0.5–11.5 | All first of the month 0000 UTC | Rosati (Zhang et al. 2007) |
| IRI-ECHAM4f* | 1982–2010 | 12 | 0.5–7.5 | All first of the month 0000 UTC | DeWitt (DeWitt 2005) |
| IRI-ECHAM4a* | 1982–2010 | 12 | 0.5–7.5 | All first of the month 0000 UTC | DeWitt (DeWitt 2005) |
| CCSM3 | 1982–2010 | 6 | 0.5–11.5 | All first of the month 0000 UTC | Kirtman (Kirtman and Min 2009) |
| Goddard Earth Observing System, version 5 (GEOS5) | 1981–2010 | 11** | 0.5–9.5 | One member every fifth day | Schubert (G. Vernieres et al. 2011, unpublished manuscript) |
| Third Generation Canadian Coupled Global Climate Model (CMC1-CanCM3) | 1981–2010 | 10 | 0.5–11.5 | All first of the month 0000 UTC | Merryfield (Merryfield et al. 2013) |
| Fourth Generation Canadian Coupled Global Climate Model (CMC2-CanCM4) | 1981–2010 | 10 | 0.5–11.5 | All first of the month 0000 UTC | Merryfield (Merryfield et al. 2013) |

* Real-time forecasts terminated in Jul 2012.

** The number of forecast and hindcast ensemble members is not constant during the period. It has grown from 6 for the initial Aug 2011 forecasts (and associated hindcasts) to 11 starting with our Jun 2012 forecasts. The additional (beyond 6 initialized every fifth day) ensemble members are based on breeding and other perturbations applied on the day closest to the beginning of the month.

experimental protocol. While the NMME-1 data are limited to monthly-mean data, it is a research tool (or testbed) that is proving extremely useful in supporting the basic prediction and predictability research needs of the project participants. This database also serves as "quick look" easy access data that are the external face of the NMME experiment to the research community.

**RESULTS: NMME-1.** Here, we show some results from the 28 years of hindcasts that cover a common period (i.e., 1982–2009) for all the models and the real-time experimental forecast from the NMME of opportunity (i.e., NMME-1). The results help provide evidence of the benefit of a multimodel ensemble of predictions, as compared with the ensemble predictions of just one high-performing model. Figure 1 shows the range spanned by the individual ensemble members from each forecast system in NMME-1, for 0.5-month-lead[2] hindcasts for the Niño-3.4 SST index. This presentation of the range assumes that each ensemble member of each model is equally likely to occur. To calculate anomalies, the forecast bias or systematic error has been removed and is calculated separately for each model using all ensemble members for that particular model. See Saha et al. (2006) or Kirtman and Min (2009) for a discussion of how the systematic error is removed. At this short lead time, the hindcasts tend to agree with one another and with the observations, to a great extent, although there is also some disagreement, particularly at certain times (e.g., near the end of 1988 and in the middle of 1998). However, it is worth noting that nowhere do the observations lie noticeably outside the envelope of the predictions.

Figure 2 shows the same results except for 5.5-month-lead predictions, with appropriately greater uncertainties shown by the larger range—often in excess of 2°C. We will show that it is just such dispersion in the individual predictions that best reflects forecast uncertainty, as well as the "best guess" multimodel-mean prediction.

Figure 3 shows the spatial distribution of the anomaly correlation between the 5.5-month lead of the grand ensemble monthly-mean hindcast and observed SST over 1982–2009. Here, the grand ensemble mean is defined as the average of all the hindcasts, assuming that each ensemble member of each model is equally probable. This is distinct from assuming that each model should be weighted equally. High skill is evident in the central and eastern tropical Pacific Ocean, as well as portions of the tropical Atlantic and Indian Oceans and some isolated regions in the extratropics.

One of the important motivating factors for both phases on the NMME project is to understand the complementary sources of skill among the models. Essentially, we seek to understand the "where and why" in how the multimodel approach improves

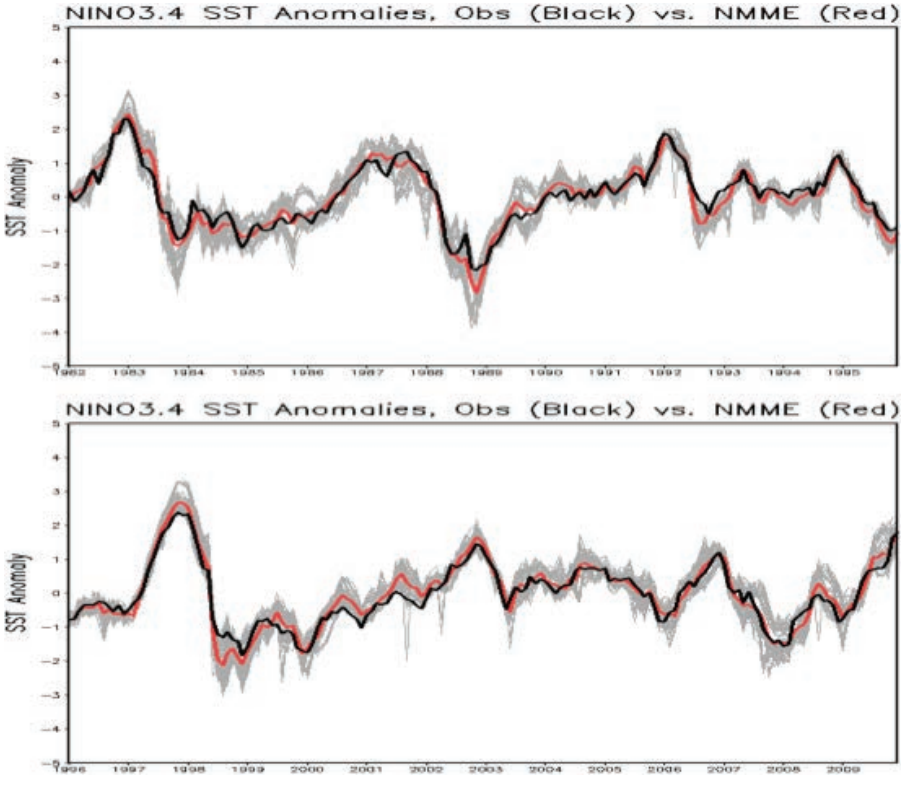


**Fig. 1. Niño-3.4 (area-averaged SSTA 5°S–5°N, 170°–120°W) plumes for 0.5-month lead: (top) 1982–95 and (bottom) 1996–2010.**

[2] The real-time forecasts are issued on the 15th of the month, so that, for example, a January 2013 monthly-mean forecast issued on 15 January 2013 is the 0.5-month lead, and the February 2013 monthly-mean forecast issued on 15 January 2013 is the 1.5-month lead and so on. The retrospective forecasts also follow this convention.

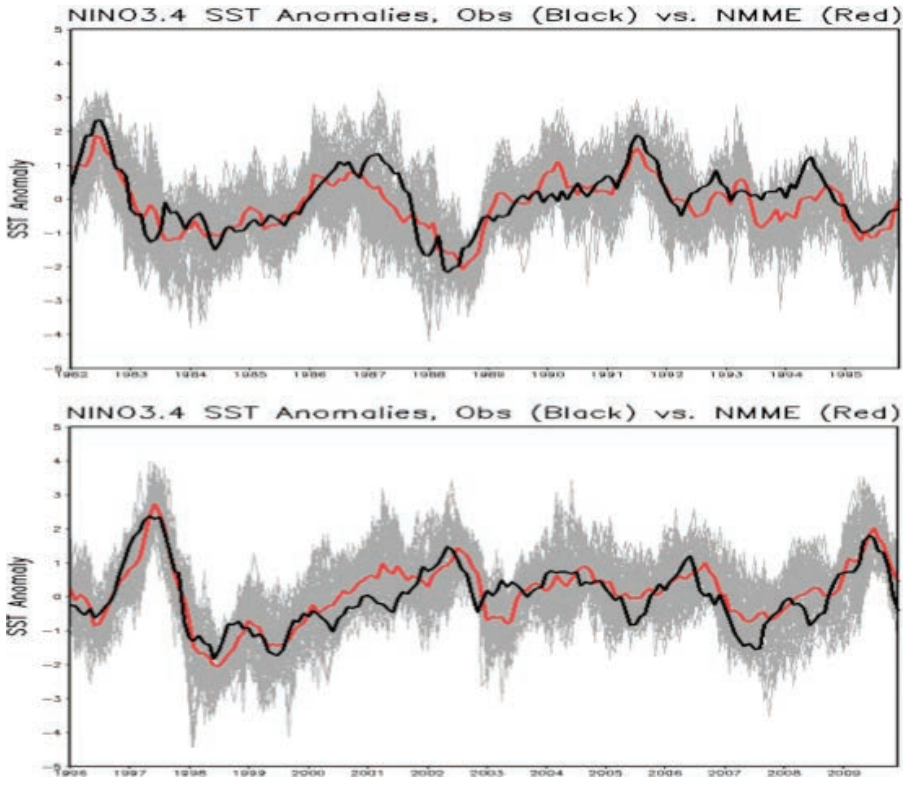


**Fig. 2. As in Fig. 1, but for 6.5-month lead.**

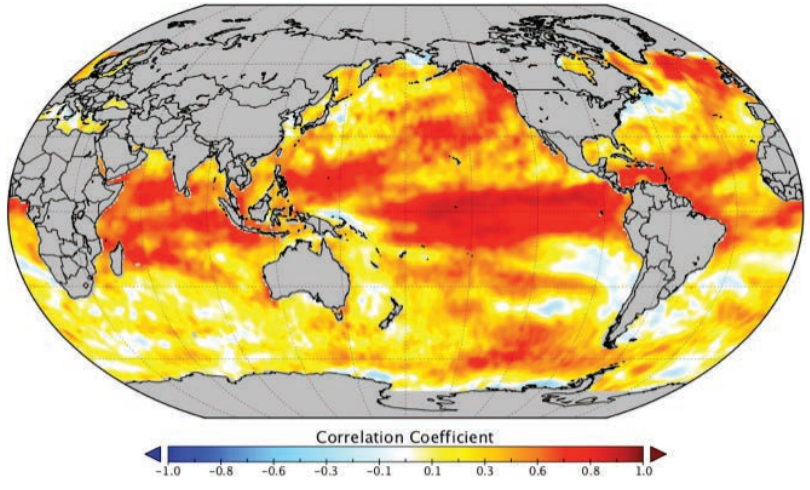


**Fig. 3. SSTA correlation coefficient with each ensemble member weighted equally. Retrospective forecasts are initialized in Aug 1982–2009 and verified in the following Feb (i.e., 5.5-month lead).**

the SST anomaly (SSTA) for individual models' 0.5- to 5.5-month-lead ensemble-mean hindcasts versus the corresponding multimodel ensemble-mean hindcasts for tropical SST for September starts. The percentage noted in each panel corresponds to the number of points where the individual model beat the multimodel. For every single individual model, most of the points are above the diagonal (i.e., the percentage of points below the diagonal is less than 50%), indicating that the multimodel tends to have smaller errors than the individual models. Generally, the models cluster around 26%–48%. The Community Climate System Model, version 3 (CCSM3), is an outlier and is being replaced with the Community Climate System Model, version 4 (CCSM4) in NMME-2.

Preliminary examination (not shown) has suggested that in general the individual model having the highest anomaly correlation skill is Climate Forecast System, version 2 (CFSv2). However, this identification of the generally best model does not suggest that the other models, when allowed to contribute to the multimodel-mean forecast, do not further enhance the performance. To demonstrate the benefit reaped by using the multimodel ensemble over the single best-performing model, the ranked probability skill score (RPSS)[3] of the multimodel ensemble hindcasts and the CFSv2 hindcasts of SST for December–

forecast quality. Here, we show the first step in this process—simply documenting how the multimodel compares to any single model. For example, Fig. 4 shows scatterplots of the root-mean-square error of

[3] RPSS is a probabilistic forecast skill metric [see Weigel et al. (2007) for details]. The RPSS evaluates the hindcasts probabilistically (using tercile-based categories and the equal-odds climatology forecasts as the reference forecast). A good rule of thumb is that an RPSS of 0.08 corresponds to a deterministic correlation of 0.4.

February (DJF) for forecasts initialized in early July are shown in Fig. 5, while those for June–August (JJA) initialized in early January are shown in Fig. 6. In the case of both seasons, the multimodel ensemble produces higher mean skill. There are isolated areas where CFSv2 outperforms the multimodel ensemble, such as in the DJF forecasts (Fig. 5) just south of the equator near 85°, south of Sri Lanka. However, the multimodel ensemble has higher, and more reliably positive, skill over most of the globe than that of any of the individual model forecasts—even the best of them.

The comparatively better RPSS results of the multimodel ensemble hindcasts than those of the CFSv2 forecasts are not limited to SST hindcasts but generalize to predictions for land surface temperature and precipitation as well. Figure 7, for example, shows the spatial distribution of RPSS for land surface temperature for JJA initialized in early January for the multimodel ensemble (top) and CFSv2 (bottom). Again, the multimodel mean has considerably less area with negative skill while maintaining the skill levels at many of the areas where CFSv2 has the highest skill. Multimodel skill at the locations of the most extreme peaks of CFSv2 skill tends to be slightly attenuated (e.g., northeastern Brazil and parts of the Middle East), but mean skill is clearly enhanced.

Figure 8 shows the spatial distribution of RPSS for hindcasts of precipitation for DJF (initialized in July) over North America using the multimodel ensemble (left) and CFSv2 alone (right). Figure 9 is the same as Fig. 8, but for the JJA season (initialized in January). The comparative superiority of the multimodel forecast over CFSv2 alone is noted for both seasons. This is most obvious in the relative lack of negative skill in the multimodel hindcasts but also in the maintenance or even enhancement of areas of peak skill. Additional results for NMME are shown in Yuan and Wood (2012).

It is worth noting that in the case of probabilistic verification, a larger ensemble size has a stronger positive influence on skill than it does for deterministic verification (e.g., using anomaly correlation). This ensemble size effect is described in detail in Richardson (2001), and this greater sensitivity in probability forecasts is due to the larger role of sampling variability in defining tercile probabilities (particularly when done by counting the fraction of ensemble members falling into each category) than in forming an ensemble mean. Indeed, Richardson (2001) shows that a Brier skill score (BSS) of, say, 0.2 for a 100-member ensemble of a single model would be about 0.1 for a 10-member ensemble and 0.17 for a 25-member ensemble. Hence, in addition to the balancing or cancellation of individual model biases,

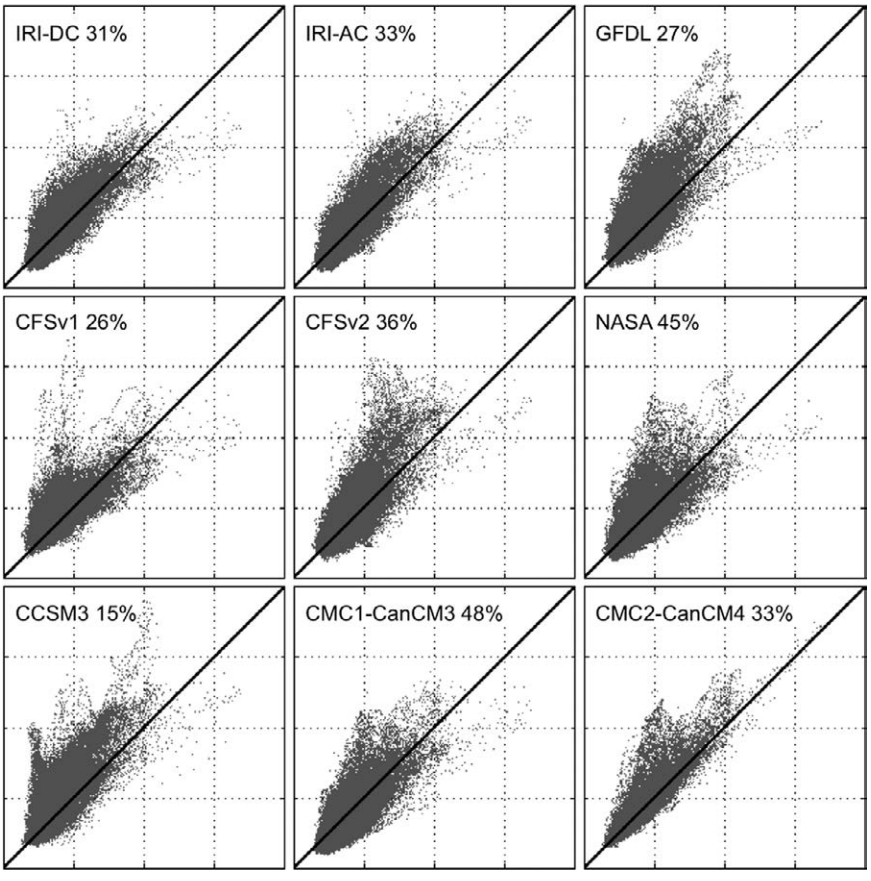


**Fig. 4. SSTA RMSE 20°S–20°N for each individual model compared to the multimodel mean; Sep starts 1982–2009, leads 0.5–5.5 months. The *x* axis ranges from 0° to 2°C and corresponds to the NMME RSME, and the *y* axis ranges from 0° to 2°C and corresponds to the individual model RMSE. Dots above the diagonal imply NMME has smaller RMSE. The percentage of points below the diagonal is noted in each panel. IRI-AC corresponds to IRI-ECHAM4a and IRI-DC corresponds to IRI-ECHAM4f in Table 1.**

a secondary reason for the relatively better performance of the multimodel hindcasts than CFSv2 is the much larger ensemble size of all the models together than of any single model.

A tool used to diagnose a set of probabilistic forecasts is reliability analysis, which measures the correspondence between the forecast probabilities and their subsequent observed relative frequencies, spanning the full range of issued forecast probabilities for each of the three climatologically equiprobable categories (below, near, or above normal). If one collected all instances of forecasts of 45% probability for "above normal," for example, and that category were actually later observed in 45% of the cases, the forecasts for that particular probability bin would be shown to have perfect reliability. Results of reliability analysis for forecasts initialized in October and verified in the following January–March (JFM) for 2-m temperature anomalies over the globe are shown in Fig. 10 for the multimodel ensemble hindcasts over the 28-yr period for the below-normal and above-normal categories. The light dotted line denotes perfect reliability.

Two aspects of common interest in reliability diagnosis are 1) the overall position of the lines relative to the ideal 45° line and 2) the slope of the lines relative to unity. The general positions of the lines in Fig. 10 are near that of the ideal line, but the line representing above-normal (below normal) forecasts is just slightly higher (lower) than ideal. This indicates a slight tendency to underforecast above-normal and to overforecast below-normal temperature. The observed mean relative frequency of occurrence of the categories, shown as colored dots on the *y* axis, indicates that above normal occurred in about 39% of cases, while below normal (and near normal) occurred in about 30% of cases. However, this weak shift toward above-normal temperature in the mean

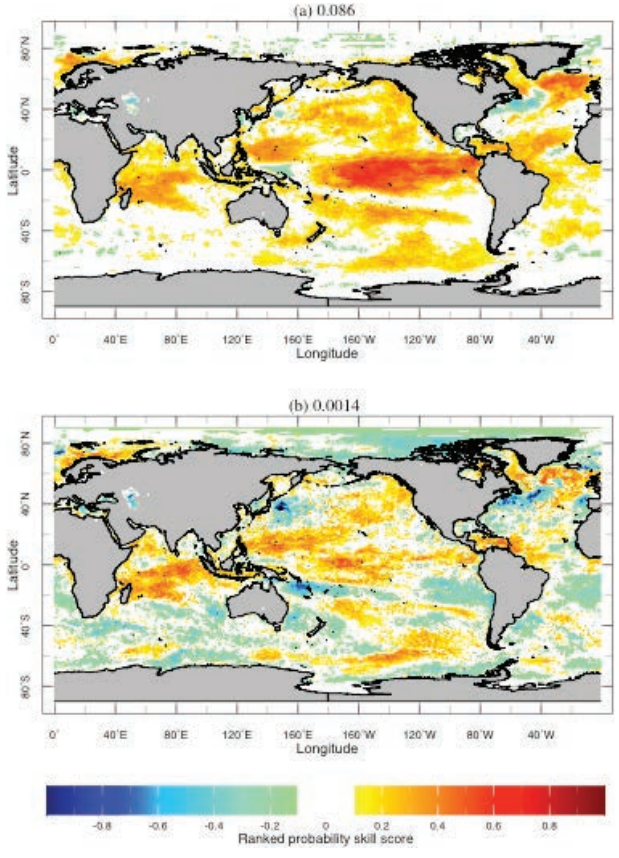


**Fig. 5. SSTA RPSS for the (a) grand NMME multimodel ensemble and for (b) CFSv2. The skill is based on hindcasts initialized in Jul 1982–2009 and verified in the following DJF seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better than climatology, and negative values indicate probabilistic skill that is worse than a climatological forecast. Global-averaged RPSS is noted in the figure.**

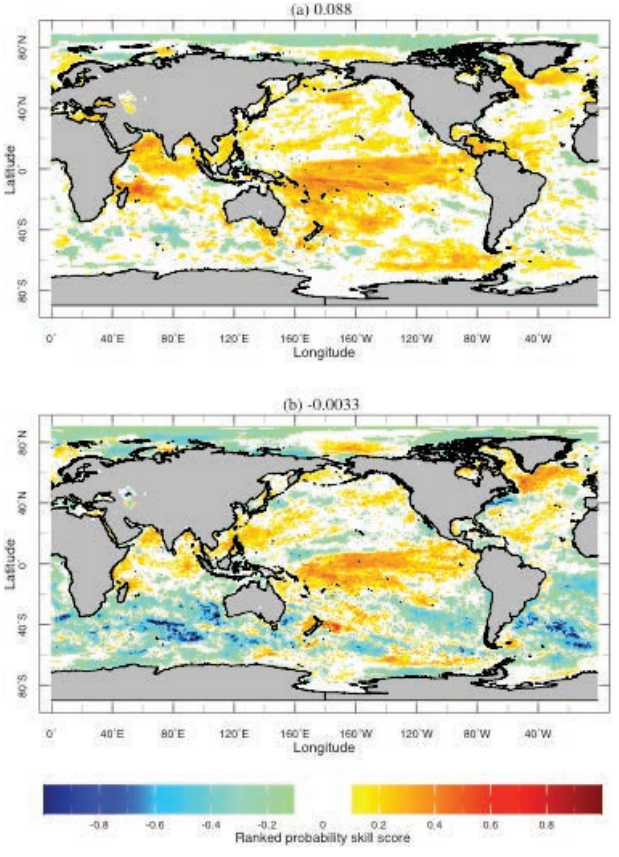


**Fig. 6. SSTA RPSS for the (a) grand NMME multimodel ensemble and for (b) CFSv2. The skill is based on hindcasts initialized in Jan 1982–2009 and verified in the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better than climatology, and negative values indicate probabilistic skill that is worse than a climatological forecast. Global-averaged RPSS is noted in the figure.**

climate over the 28-yr period was induced by a slight offset in the base period of the observations and the model hindcasts: for the observations, the period is 1981–2010, while for the model forecasts it is 1982–2009. Thus, the overall position of the reliability curves, while usually indicative of the model bias, is influenced here by the slight model versus observational base period offset.

The slope of the lines is related to the confidence level of the probability forecasts. Lines with slopes of less than 1 indicate forecast overconfidence, with greater relative differences in forecast probability than the corresponding differences in observed frequencies. A bias toward overconfidence has been noted in many individual dynamical models. Figure 10 indicates that this problem, while present, is very mild in the multimodel ensemble hindcasts compared to the individual models shown in Fig. 11. The amelioration of the overconfidence problem is undoubtedly a consequence of partial cancellation of somewhat conflicting signals that are overconfident in many of the individual models, resulting in an appropriately more probabilistically conservative forecast when the models are combined.

The offsetting of potentially overconfident forecasts of individual models when combined into a multimodel ensemble is illustrated by an example of a recent real-time prediction of the Niño-3.4 SST index (Fig. 12). The predictions of the individual ensemble members express the uncertainty distribution within each model, while the overall plume of forecasts express the uncertainty of the full multimodel ensemble. It is noted that the uncertainty distributions of the individual models is smaller than that of the collection of members of all models. The multimodel ensemble is probabilistically less overconfident than the ensembles of most of the individual models, because each individual model is imperfect, but has a higher than realistic confidence level in its "model world." Combining many models serves to offset differing biases, resulting in a more balanced and probabilistically reliable prediction.

One measure of the success of the NMME project is whether it will advance hydrologic applications,

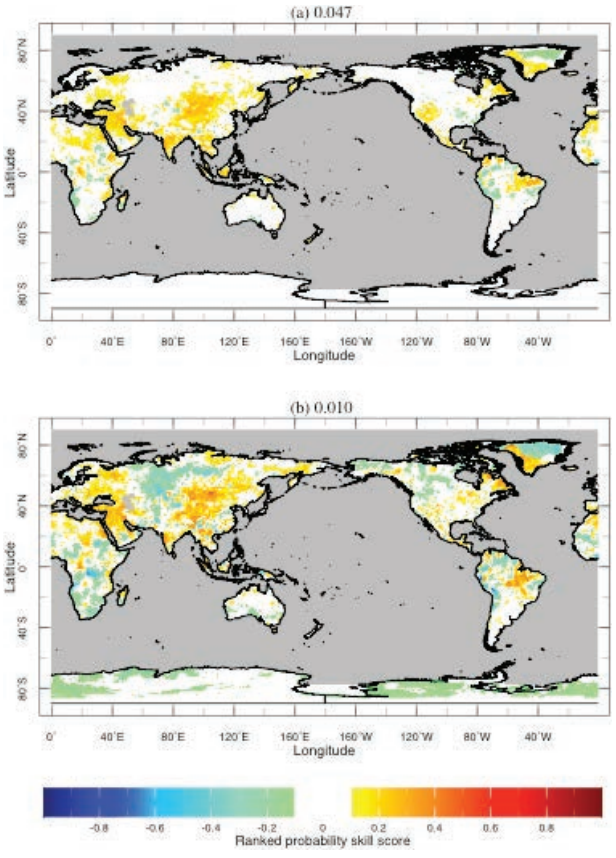


**Fig. 7. Surface atmospheric temperature (2 m) RPSS for the (a) grand NMME multimodel ensemble and for (b) CFSv2. The skill is based on hindcasts initialized in Jan 1982–2009 and verified in the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better than climatology, and negative values indicate probabilistic skill that is worse than a climatological forecast. Global-averaged RPSS is noted in the figure.**

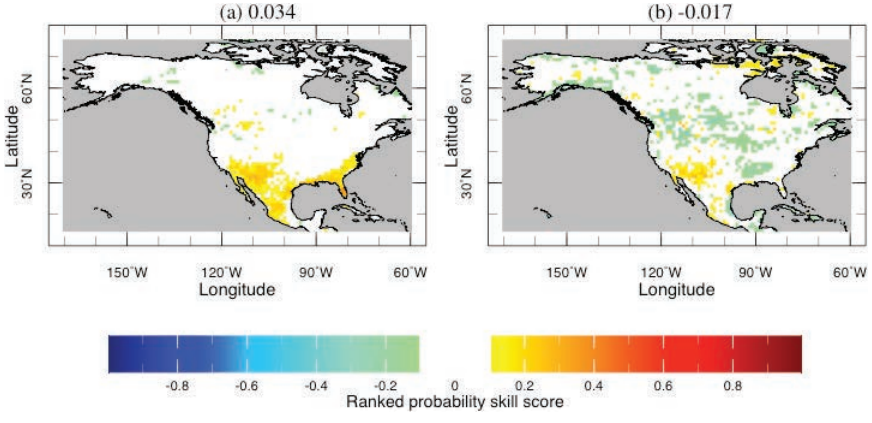


**Fig. 8. Precipitation forecast RPSS for the (a) grand NMME multimodel ensemble and for (b) CFSv2. The skill is based on hindcasts initialized in Jul 1982–2009 and verified in the following DJF seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better than climatology, and negative values indicate probabilistic skill that is worse than a climatological forecast. Global-averaged RPSS is noted in the figure.**

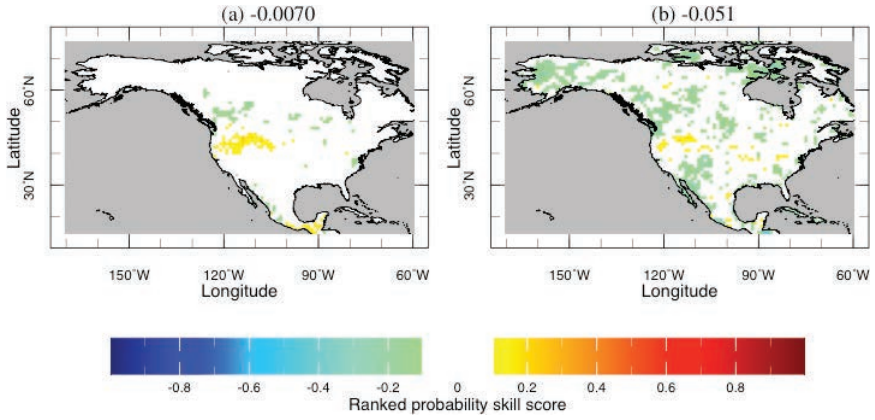


**Fig. 9. Precipitation forecast RPSS for the (a) grand NMME multimodel ensemble and for (b) CFSv2. The skill is based on hindcasts initialized in Jan 1982–2009 and verified in the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better than climatology, and negative values indicate probabilistic skill that is worse than a climatological forecast. Global-averaged RPSS is noted in the figure.**

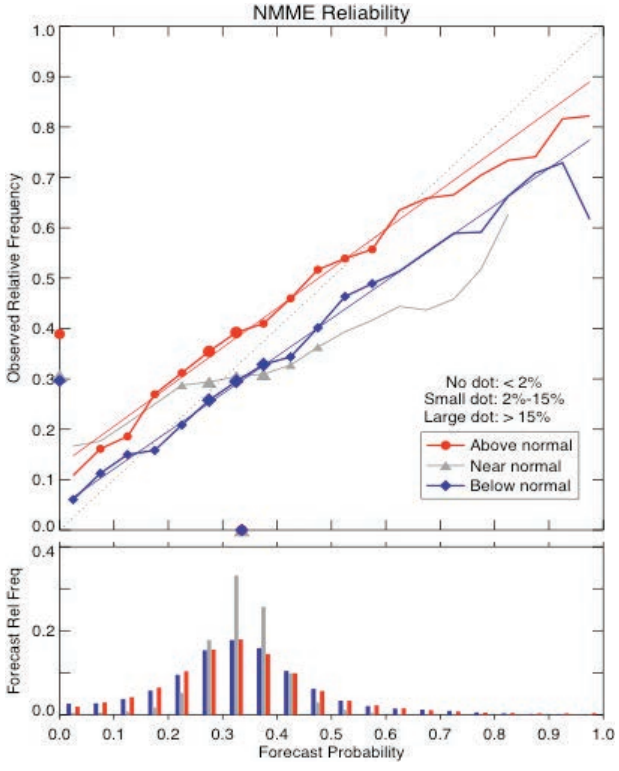


**Fig. 10. NMME reliability diagram for T2m anomalies throughout the globe. The reliability corresponds to forecasts initialized in Oct 1982–2009 and verified in following JFM season.**

which include streamflow and drought forecasting. Drought forecasting includes not only meteorological drought but also agricultural and hydrological drought. Meteorological drought is assessed through precipitation deficits with indices like the standardized precipitation index (SPI) determined over a window centered on the initial forecast date. Agricultural drought focuses on soil moisture deficits or indices such as their percentiles (Sheffield et al. 2004) and hydrological drought on streamflow. Collectively under the NMME project, seasonal hydrologic forecasting will include drought forecasting as well as related hydrological seasonal forecasting such as persistent wet conditions. Since hydrological applications usually require information at smaller spatial scales than that provided by the seasonal forecast models, the climate forecasts from the multimodel ensemble will be downscaled and bias corrected, using the approach of Luo et al. (2007), and used to drive a calibrated land surface model. The output of the land surface model is then used to for hydrologic forecasts, including drought. This approach has been well developed (Luo and Wood 2007, 2008; Yuan et al. 2013). Figure 13 shows the results for streamflow forecast skill from NMME relative to the skill from the often-used ensemble streamflow prediction (ESP) approach where hydrological model forcings come from historical resampling. The results are presented over the National Integrated Drought Information System (NIDIS) Colorado and southeastern U.S. testbeds. For the Colorado domain, NMME is more skillful than ESP, particularly in the summer with the skill coming primarily from increased precipitation skill. Not shown is the comparison between CFSv2 alone and NMME in which CFSv2 has slightly lower precipitation skill. For the southeast NIDIS domain, ESP is more skillful for 1-month leads due to low NMME precipitation skill, but the situation changes for longer leads when the full resolution is downscaled; bias-corrected forecasts are used in the hydrological model. For both ESP and NMME hydrological forecasts, observed hydrologic initial states are used at the initial forecast time. These can be provided from the North American Land Data Assimilation System (NLDAS) (Mitchell et al. 2004).

For meteorological drought assessed at continental-to-global scales, the 1° NMME model precipitation forecasts can be used. Figure 14 shows the NMME 6-month SPI (SPI6) forecast initiated

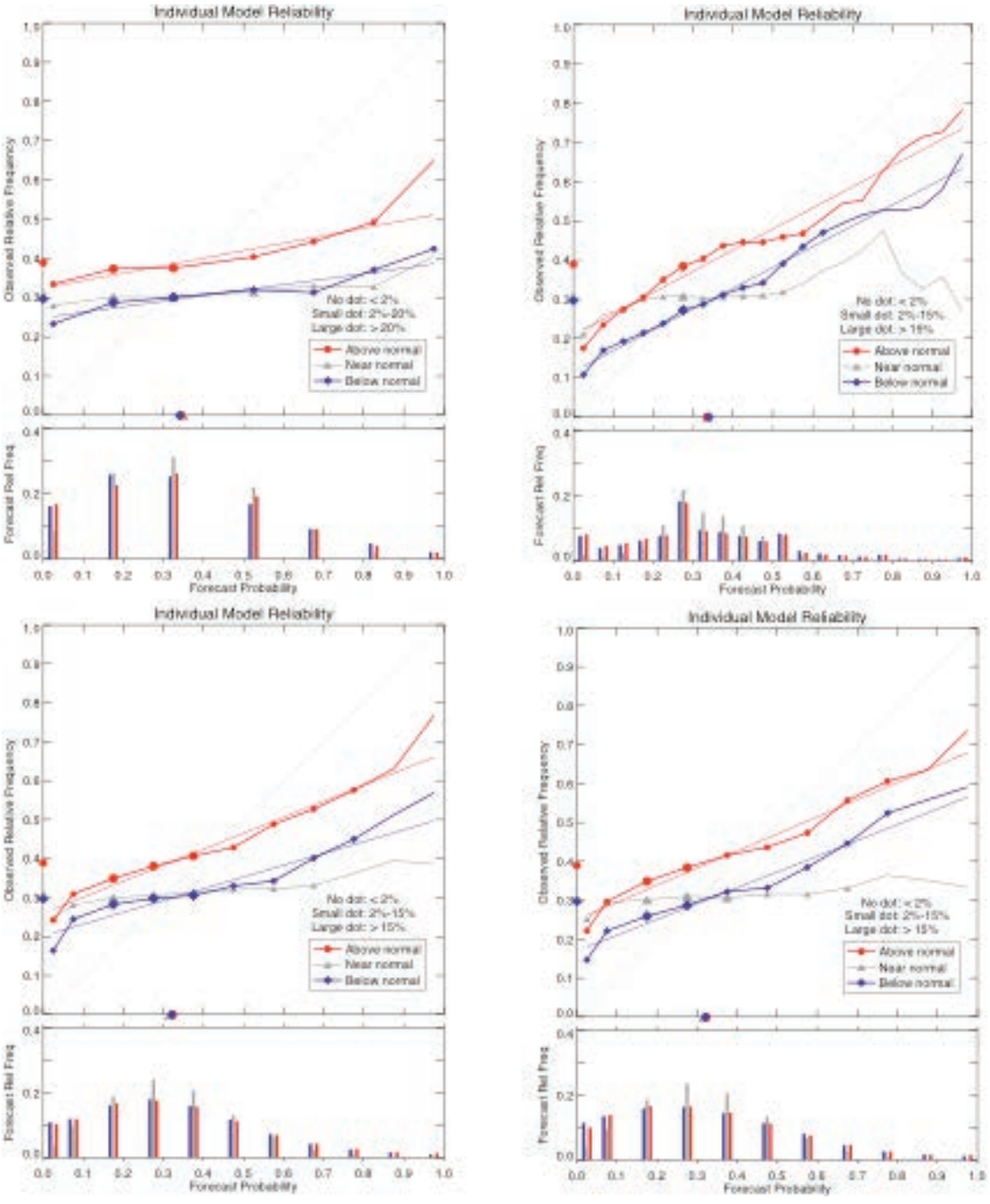


**Fig. 11. Reliability diagram for T2m anomalies throughout the globe from a sample of individual models. The reliability corresponds to forecasts initialized in Oct 1982–2009 and verified in following JFM season.**

on 1 June 2011 and 2012 for six models (ensemble mean), the equally weighted multimodel mean, and the observed SPI6 from the CPC-merged gauge radar precipitation analysis. As is done with SPI forecasts, observed March–May (MAM) precipitation is combined with JJA-precipitation forecasts to provide the SPI6 forecast. This methodology of combining 50% observational data with 50% forecast data is described in Quan et al. (2012).

**THE PHASE-2 NMME.** The NMME-2 project was awarded in August 2012 so results to present here are limited. However, there are some specific issues to highlight. In particular, we provide some

preliminary results indicating that both modeling system improvements and data assimilation system improvements will contribute to improved NMME-2 forecast quality. We also describe an example of how some lessons learned regarding the retrospective forecast protocol in NMME-1 contribute to the NMME-2 forecast protocol. Finally, we provide some details regarding the data dissemination strategy on NMME-2.

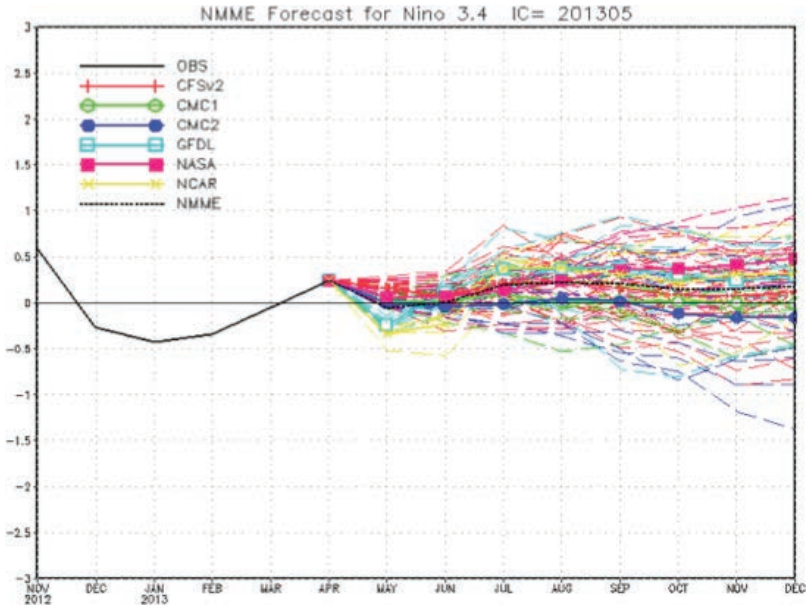


**Fig. 12. Real-time Niño-3.4 predictions initialized in May 2013.**

*Prediction system improvement.* The NMME team will transition from CCSM3 (T85) to CCSM4 (0.9 × 1.25_g1v6 resolution), although if CCSM3 continues to be a useful contributor to the NMME, we will continue the real-time predictions. CCSM4 has significant improvements in the simulation of tropical variability relative to CCSM3 (Neale et al. 2008; Jochum et al. 2008; Gent et al. 2010). The initialization procedure differs from CCSM3 in that we will use the operational Climate Forecast System Reanalysis (CFSR) ocean, land, and atmospheric states to initialize CCSM4 as opposed to ocean-only initialization using optimal interpolation from the Geophysical Fluid Dynamics Laboratory (GFDL) (i.e., Derber and Rosati 1989). We have begun testing the CFSR ocean states in CCSM4 hindcast experiments, and Fig. 15 shows the hindcast SSTA correlation for a parallel set of experiments using CCSM3 with the original GFDL ocean states (bottom panel) and using the CFSR ocean states (top panel). The correlation is notably larger with CCSM4 using CFSR ocean states. We separately examined the impact of the model changes (i.e., CCSM3 vs CCSM4) and the changes associated with the different ocean

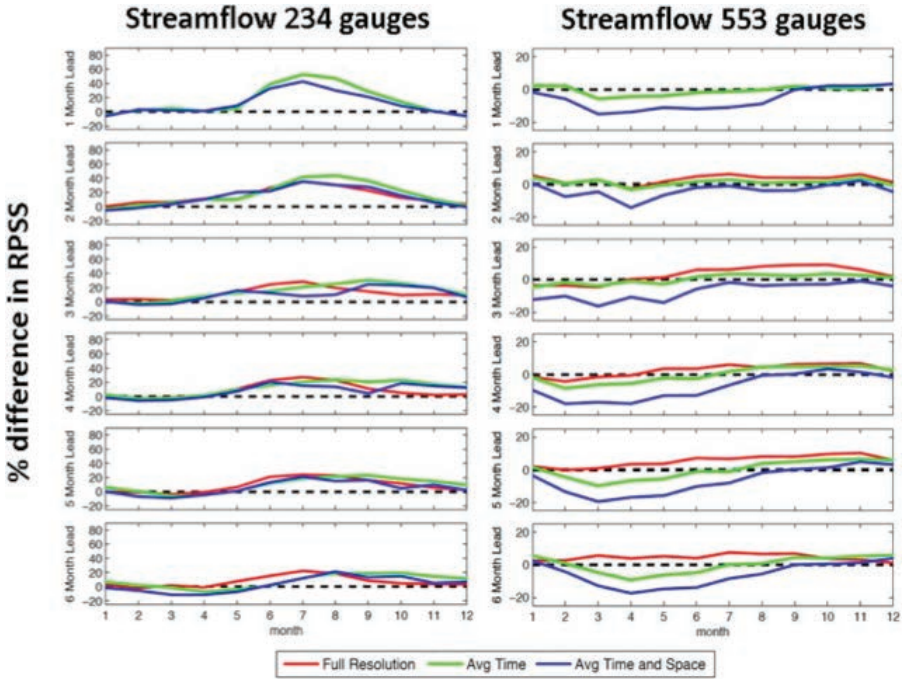


**Fig. 13. Percent difference in RPSS skill of streamflow forecasts over the (left) Colorado NIDIS testbed and (right) southeastern U.S. NIDIS testbed with lead times out to 6 months. Skill differences above 0 indicates NMME forecasts are more skillful than ESP. Full resolution indicates using the downscaled 1/8°, daily seasonal climate model variables; Avg Time indicates the forecasts are averaged over the lead time; and Avg Time and Space indicates that the forecasts are averaged over the lead times and domain.**

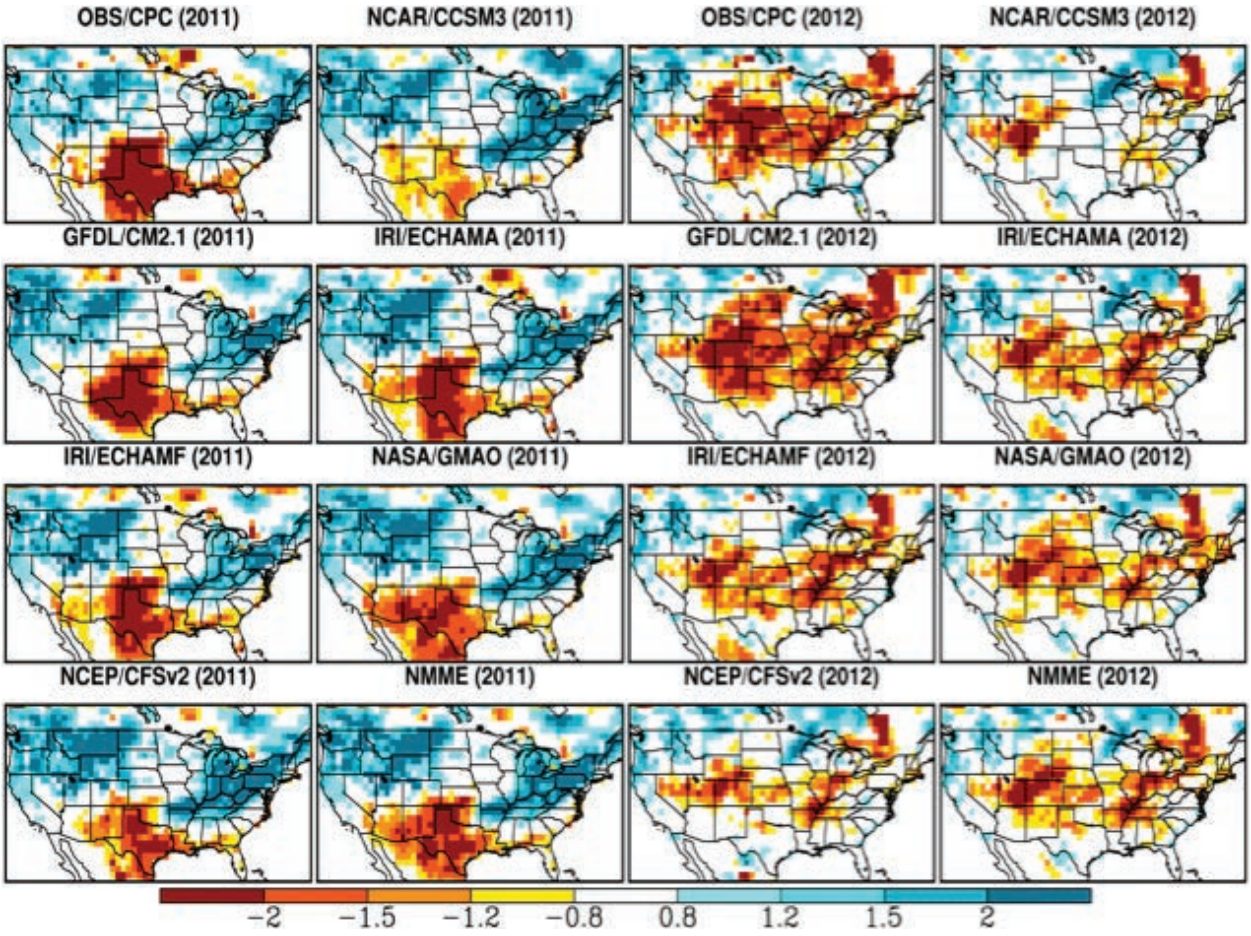


**Fig. 14. NMME SPI6 forecasts initialized 1 Jun 2011 and 2012. Observed MAM precipitation is combined with JJA model ensemble-mean forecast. The NMME forecast is the equally weighted ensemble model average.**

state. Both changes contribute to the increases in the correlation but are dominated by the model changes. We have also developed procedures for using CFSR data for the atmosphere and land initial states (e.g., Paolino et al. 2012).

The GFDL NMME contribution will transition from the GFDL Climate Model, version 2.1 (CM2.1) to the high-resolution coupled GFDL Climate Model, version 2.5 (CM2.5) (described below). The atmospheric component of CM2.5 is derived from the atmospheric component of the coupled GFDL CM2.1. The horizontal resolution has been refined from roughly 200 km to approximately 50 km. The ocean model is substantially different from that used in CM2.1. The ocean grid is considerably finer, with horizontal spacing varying from 28 km at the equator to 8 km in high latitudes. In addition, the grid boxes maintain an aspect ratio close to one, in contrast to CM2.1 where the aspect ratio can exceed 2 at high latitudes due to the convergence of the meridians. The ocean component uses 50 levels in the vertical as in CM2.1. The land model (Dunne et al. 2013) in CM2.5 is called LM3 and represents a major change from the land model used in CM2.1. LM3 is a new model for land water, energy, and carbon balance. The sea ice component used in CM2.5 is almost identical to that used in CM2.1, called the GFDL Sea Ice Simulator (SIS).

*Data dissemination strategy.* One of the major challenges for both NMME-1 and NMME-2 is to provide rapid and open access to all the hindcasts and real-time forecasts. The strategy developed includes two major components. First, NOAA/CPC will obtain and store the monthly-mean data (hindcasts and real-time forecasts) for the three [expanding to eight, that is, SST, precipitation, T2m, 500-mb geopotential, maximum temperature (Tmax), minimum temperature (Tmin), and soil moisture and runoff] required variables from all the participating models, and the IRI will maintain a NMME website serving this minimal dataset to the broader research and applications communities in real time. This rapid and open access to the data is a critical element distinguishing the NMME activity.

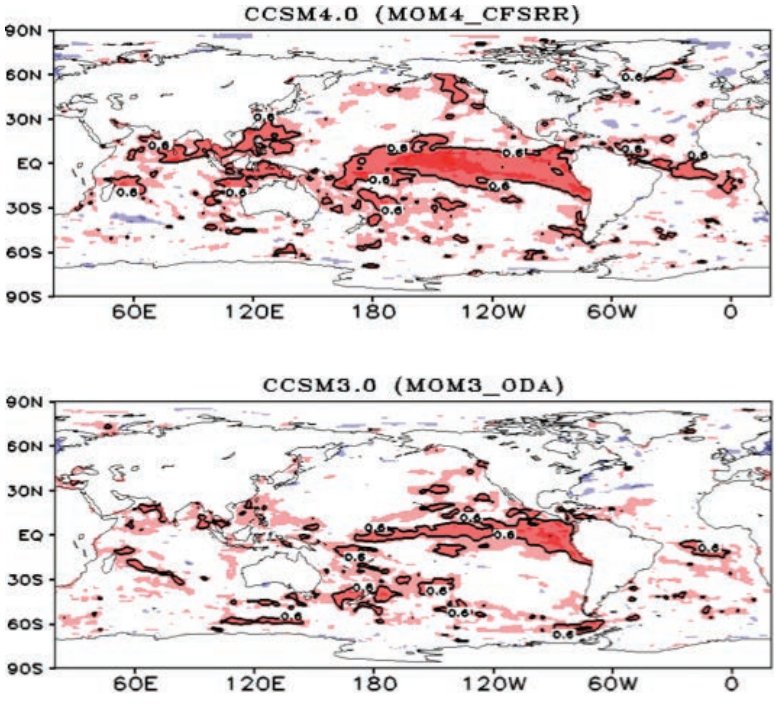


**Fig. 15. SSTA correlation coefficient for forecasts initialized in early Jan and verified for May (1982–2000). The top panel shows results using CCSM4 and CFSR initial states for the ocean and the bottom panel shows results for CCSM3 using MOM3 ODA initial states.**

The second component of the approach recognizes that the data and possibly the number of participating models will grow; a more robust centralized data strategy is required to meet the needs of the broader research and applications communities. As such, we have developed an NMME-2 data server to be housed at the new National Center for Atmospheric Research (NCAR) Wyoming Supercomputing Center (NWSC). This NMME-2 data server will include high frequency (e.g., 3 hourly and daily) and a much more complete three-dimensional distribution of the data.

**NMME-2 RESEARCH.** A major challenge to the NMME experiment is to quantitatively document the success of the project. Here, we briefly summarize some elements of our strategy but also welcome the broader research community to rigorously assess and use the data. Indeed, we assert that making the data readily available to all interested parties is the best approach for evaluating the utility of the multimodel approach advocated here. The measures of success envisioned by the NMME-2 team include a spectrum of quantitative metrics such as forecast skill assessment as a function of the number of models and ensemble members to identifying complementary skill among the models to assessing phenomenological skill.

For example, to determine the forecast skill as a function of the number of models and the number of ensemble members, we will assess a hierarchy of methods of varying complexity using a variety of deterministic and probabilistic verification measures. The deterministic verifications will be applied to the multimodel ensemble-mean forecast, while the probabilistic verifications will be applied to the forecast probabilities of tercile-based categories (hereafter called terciles) and of the extreme 15% tails of the climatological distribution. To facilitate this analysis the NMME project is developing an open access "verification map room" (http://iri.columbia.edu/~tippett/NMME/) that will also be easily accessible via smartphone. The reader is also encouraged to visit this website and the developing reliability website (http://iri.columbia.edu/~shuhua/mis-html/Reliability_nmme.html), both of which are already delivering results.

The above forecast skill assessment is applied without any mechanistic or phenomenological perspective. A second important measure of success is the extent to which we provide a better understanding of the mechanisms and sources of predictive skill. In this second category, we confront the forecasts with observations from a mechanistic and phenomenological perspective that also has the advantage of entraining some additional user communities into the skill assessment. We already have in place commitments to use the NMME data for the U.S. drought briefing, to derive standardized drought precipitation indices (K. Mo 2012, personal communication), and for the emerging Global Drought Information System (GDIS). Feedback from these applications will aid in assessing forecast skill from a drought user perspective, and the use of the NMME data in this regard is a clear measure of success.

An NMME, or any combination of forecast methods, begs the question as to how many models and ensemble members we really need for the problem at hand [this question also comes up in the Intergovernmental Panel on Climate Change (IPCC) context]. For example, do the $N + 1$ models always provide more skill than $N$ models? The NMME phase-2 hindcasts provide an excellent

opportunity to research this issue for subseasonal-to-seasonal time scales (beyond 2 weeks, excluding the weather prediction portion of each forecast period). Well-known notions with respect to the effective number of degrees of freedom in space and time (often approximated by how many EOFs it takes to explain, say, 90% of the variance of a dataset) can be applied here where an additional dimension "space" is taken to be across all the ensemble members. This way we could find that it takes only $n$ models with $k$ ensemble members to describe 90% of the information we have generated by $K$ members of $N$ models. This information content approach can be applied straightforwardly and is directly related to the notion of orthogonality/independence. It will take more originality to combine this with the skill of the forecasts; that is, add the observational dataset (1 single realization) to arrive at those components of a huge forecast dataset that are orthogonal with respect to their ability to add skillful information. These questions and many others can be addressed with the NMME phase-2 data that will be available to researchers beyond the NMME team.

**CONCLUDING REMARKS.** The purpose of this paper is to introduce the weather and climate research and applications communities to the NMME experiment. Here, we have provided a description of the NMME project and its expected evolution over the next 18–24 months (i.e., NMME-2). Part of the description emphasized both deterministic and probabilistic retrospectives in forecast verification. We chose to compare the NMME system (which includes the NOAA operational CFSv2) to CFSv2 alone. This choice was pragmatic and based on addressing the question of whether the NMME project can enhance the NOAA operational system. Overall, the various skill metrics (correlation, RMSE, RPSS, and reliability) all suggest that the NMME system improves the skill over the CFSv2. Admittedly, we have not clearly shown whether the improvement is due to a larger ensemble size or the use of multiple models (or both); nevertheless, the distribution of the forecast production to a number of different groups and centers is an effective strategy for economically increasing the forecast skill.

The assertion that the use of multiple models is an important aspect of the improved skill is supported by a number of previous efforts [e.g., Climate-System Historical Forecast Project (CHFP; www.wcrp-climate.org/wgsip/chfp/index.shtml), North American Ensemble Forecast System (NAEFS; www.emc.ncep.noaa.gov/gmb/ens/NAEFS.html), The Observing System Research and Predictability Experiment (THORPEX) Interactive Grand Global Ensemble (TIGGE; http://tigge.ecmwf.int/), Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER; www.ecmwf.int/research/demeter/index.html), and Ensemble-Based Predictions of Climate Changes and their Impacts (ENSEMBLES; www.ecmwf.int/research/EU_projects/ENSEMBLES/index.html)]. Indeed, much like the NMME activity, the International Multimodel Ensemble [IMME; the IMME project is an expansion of the European Seasonal to Interannual Prediction (EUROSIP) superensemble to include the CFSv2; see www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/eurosip/] is motivated by the results of these early studies. The distinction of the NMME project is twofold. First, the previous efforts focus entirely on retrospective forecasts, whereas the NMME project includes both real-time and retrospective forecasts. Second, the NMME project is committed to provide easy access to all the data (in near–real time), whereas the access to data is restricted in the IMME project. There is an important caveat here; namely, while multimodels' approaches are the pragmatic approach, we recognize that they do not adequately resolve the uncertainty due to model formulation.

Finally, we note that the NMME models that are retained as we enter phase-2 of the project are from major national modeling centers [i.e., NOAA–GFDL, NOAA–NCEP, NASA, NCAR, and the Canadian Meteorological Centre (CMC)], and it is our expectation that these efforts have critical mass in terms of human resources for continued evaluation and testing and that participation by the various NMME partners is mutually beneficial. For example, the project leverages all the model, assimilation, and data development activities at the various centers. The various centers, in turn, test their models against other state-of-the-art prediction systems in both retrospective and real-time mode and potentially have a much wider user community examine the predictions in various applications. We also believe that this continual enhanced collaboration among a broad base of researchers will lead to improved specific operational prediction products. Just as important, the core research collaboration that is at the heart of the NMME project will lead to a better understanding of mechanism of and sources for predictability and better estimates of the inherent limits of predictability. Moreover, some of these national efforts have distinct science missions, and the NMME

project provides common experimental framework to evaluate model performance. Nevertheless, it remains a challenge to demonstrate that the research results from the NMME experiment feedback to model development, and the success of the project should be evaluated in this regard.

**ACKNOWLEDGMENTS.** The phase-1 NMME project was supported by the NOAA MAPP program, and the phase-2 NMME project is support by NOAA MAPP, NSF, NASA, and the DOE.

## REFERENCES

Berner, J., F. J. Doblas-Reyes, T. N. Palmer, G. Shutts, and A. Weisheimer, 2008: Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. *Philos. Trans. Roy. Soc. London,* **A366,** 2561–2579, doi:10.1098/rsta.2008.0033.

Challinor, and Coauthors, 2005: Probabilistic simulations of crop yield over western India using the DEMETER seasonal hindcast ensembles. *Tellus,* **57A,** 498–512.

Derber, J., and A. Rosati, 1989: A global oceanic data assimilation system. *J. Phys. Oceanogr.,* **19,** 1333–1347.

DeWitt, D. G., 2005: Retrospective forecasts of interannual sea surface temperature anomalies from 1982 to present using a directly coupled atmosphere–ocean general circulation model. *Mon. Wea. Rev.,* **133,** 2972–2995.

Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—II. Calibration and combination. *Tellus,* **57A,** 234–252, doi:10.1111/j.1600-0870.2005.00104.x.

Dunne, J. P., and Coauthors, 2013: GFDL's ESM2 global coupled climate–carbon Earth System Models. Part II: Carbon system formulation and baseline simulation characteristics. *J. Climate,* **26,** 2247–2267.

Gent P. R., S. G. Yeager, R. B. Neale, S. Levis, and D. A. Bailey, 2010: Improvements in a half degree atmosphere/land version of the CCSM. *Climate Dyn.,* 34, 819–833.

Goddard, L., S. J. Mason, S. E. Zebiak, C. F. Ropelewski, R. Basher, and M. A. Cane, 2001: Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.,* **21,** 1111–1152.

Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. *Tellus,* **57A,** 219–233, doi:10.1111/j.1600-0870.2005.00103.x.

Jochum M., G. Danabasoglu, M. M. Holland, Y. O. Kwon, and W. G. Large, 2008: Ocean viscosity and climate. *J Geophys Res.,* **113,** C06017, doi:10.1029/2007JC004515.

Kirtman, B. P., 2003: The COLA anomaly coupled model: Ensemble ENSO prediction. *Mon. Wea. Rev.,* **131,** 2324–2341.

——, and D. Min, 2009: Multimodel ensemble ENSO prediction with CCSM and CFS. *Mon. Wea. Rev.,* **137,** 2908–2930.

Luo, L., and E. F. Wood, 2007: Monitoring and predicting the 2007 U.S. drought. *Geophys. Res. Lett.,* **34,** L22702, doi:10.1029/2007GL031673.

——, and ——, 2008: Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. *J. Hydrometeor.,* **9,** 866–884.

——, ——, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.,* **112,** D10102, doi:10.1029/2006JD007655.

Merryfield, W. J., and Coauthors, 2013: The Canadian seasonal to interannual prediction system. Part I: Models and initialization. *Mon. Wea. Rev.,* **141,** 2910–2945.

Mitchell, K. E., and Coauthors, 2004: The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *J. Geophys. Res.,* **109,** D07S90, doi:10.1029/2003JD003823.

Morse, A. P., F. J. Doblas-Reyes, M. B. Hoshen, R. Hagedorn, and T. N. Palmer, 2005: A forecast quality assessment of an end-to-end probabilistic multimodel seasonal forecast system using a malaria model. *Tellus,* **57A,** 464–475, doi:10.1111/j.1600-0870.2005.00124.x.

Neale, R. B., J. H. Richter, and M. Jochum, 2008: The impact of convection on ENSO: From a delayed oscillator to a series of events. *J. Climate,* **21,** 5904–5924, doi:10.1175/2008JCLI2244.1.

Palmer, T. N., and Coauthors, 2004: Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER). *Bull. Amer. Meteor. Soc.,* **85,** 853–872.

——, C. Brankovic, and D. S. Richardson, 2000: A probability and decision-model analysis of PROVOST seasonal multimodel ensemble integrations. *Quart. J. Roy. Meteor. Soc.,* **126,** 2013–2034.

——, F. J. Doblas-Reyes, A. Weisheimer, and M. J. Rodwell, 2008: Toward seamless prediction: Calibration of climate change projections using seasonal forecast. *Bull. Amer. Meteor. Soc.,* **89,** 459–470.

Paolino, D. A., and Coauthors, 2012: The impact of land surface and atmospheric initialization on seasonal forecasts with CCSM. *J. Climate,* **25,** 1007–1021.

Pegion, K., and P. D. Sardeshmukh, 2011: Prospects for Improving Subseasonal Predictions. *Mon. Wea. Rev.,* **139,** 3648–3666, doi:10.1175/MWR-D-11-00004.1.

Quan, X., M. P. Hoerling, B. Lyon, A. Kumar, M. A. Bell, M. K. Tippett, and H. Wang, 2012: Prospects for dynamical prediction of meteorological drought. *J. Appl. Meteor. Climatol.,* **51,** 1238–1252.

Richardson, D. S., 2001: Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quart. J. Roy. Meteor. Soc.,* **127,** 2473–2489, doi:10.1002/qj.49712757715.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–351.

——, and Coauthors, 2014: The NCEP Climate Forecast System version 2. *J. Climate,* in press.

Sheffield, J., G. Goteti, F. Wen, and E. F. Wood, 2004; A simulated soil moisture based drought analysis for the United States, *J. Geophys. Res.,* **109,** D24108, doi:10.1029/2004JD005182.

Vernieres, G., C. Keppenne, M.M. Rienecker, J. Jacob, and R. Kovach, 2012: The GEOS-ODAS, description and evaluation. NASA Technical Report Series on Global Modeling and Data Assimilation, NASA/TM–2012–104606, Vol. 30.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete brier and ranked probability skill scores. *Mon. Wea. Rev.,* **135,** 118–124.

Yuan, X., and E. F. Wood, 2012: On the clustering of climate models in ensemble seasonal forecasting. *Geophys. Res. Lett.,* **39,** L18701, doi:10.1029/2012GL052735.

——, ——, J. K. Roundy, and M. Pan, 2013: CFSv2-based seasonal hydro climatic forecasts over conterminous United States, *J. Climate,* **26,** 4828–4847, doi:10.1175/JCLI-D-12-00683.1.

Zhang, S., M. J. Harrison. A. Rosati, and A. T. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.,* **135,** 3541–3564, doi:10.1175/MWR3466.1.