AMERICAN
METEOROLOGICAL
SOCIETY

Bulletin of the American Meteorological Society

The DOI for this manuscript is doi: 10.1175/BAMS-D-12-00050.1

The North American Multi-Model Ensemble (NMME): Phase-1 Seasonal to Interannual Prediction, Phase-2 Toward Developing Intra-Seasonal Prediction

Ben P. Kirtman*, Dughong Min and Johnna M. Infanti
University of Miami, Rosenstiel School for Marine and Atmospheric Science

James L. Kinter III and Daniel A. Paolino
Center for Ocean-Land-Atmosphere Studies

Qin Zhang, Huug van den Dool, Suranjana Saha, Malaquias Pena Mendez, Emily Becker, Peitao Peng, Patrick Tripp and Jin Huang
NOAA National Centers for Environmental Prediction

David G. DeWitt#, Michael K. Tippett, Anthony G. Barnston, Shuhua Li
International Research Institute for Climate and Society

Anthony Rosati
NOAA Geophysical Fluid Dynamics Laboratory

Siegfried D. Schubert, Michele Rienecker, Max Suarez, Zhao E. Li, Jelena Marshak, Young-Kwon Lim
NASA Goddard Space Flight Center

Joseph Tribbia
National Center for Atmospheric Research

Kathleen Pegion
CIRES – University of Colorado

William J. Merryfield, Bertrand Denis
Environment Canada

Eric F. Wood
Princeton University

*Corresponding Author:
Rosenstiel School for Marine and Atmospheric Science
4600 Rickenbacker Causeway
Miami FL 33149
bkirtman@rsmas.miami.edu

#Current Affiliation: NOAA National Weather Service

Abstract

The recent US National Academies report "Assessment of Intraseasonal to Interannual Climate Prediction and Predictability" was unequivocal in recommending the need for the development of a North American Multi-Model Ensemble (NMME) operational predictive capability. Indeed, this effort is required to meet the specific tailored regional prediction and decision support needs of a large community of climate information users.

The multi-model ensemble approach has proven extremely effective at quantifying prediction uncertainty due to uncertainty in model formulation, and has proven to produce better prediction quality (on average) then any single model ensemble. This multi-model approach is the basis for several international collaborative prediction research efforts, an operational European system and there are numerous examples of how this multi-model ensemble approach yields superior forecasts compared to any single model.

Based on two NOAA Climate Test Bed (CTB) NMME workshops (February 18, and April 8, 2011) a collaborative and coordinated implementation strategy for a NMME prediction system has been developed and is currently delivering real-time seasonal-to-interannual predictions on the NOAA Climate Prediction Center (CPC) operational schedule. The hindcast and real-time prediction data is readily available (e.g., http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/) and in graphical format from CPC (http://origin.cpc.ncep.noaa.gov/products/people/wd51yf/NMME/index.html). Moreover, the NMME forecast are already currently being used as guidance for operational forecasters. This paper describes the new NMME effort, presents an overview of the multi-model forecast quality, and the complementary skill associated with individual models.

71 **Capsule Summary:** The paper describes the North American Multi-Model Ensemble prediction
72 experiment include how to access the data in digital and graphical form, and some discussion of
73 forecast quality.
74
75 **1. Introduction**
76 After more than three decades of research into the origins of seasonal climate
77 predictability and the development of dynamical model-based seasonal prediction systems, the
78 continuing relatively deliberate pace of progress has inspired two notable changes in prediction
79 strategy, largely based on multi-institutional international collaborations. One change in strategy
80 is the inclusion of quantitative information regarding uncertainty (i.e., probabilistic prediction) in
81 forecasts and probabilistic measures of forecast quality in the verifications (e.g., Palmer et al.
82 2000; Goddard et al. 2001; Kirtman 2003; Palmer et al. 2004; DeWitt 2005; Hagedorn et al.
83 2005; Doblas-Reyes et al. 2005; Saha et al. 2006 among many others). The other change is the
84 recognition that a multi-model ensemble strategy is a viable approach for adequately resolving
85 forecast uncertainty (Palmer et al. 2004; Hagedorn et al. 2005; Doblas-Reyes 2005; Palmer et al.
86 2008), although other techniques such as perturbed physics ensembles (currently in use at the
87 UK Met Office for their operational system) or stochastic physics (e.g., Berner et al., 2008) have
88 been developed and appear to be quite promising. The first change in prediction strategy
89 naturally follows from the fact that climate variability includes a chaotic or irregular component,
90 and, because of this, forecasts must include a quantitative assessment of this uncertainty. More
91 importantly, the climate prediction community now understands that the potential utility of
92 climate forecasts is based on end-user decision support (Palmer et al., 2000; Morse et al. 2005;
93 Challinor et al. 2005), which requires probabilistic forecasts that include quantitative information

94 regarding forecast uncertainty. The second change in prediction strategy follows from the first,
95 because, given our current modeling capabilities, a multi-model strategy is a practical and
96 relatively simple approach for quantifying forecast uncertainty due to uncertainty in model
97 formulation, although it is likely that the uncertainty is not fully resolved.

98 More recently, there has been a growing interest in forecast information on time scales
99 beyond 10 days, but less then a season. For example, the Climate Prediction Center of the
100 National Centers for Environmental Prediction (NCEP/CPC) in the United States currently
101 makes "outlook" type forecasts for extended weather forecast ranges (i.e., two weeks) such as
102 the NCEP/CPC *Global Tropical Hazards/Benefits Assessment* provides forecasts of anomalous
103 tropical temperature and precipitation. The *U.S. Hazards Assessment* product, also issued by
104 NCEP/CPC, includes outlooks of potential hazards in the U.S up to 16 days. At present such
105 outlook-style forecast products are based on a subjective combination of various statistical and
106 dynamical methods, although there is momentum to make the process more objective using real-
107 time dynamic model forecasts (Gottshalck 2008). These developments demonstrate the demand
108 for such dynamical forecast information.

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

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

Given the pragmatic utility of the multi-model approach, there is multi-agency (NOAA, NSF, NASA, and DOE) support for a North American Multi-Model Ensemble (NMME) Intra-seasonal to Seasonal to Inter-annual (ISI) prediction experiment. This experiment leverages an NMME team that has already formed and began producing routine real-time multi-model ensemble ISI predictions since August 2011. The forecasts are provided to the NOAA Climate Prediction Center (CPC) on an experimental basis for evaluation and consolidation as a multi-model 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 season-to-interannual time-scales in that the data that is exchanged is monthly.

The newly funded multi-agency 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-I project.

The NMME-2 experiment will:

i. Build on existing state-of-the-art US 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 new forecast system, and ensure interoperability so as to easily incorporate future model developments.

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 US 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 intra-seasonal or sub-seasonal multi-model 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 multi-model forecast quality assessment and our strategy for evaluating how the multi-model approach contributes to the forecast quality. We also describe how NMME-2 will evolve from NNME-1 and the coordinated research activities and data dissemination strategy envisaged.

2. The Phase 1 NMME

Based on two Climate Test Bed (CTB) NMME workshops (February 18, and April 8, 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.

a. 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:

- 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 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-meter Temperature (T2m), and precipitation rate. More fields will be added based on experience and demand. It is also recognized that higher frequency data is desirable and this will be implemented as feasible.
- Routine real-time forecast data must be available by the 8th of each month.

185 The NMME-1 activity began in February 2011 and became an experimental real-time system
186 in August 2011. Specifically, on August 8th, NCEP (CPC and EMC) collected from the
187 respective ftp sites of the NMME partners the real time seasonal predictions. In the months
188 before August 2011 the hindcast data were collected, and climatologies and skill assessments for
189 each model to be applied to subsequent real time predictions were calculated. Graphical forecast
190 guidance based on the NMME was prepared and given to NOAA operational forecasters in time
191 for the CPC seasonal prediction cycle. The graphical forecast guidance includes North American
192 and global domains, and T2m (T), Precipitation (P) and SST fields, and the plots are for monthly
193 and seasonal means with and without a skill mask applied. All NMME forecasts are bias-
194 corrected (making use of the hindcasts) using cross-validation (see Kirtman and Min 2009 for
195 details of how to make the bias correction).
196 The effort is significant because, although experimental, the NMME protocol adheres to
197 CPC's operational schedule, so the forecasters can use the information for operational guidance.
198 The scripts for the data ingest and graphical outputs are intended to be robust, i.e. any number of
199 models, with any number of ensemble members can be used. A major element of the NMME
200 experiment is to continue this effort for the benefit of operations. Meanwhile we have built up a
201 "live" hindcast data set of about 30 years that is open to anybody, and can be used for research.
202 Quite probably, this NMME data set is now the most extensive multi-model seasonal prediction
203 archive currently available that includes models that are continuing to make real-time
204 predictions. The following table summarizes the NMME-1 hindcast data sets and identifies the
205 point of contact for each prediction system.
206 In addition, NOAA/CPC has agreed to evaluate the hindcasts, combine the forecasts,
207 perform verification, provide an NMME web site (http://cpc.ncep.noaa.gov/products/NMME/)

208 and make the real-time NMME forecast delivery to NOAA forecasters. CPC is also maintaining
209 a NMME newsletter. The hindcast data and real-time forecast data is also available for download
210 or analysis at the IRI (http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/). The CPC site
211 primarily serves the real-time needs of the project, and the IRI site, along with the analysis tools
212 that are being developed at the IRI
213 (http://iridl.ldeo.columbia.edu/home/.tippett/.NMME/.Verification/), primarily serves research
214 needs in terms of assessing the prediction skill and predictability limits associated with NMME-I
215 and in terms of designing the NMME-II experimental protocol. While the NMME-I data is
216 limited to monthly mean data, it is a research tool (or test-bed) that is proving extremely useful in
217 supporting the basic prediction and predictability research needs of the project participants. This
218 database also serves as "quick look" easy access data that is the external face of the NMME
219 experiment to the research community.
220
221 *b. Results – NMME-1*
222 Here we show some results from 28 years of hindcasts that cover a common period (i.e.,
223 1982-2009) for all the models, and the real-time experimental forecast from the NMME of
224 opportunity (i.e., NMME-1). The results help provide evidence of the benefit of a multi-model
225 ensemble of predictions, as compared with the ensemble predictions of just one high performing
226 model. Figure 1 shows the range spanned by the individual ensemble members from each
227 forecast system in NMME-1, for 0.5-month lead[2] hindcasts for the Nino3.4 SST index. This
228 presentation of the range assumes that each ensemble member of each model is equally likely to
229 occur. In order to calculate anomalies, the forecast bias or systematic error has been removed and

[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 1.5 month lead and so on. The retrospective forecasts also follow this convention.

230 is calculated separately for each model using all ensemble members for that particular model.
231 See Saha et al. (2006) or Kirtman and Min (2009) for discussion of how the systematic error is
232 removed. At this short lead-time the hindcasts tend to agree with one another and with the
233 observations, to a great extent, although there is also some disagreement, particularly at certain
234 times (e.g. near the end of 1988 and in the middle of 1998). However, it is worth noting that
235 nowhere do the observations lie noticeable outside the envelope of the predictions.
236 Figure 2 shows the same results except for 5.5-month lead predictions, with appropriately
237 greater uncertainties shown by the larger range - often in excess of 2°C. We will show that it is
238 just such dispersion in the individual predictions that best reflects forecast uncertainty, as well as
239 the "best guess" multi-model mean prediction.
240 Figure 3 shows the spatial distribution of the anomaly correlation between the 5.5-month
241 lead of the grand ensemble monthly mean hindcast and observed SST over 1982-2009. Here the
242 grand ensemble mean is defined as the average of all the hindcasts assuming that each ensemble
243 member of each model is equally probable. This is distinct from assuming that each model
244 should be weight equally. High skill is evident in the central and eastern tropical Pacific Ocean,
245 as well as portions of the tropical Atlantic and Indian oceans and some isolated regions in the
246 extratropics.
247 One of the important motivating factors for both phases on the NMME project is to
248 understand the complementary sources of skill among the models. Essentially, we seek to
249 understand the "where and why" in how the multi-model approach improves forecast quality.
250 Here we show the first step in this process – simply documenting how the multi-model compares
251 to any single model. For example, Fig. 4 shows scatterplots of the root mean squared error of the
252 SSTA for individual models 0.5-to-5.5 month lead ensemble mean hindcasts versus the

253 corresponding multi-model ensemble mean hindcasts for tropical SST for September starts. The
254 percentage noted in each panel corresponds to the number of points where the individual model
255 beat the multi-model. For every single individual model most of the points are above the
256 diagonal (i.e., the percentage of points below the diagonal is less than 50%), indicating that the
257 multi-model tends to have smaller errors than the individual models. Generally, the models
258 cluster around 26-48%. CCSM3 is an outlier, and is being replaced with CCSM4 in NMME-2.
259 Preliminary examination (not shown) has suggested that in general, the individual model
260 having highest anomaly correlation skill is CFSv2. However, this identification of the generally
261 best model does not suggest that the other models, when allowed to contribute to the multi-model
262 mean forecast, do not further enhance the performance. To demonstrate the benefit reaped by
263 using the multi-model ensemble over the single best performing model, the Ranked Probability
264 Skill Score (RPSS)[3] of the multi-model ensemble hindcasts and the CFSv2 hindcasts of SST for
265 DJF for forecasts initialized in early July are shown in Fig. 5, while those for JJA initialized in
266 early January are shown in Fig. 6. In the case of both seasons, the multi-model ensemble
267 produces higher mean skill. There are isolated areas where CFSv2 outperforms the multi-model
268 ensemble, such as in the DJF forecasts (Fig. 5) just south of the equator near 85°, south of Sri
269 Lanka. However, the multi-model ensemble has higher, and more reliably positive, skill in over
270 most of the globe that of any of the individual model forecasts—even the best of them.
271 The comparatively better RPSS results of the multi-model ensemble hindcasts than those
272 of the CFSv2 forecasts are not limited to SST hindcasts, but generalize to predictions for land
273 surface temperature and precipitation as well. Figure 7, for example, shows the spatial
274 distribution of RPSS for land surface temperature for JJA initialized in early January for the

[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 using the equal-odds climatology forecasts as the reference forecast). A good rule of thumb is that an RPSS of 0.08 corresponds to deterministic correlation of 0.4.

multi-model ensemble (top) and CFSv2 (bottom). Again, the multi-model mean has considerably less area with negative skill while maintaining the skill levels at many of the areas where CFSv2 has highest skill. Multi-model skill at the locations of the most extreme peaks of CFSv2 skill tends to be slightly attenuated (e.g. northeastern Brazil, 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 (left) the multi-model ensemble and (right) CFSv2 alone. Figure 9 is the same a Fig. 8, but for JJA season (initialized in Janurary). The comparative superiority of the multi-model forecast over CFSv2 alone is noted for both seasons. This is most obvious in the relative lack of negative skill in the multi-model 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, a secondary reason for the relatively better performance of the multi-model hindcasts than CFSv2 is the much larger ensemble size of all the models together than of any single model.

298 A tool used to diagnose a set of probabilistic forecasts is reliability analysis, which
299 measures the correspondence between the forecast probabilities and their subsequent observed
300 relative frequencies, spanning the full range of issued forecast probabilities for each of the three
301 climatologically equi-probable categories (below, near, or above normal). If one collected all
302 instances of forecasts of 45% probability for "above normal", for example, and that category
303 were actually later observed in 45% of the cases, the forecasts for that particular probability bin
304 would be shown to have perfect reliability. Results of reliability analysis for forecasts initialized
305 in October and verifying in the following JFM for 2-meter temperature anomalies over the globe
306 are shown in Fig. 10 for the multi-model ensemble hindcasts over the 28-year period for the
307 below-normal and above normal categories. The light dotted line denotes perfect reliability.
308 Two aspects of common interest in reliability diagnosis are (1) the overall position of the
309 lines relative to the ideal 45° line, and (2) the slope of the lines relative to unity. The general
310 positions of the lines in Fig. 10 are near that of the ideal line, but the line representing above
311 (below) normal forecasts is just slightly higher (lower) than ideal. This indicates a slight
312 tendency to under-forecast above normal and over-forecast below normal temperature. The
313 observed mean relative frequency of occurrence of the categories, shown as colored dots on the
314 y-axis, indicates that above normal occurred in about 39% of cases, while below normal (and
315 near normal) occurred in about 30% of cases. However, this weak shift toward above normal
316 temperature in the mean climate over the 28-year period was induced by a slight offset in the
317 base period of the observations and the model hindcasts: for the observations the period is 1981-
318 2010, while for the model forecasts it is 1982-2009. Thus, the overall position of the reliability
319 curves, while usually indicative of the model bias, is influenced here by the slight model versus
320 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 multi-model 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 multi-model ensemble is illustrated by an example of a recent real-time prediction of the Nino3.4 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 multi-model 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 multi-model 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, probabilistically reliable prediction.

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

344 Standardized Precipitation Index (SPI) determined over a window centered on the initial forecast
345 date. Agricultural drought focuses on soil moisture deficits or indices such as their percentiles
346 (Sheffield et al., 2004) and hydrological drought on streamflow. Collectively under the NMME
347 project seasonal hydrologic forecasting will include drought forecasting as well as related
348 hydrological seasonal forecasting such as persistent wet conditions. Since hydrological
349 applications usually require information at smaller spatial scales than that provided by the
350 seasonal forecast models, the climate forecasts from the multi-model ensemble will be
351 downscaled and bias corrected, using the approach of Luo et al. (2007), and used to drive a
352 calibrated land surface model. The output of the land surface model is then used to for
353 hydrologic forecasts, including drought. This approach has been well developed (Lou and
354 Wood, 2007, 2008; Yuan et al., 2013). Figure 13 shows the results for streamflow forecast skill
355 from NMME relative to the skill from the often used Extended Streamflow Prediction (ESP)
356 approach where hydrological model forcings come from historical resampling. The results are
357 presented over the National Drought Information System (NIDIS) Colorado and SE US testbeds.
358 For the Colorado domain, NMME is more skillful than ESP, particularly in the summer with the
359 skill coming primarily from increased precipitation skill. Not shown is the comparison between
360 CFSv2 alone and NMME in which CFSv2 has slightly lower precipitation skill. For the SE
361 NIDIS domain, ESP is more skillful in for month-1 leads due to low NMME precipitation skill,
362 but the situation changes for longer leads when the full resolution dowscaled, bias corrected
363 forecasts are used in the hydrological model. For both ESP and NMME hydrological forecasts,
364 observed hydrologic initial states are used at the initial forecast time. These can be provided
365 from the National Land Data Assimilation System (NLDAS) (Mitchell et al., 2004).

For meteorological drought assessed at continental-to-global scales, the 1-degree NMME model precipitation forecasts can be used. Figure 14 shows the NMME 6-month SPI (SPI6) forecast initiated on the 1st of June 2011 and 2012 for 6 models (ensemble mean), the equally weighted multi-model mean, and the observed SPI6 from the CPC merged gauge-radar precipitation analysis. As is done with SPI forecasts, observed 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).

3. 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 here. 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.

a. Prediction System Improvement

The NMME team will transition from CCSM3 (T85) to CCSM4 (0.9x1.25_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 CCSM3.0 (Neale et al. 2008; Jochum et al. 2008; Gent et al. 2009). The initialization procedure differs from CCSM3 in that we will use the operational CFSR

389 ocean, land and atmospheric states to initialize CCSM4 as opposed to ocean only initialization
390 using optimal interpolation from GFDL (i.e., Derber and Rosati 1989). We have begun testing
391 the CFSR ocean states in CCSM4 hindcast experiments, and Fig. 15 shows the hindcast SSTA
392 correlation for a parallel set of experiments using CCSM3 with the original GFDL ocean states
393 (bottom panel) and using the CFSR ocean states (top panel). The correlation is noteable larger
394 with CCSM4 using CFSR ocean states. We separately examined the impact of the model
395 changes (i.e., CCSM3 vs. CCSM4) and the changes associated with the different ocean state.
396 Both changes contribute to the increases in the correlation, but are dominated by the model
397 changes. We have also developed procedures for using CFSR data for the atmosphere and land
398 initial states (e.g., Paolino et al. 2012).
399 The GFDL NMME contribution will transition from the CM2.1 model to the high-
400 resolution coupled model CM2.5 (described below). The atmospheric component of CM2.5 is
401 derived from the atmospheric component of the GFDL CM2.1 coupled model. The horizontal
402 resolution has been refined from roughly 200 km to approximately 50 km. The ocean model is
403 substantially different from that used in CM2.1. The ocean grid is considerably finer, with
404 horizontal spacing varying from 28 km at the equator to 8 km in high latitudes. In addition, the
405 grid boxes maintain an aspect ratio close to one, in contrast to CM2.1 where the aspect ratio can
406 exceed 2 at high latitudes due to the convergence of the meridians. The ocean component uses 50
407 levels in the vertical as in CM2.1. The land model (Dunne et al. 2013) in CM2.5 is called “LM3”
408 and represents a major change from the land model used in CM2.1. LM3 is a new model for land
409 water, energy, and carbon balance. The sea ice component used in CM2.5 is almost identical to
410 that used in CM2.1, called the GFDL Sea Ice Simulator (SIS).
411

b. 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 the 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, Tmax, Tmin, Soil Moisture and Runoff) required variables from all the participating models and the IRI will maintain a NMME web site serving this minimal data set 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. 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 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.

4. 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 multi-model 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

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, and 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 multi-model 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 smart phone. The reader is also encouraged to visit this web site and the developing reliability web site (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 US drought briefing, to derive standardized drought precipitation indices (K. Mo 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.

458 An NMME, or any combination of forecast methods, begs the question as to how many
459 models and ensemble members we really need for the problem at hand (this question also comes
460 up in the IPCC context). For example, does the N+1 models always provide more skill than N
461 models? The NMME phase-2 hindcasts provide an excellent opportunity to research this issue
462 for sub-seasonal to seasonal time scales (beyond 2 weeks, excluding the weather prediction
463 portion of each forecast period). Well-known notions with respect to the effective number of
464 degrees of freedom in space and time (often approximated by how many EOFs it takes to explain
465 say 90% of the variance of a data set) can be applied here where an additional dimension 'space'
466 is taken to be across all the ensemble members. This way we could find that it takes only n
467 models with k ensemble members to describe 90% of the information we have generated by K
468 members of N models. This information content approach can be applied straightforwardly and is
469 directly related to the notion of orthogonality/independence. It will take more originality to
470 combine this with the skill of the forecasts, i.e. add the observational data set (1 single
471 realization) to arrive at those components of a huge forecast data set that are orthogonal with
472 respect to their ability to add skillful information. These questions and many others can be
473 addressed with the NMME phase-2 data that will be available to researchers beyond the NMME
474 team.
475
476 **5. Concluding Remarks**
477 The purpose of this paper is to introduce the weather and climate research and
478 applications communities to the NMME experiment. Here we have provided a description of the
479 NMME project and its expected evolution over the next 18 to 24 months (i.e., NMME-II). Part of
480 the description emphasized both deterministic and probabilistic retrospectives in forecast

481 verification. We chose to compare the NMME system (which includes the NOAA operational
482 CFSv2) to CFSv2 alone. This choice was pragmatic and based on addressing the question of
483 whether the NMME project can enhance the NOAA operational system. Overall, the various
484 skill metrics (correlation, RMSE, RPSS and reliability) all suggest that the NMME system
485 improves the skill over the CFSv2. Admittedly, we have not clearly shown whether the
486 improvement is due to a larger ensemble size or the use of the multiple models (or both);
487 nevertheless, the distribution of the forecast production to a number of different groups and
488 centers is an effective strategy for economically increasing the forecast skill.

489 The assertion that the use of multiple models is an important aspect of the improved skill
490 is supported by a number of previous efforts (e.g., CHFP[4], NAEFS[5], TIGGE[6], DEMETER[7],
491 ENSEMBLES[8]). Indeed, much like the NMME activity, the International Multi-Model Ensemble
492 (IMME)[9] is motivated by the results of these early studies. The distinction of the NMME project
493 is two fold. First, the previous efforts focus entirely on retrospective forecasts, whereas the
494 NMME project includes both real-time and retrospective forecasts. Second, the NMME project is
495 committed to provide easy access to all the data (in near real-time), whereas the access to data is
496 restricted in the IMME project. There is an important caveat here, namely, while multi-models
497 approaches are pragmatic approach, we recognize that they do not adequately resolve the
498 uncertainty due to model formulation.

499 Finally, we note that the NMME models that are retained as we enter phase-2 of the
500 project are from major national modeling centers (i.e., NOAA-GFDL, NOAA-NCEP, NASA,

[4] http://www.wcrp-climate.org/wgsip/chfp/index.shtml
[5] http://www.emc.ncep.noaa.gov/gmb/ens/NAEFS.html
[6] http://tigge.ecmwf.int/
[7] http://www.ecmwf.int/research/demeter/index.html
[8] http://www.ecmwf.int/research/EU_projects/ENSEMBLES/index.html
[9] The IMME project is an expansion of the EUROSIP (http://www.ecmwf.int/products/forecasts/d/charts/seasonal/forecast/eurosip/) to include the CFSv2.

501 NCAR, CMC) and it is our expectation is that these efforts have critical mass in terms of human
502 resources for continued evaluation and testing, and that participation by the various NMME
503 partners is mutually beneficial. For example, the project leverages all the model, assimilation and
504 data development activities at the various centers. The various centers, in turn, test their models
505 against other state-of-the-art prediction systems in both retrospective and real-time mode, and
506 potentially have a much wider user community examine the predictions in various applications.
507 We also believe that this continual enhanced collaboration among a broad base of researchers
508 will lead to improved specific operational prediction products. Just as important, the core
509 research collaboration that is at the heart of the NMME project will lead to a better
510 understanding of mechanism of and sources for predictability and better estimates of the inherent
511 limits of predictability. Moreover, some of these national efforts have distinct science missions,
512 and the NMME project provides common experimental framework to evaluate model
513 performance. Nevertheless, it remains a challenge to demonstrate that the research results from
514 the NMME experiment feedback to model development, and the success of the project should be
515 evaluated in this regard.

Acknowledgements: The phase-I NMME project was supported by the NOAA MAPP program, and phase-II NMME project is support by NOAA-MAPP, NSF, NASA and the DOE.

6. References

518 Anderson, J., and co-authors, 2009: The Data Assimilation Research Testbed, A community
519 facility. BAMS, 90, 1283-1296, doi: 10.1175/2009BAMS2618.1
520 Barnston, A. G., M. Glantz, and Y. He, 1999: Predictive skill of statistical and dynamical climate
521 models in SST forecasts during the 1997–98 El Nino and the 1998 La Nina onset. *Bull.*
522 *Amer. Meteor. Soc.,* **80,** 217–243.

Berner, J., and co-authors, 2008 Impact of a quasi-stochastic cellular automaton backscatter scheme on the systematic error and seasonal prediction skill of a global climate model. Phil. Trans. R. Soc. A 366, 2561–2579. (doi:10.1098/rsta.2008.0033)

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

Danabasoglu, G., and coauthors, 2006: Diurnal coupling in the tropical oceans of CCSM3. *J. Climate*, **19**, 2347-2365.

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

Dunne, John 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. doi: http://dx.doi.org/10.1175/JCLI-D-12-00150.1

Goddard, L., and co-authors, 2001: Current approaches to seasonal-to-interannual climate predictions. *Int. J. Climatol.,* **21,** 1111–1152.

Griffies, S. M., and Coauthors, 2005: Formulation of an ocean model for global climate simulations. Ocean Science,45-79.

546 Hagedorn, R. F. J. Doblas-Reyes, T. N. Palmer, 2005: The rationale behind the success of multi-
547 model ensembles in seasonal forecasting - I. Basic concept *Tellus A*, **57**, 219–233
548 doi:10.1111/j.1600-0870.2005.00103.x
549 Hunke, E. C., and W.H. Lipscomb, 2008: CICE: The Los Alamos Sea Ice Model, Documentation
550 and Software Manual, Version 4.0. Technical Report, Los Alamos National Laboratory.
551 Kirtman, B. P., 2003: The COLA anomaly coupled model: Ensemble ENSO prediction. *Mon.*
552 *Wea. Rev*., **131**, 2324-2341.
553 Kirtman, B. P., and D. Min, 2009: Multi-model ensemble ENSO prediction with CCSM and
554 CFS. *Mon. Wea. Rev*., DOI: 10.1175/2009MWR2672.1.
555 Krishnamurti, T.N., and co-authors, 2000: Multi-Model ensemble forecasts for weather and
556 seasonal climate. *J. Climate*, **13,** 4196-4216.
557 Lavers, D., L. Luo, and E. F. Wood, 2009: A multiple model assessment of seasonal climate
558 forecast skill for applications. *Geophys. Res. Lett.*, **36**, L23711,
559 doi:10.1029/2009GL041365.
560 Landsea, C. W., and J. A. Knaff, 2000: How much skill was there in forecasting the very strong
561 1997–98 El Nino? *Bull. Amer. Meteor. Soc.,* **81,** 2107–2120.
562 Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing
563 uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.*, **114**, D04114,
564 doi:10.1029/2008JD010969.
565 Li, H., J. Shefffield, and E. F. Wood, 2010: Bias correction of monthly precipitation and
566 temperature fields from Intergovernmental Panel on Climate Change AR4 models using
567 equidistant quantile matching. *J. Geophys. Res.*, **115**, D10101,
568 doi:10.1029/2009JD012882.

569 Luo, L., and E. F. Wood, 2006: Assessing the idealized predictability of precipitation and
570 temperature in the NCEP Climate Forecast System. *Geophys. Res. Lett.*, **33**, L04708,
571 doi:10.1029/2005GL025292.
572 Luo, L., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts
573 for seasonal hydrological predictions. *J. Geophys. Res.*,**112**, D10102,
574 doi:10.1029/2006JD007655.
575 Luo, L., and E. F. Wood, 2007: Monitoring and predicting the 2007 U.S. drought. *Geophys. Res.*
576 *Lett.*, **34**, L22702, doi:10.1029/2007GL031673
577 Luo L. and E. F. Wood, 2008: Use of Bayesian merging techniques in a multimodel seasonal
578 hydrologic ensemble prediction system for the Eastern United States. *Journal of*
579 *Hydrometeorology*, **9**, 866-884.
580 Merryfield, W. J., W.-S. Lee, G. J. Boer, V. V. Kharin, J. F. Scinocca, G. M. Flato, R. S.
581 Ajayamohan, J. C. Fyfe, Y. Tang, and S. Polavarapu, 2013. The Canadian Seasonal to
582 Interannual Prediction System. Part I: Models and Initialization, Monthly Weather
583 Review, in press.
584 Mitchell, K.E., et al., 2004: The multi-institution North American Land Data Assimilation
585 System (NLDAS): Utilizing multiple GCIP products and partners in a continental
586 distributed hydrological modeling system. *J. Geophys. Res.* **109**(D7):
587 10.1029/2003JD003823.
588 Morse, A. P., and co-authors, 2005: A forecast quality assessment of an end-to-end probabilistic
589 multi-model seasonal forecast system using a malaria model *Tellus A* **57**, 464–475
590 doi:10.1111/j.1600-0870.2005.00124.x

591 Pacanowski, R. C., and S. M. Griffies, 1998: MOM 3.0 manual. NOAA/Geophysical Fluid
592 Dynamics Laboratory, Princeton, NJ, 608 pp.
593 Palmer, T. N., C . Brankovic, and D. S. Richardson, 2000: A probability and decision-model
594 analysis of PROVOST seasonal multimodel ensemble integrations. *Quart. J. Roy.*
595 *Meteor. Soc.,* **126,** 2013–2034.
596 Palmer, T.N., and Coauthors, 2004: Development of a European multi-model ensemble system
597 for seasonal-to-interannual prediction (DEMETER), *Bull. Amer. Meteor. Soc*., **85,** 853-
598 872.
599 Palmer, T. N., and co-authors, 2008: Toward seamless prediction: Calibration of climate change
600 projections using seasonal forecast, *Bull. Amer. Metero. Soc*., **89**, 459-470.
601 Paolino, D.A., and co-authors, 2011: The Impact of Land Surface and Atmospheric Initialization
602 on Seasonal Forecasts with CCSM. Journal of Climate, in press.
603 doi:10.1175/2011JCLI3934.1
604 Vernieres, G., and co-authors: The GEOS-ODAS, description and evaluation. NASA Technical
605 Memorandum (in preparation).
606 Quan, X., M.P Hoerling, B. Lyon, A. Kumar, M.A. Bell, M.K. Tippett, and H. Wang, 2012:
607 Prospects for Dynamical Prediction of Meteorological Drought. *Journal of*
608 *Applied Meteorology and Climatology*, 51, 1238-1252.
609 Richardson, D. S. (2001), Measures of skill and value of ensemble prediction systems, their
610 interrelationship and the effect of ensemble size. Q.J.R. Meteorol. Soc., 127: 2473–2489.
611 doi: 10.1002/qj.49712757715.
612 Roeckner, E., and Coauthors, 1996: The atmospheric general circulation model ECHAM4:
613 Model description and simulation of present day climate. Rep. 218, Max-Planck-Institute

fur Meteorologie, 90 pp. [Available from MPI fur Meteorologie, Bundesstr. 55, 20146 Hamburg, Germany.]

Suranjana Saha, Shrinivas Moorthi, Xingren Wu, Jiande Wang, Sudhir Nadiga, Patrick Tripp, David Behringer, Yu-Tai Hou, Hui-ya Chuang, Mark Iredell, Michael Ek, Jesse Meng, Rongqian Yang, Malaquias Pena Mendez, Huug van den Dool, Qin Zhang, Wanqiu Wang, Mingyue Chen, Emily Becker, 2013 : The NCEP Climate Forecast System Version 2. Journal of Climate, under review, revised

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

Weigel, Andreas P., Mark A. Liniger, Christof Appenzeller, 2007: The Discrete Brier and Ranked Probability Skill Scores. *Mon. Wea. Rev.*, **135**, 118–124. doi: http://dx.doi.org/10.1175/MWR3280.1

Yuan, X, E F Wood, 2012. On the clustering of climate models in ensemble seasonal forecasting. *Geophys. Res. Letts.* 39, Art. No. L18701, doi: 10.1029/2012GL052735, Sept 19.

7. Figure Captions

Figure 1: Nino34 (area averaged SSTA 170W-120W, 5S-5N) plumes – for 0.5 months lead: 1982-1995 on top and 1996-2010 on bottom.

Figure 2: Same as Fig. 1 but for 6.5 month lead

Figure 3: SSTA correlation coefficient – each ensemble member weighted equally. Retrospective forecasts are initialized in August 1982-2009 and verifying in following February (i.e., 5.5 month lead).

Figure 4: SSTA Root Mean Squared Error (RMSE) 20S-20N for each individual model compared to the multi-model mean, September starts 1982-2009, leads 0.5-5.5. The x-axis ranges from 0 to 2^{o}C and corresponds to the NMME RSME and the y-axis ranges from 0 to 2^{o}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.

Figure 5: SSTA Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (top panel) and for CFSv2 (bottom panel). The skill is based on hindcasts initialized in July 1982-2009 and verifying the following DJF seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast. Global averaged RPSS is noted on the figure.

651 Figure 6: SSTA Rank Probability Skill Scores (RPSS) for the grand NMME multi-model
652 ensemble (top panel) and for CFSv2 (bottom panel). The skill is based on hindcasts initialized in
653 January 1982-2009 and verifying the following JJA seasonal mean for tercile forecasts. Positive
654 values indicate probabilistic skill that is better then climatology and negative values indicate
655 probabilistic skill that is worse than a climatological forecast. Global averaged RPSS is noted on
656 the figure.
657
658 Figure 7: Surface atmospheric temperature (2 meter) Rank Probability Skill Scores (RPSS) for
659 the grand NMME multi-model ensemble (top panel) and for CFSv2 (bottom panel). The skill is
660 based on hindcasts initialized in January1982-2009 and verifying the following JJA seasonal
661 mean for tercile forecasts. Positive values indicate probabilistic skill that is better then
662 climatology and negative values indicate probabilistic skill that is worse than a climatological
663 forecast. Global averaged RPSS is noted on the figure.
664
665 Figure 8: Precipitation forecast Rank Probability Skill Scores (RPSS) for the grand NMME
666 multi-model ensemble (left panel) and for CFSv2 (right panel). The skill is based on hindcasts
667 initialized in July 1982-2009 and verifying the following DJF seasonal mean for tercile forecasts.
668 Positive values indicate probabilistic skill that is better then climatology and negative values
669 indicate probabilistic skill that is worse than a climatological forecast.
670
671 Figure 9: Precipitation forecast Rank Probability Skill Scores (RPSS) for the grand NMME
672 multi-model ensemble (left panel) and for CFSv2 (right panel). The skill is based on hindcasts
673 initialized in January 1982-2009 and verifying the following JJA seasonal mean for tercile

forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast.

Figure 10: NMME reliability diagram for 2-meter temperature anomalies throughout the globe. The reliability corresponds to forecasts initialized in October 1982-2009 and verifying in following JFM season.

Figure 11: Reliability diagram for 2-meter temperature anomalies throughout the globe from a sample of individual models. The reliability corresponds to forecasts initialized in October 1982-2009 and verifying in following JFM season.

Figure 12: Real-time NINO3.4 predictions initialized in May 2013.

Figure 13: Percent difference in RPSS skill of streamflow forecasts over the Colorado NIDIS testbed (left panel) and SE US NIDIS testbed (right panel) with lead times out to 6 months. Skill differences above 0 indicates NMME forecasts are more skilful than ESP. "Full resolution" indicates using the downscaled $1/8^{th}$ degree, daily seasonal climate model variables; "Avg Time" indicates the forecasts are averaged over the lead time; and "Avg Time and Space" indicates that the forecast are averaged over the lead times and domain.

Figure 14: NMME SPI6 forecasts initialized June 1, 2011 and Jun1, 2012. Observed MAM precipitation is combined with JJA model ensemble mean forecast. The NMME forecast is the equally weighted ensemble model average.

698 Figure 15: SSTA correlation coefficient for forecasts initialized in early January and verifying
699 for May (1982-2000). The top panel shows results using CCSM4 and CFSR initial states for the
700 ocean and the bottom panel shows results for CCSM3 using MOM3 ODA initial states.
701
702

703 Table 1: NMME partner models and forecasts

Model	Hindcast Period	Ensemble Size	Lead Times	Arrangement of Ensemble Members	Contact and reference
CFSv1	1981-2009	15	0.5-8.5 Months	1st 0Z +/-2 days, 21st0Z +/-2d, 11th 0Z+/- 2d	Saha (Saha et al. 2006)
CFSv2	1982-2010	24(28)	0.5-9.5 Months	4 members (0,6,12,18Z) every 5th day	Saha (Saha et al. 2013)
GFDL-CM2.2	1982-2010	10	0.5-11.5 Months	All 1st of the month 0Z	Rosati (Zhang et al. 2007)
IRI-ECHAM4-f[10]	1982-2010	12	0.5-7.5 Months	All 1st of the month 0Z	DeWitt (DeWitt 2005)
IRI-ECHAM4-a[2]	1982-2010	12	0.5-7.5 Months	All 1st of the Month 0Z	DeWitt (Dewitt 2005)
CCSM3.0	1982-2010	6	0.5-11.5 Months	All 1st of the Month 0Z	Kirtman (Kirtman and Min 2009)
GEOS5	1981-2010	11[11]	0.5-9.5 Months	1 Member every 5th day	Schubert (Vernieres et al. 2011)
CMC1-CanCM3	1981-2010	10	0.5-11.5	All 1st of the month 0Z	Merryfield Merryfield et al. (2013)
CMC2-CanCM4	1981-2010	10	0.5-11.5	ALL 1ST of the month 0Z	Merryfield Merryfield et al. (2013)

704

[10] Real-time forecasts terminated in July 2012.

[11] The number of forecast and hindcast ensemble members is not constant during the period. It has grown from 6 for the initial August of 2011 forecasts (and associated hindcasts), to 11 starting with our June 2012 forecasts. The additional (beyond 6 initialized every 5th day) ensemble members are based on breeding and other perturbations applied on the day closest to the beginning of the month.

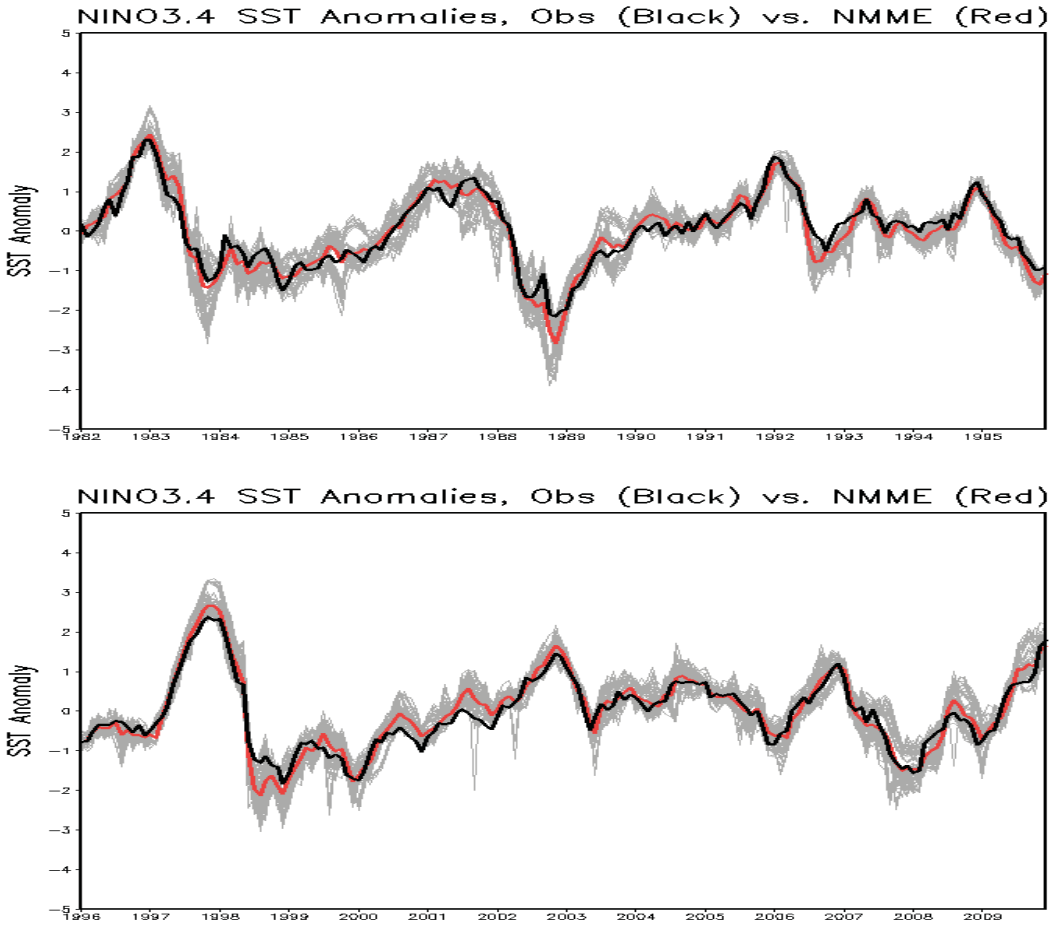

Figure 1: Nino34 (area averaged SSTA 170°W-120°W, 5°S- 5°N) plumes – for 0.5 month lead: 1982-1995 on top and 1996-2010 on bottom.

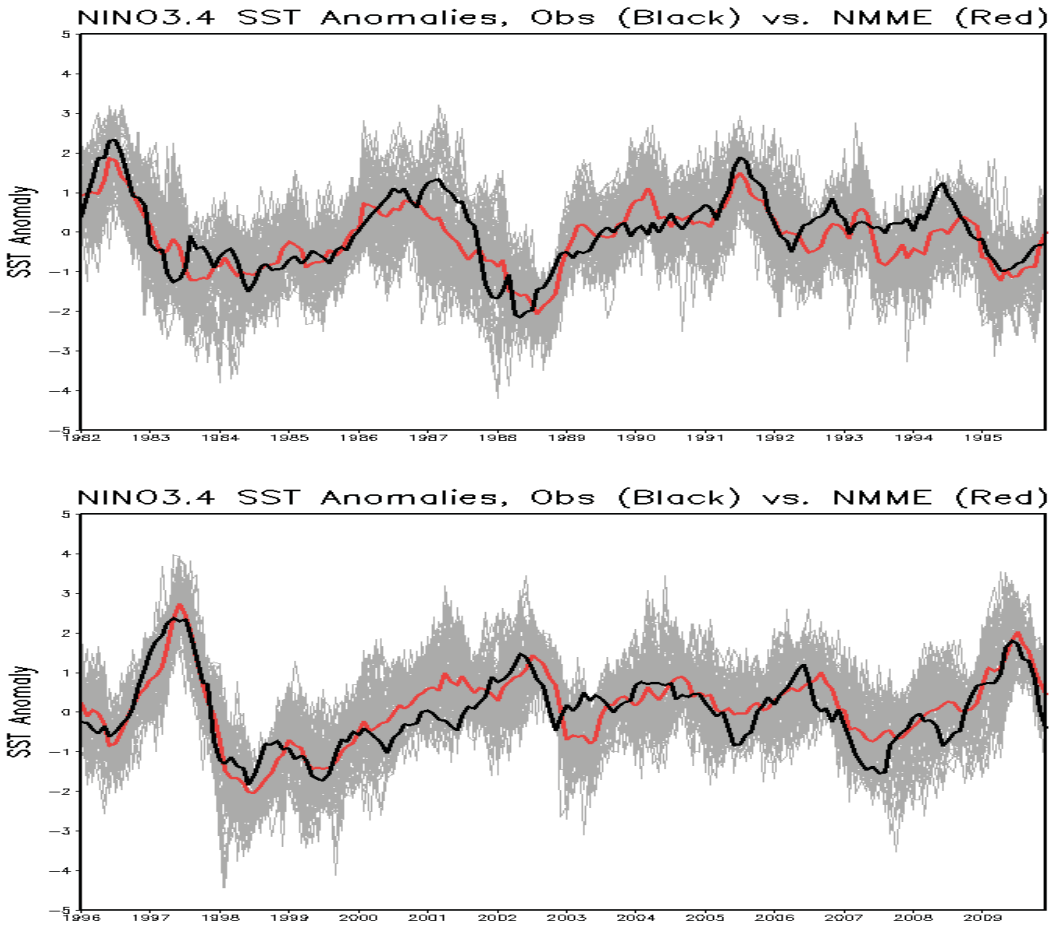

Figure 2: Same as Fig. 1 but for 6.5 month lead.

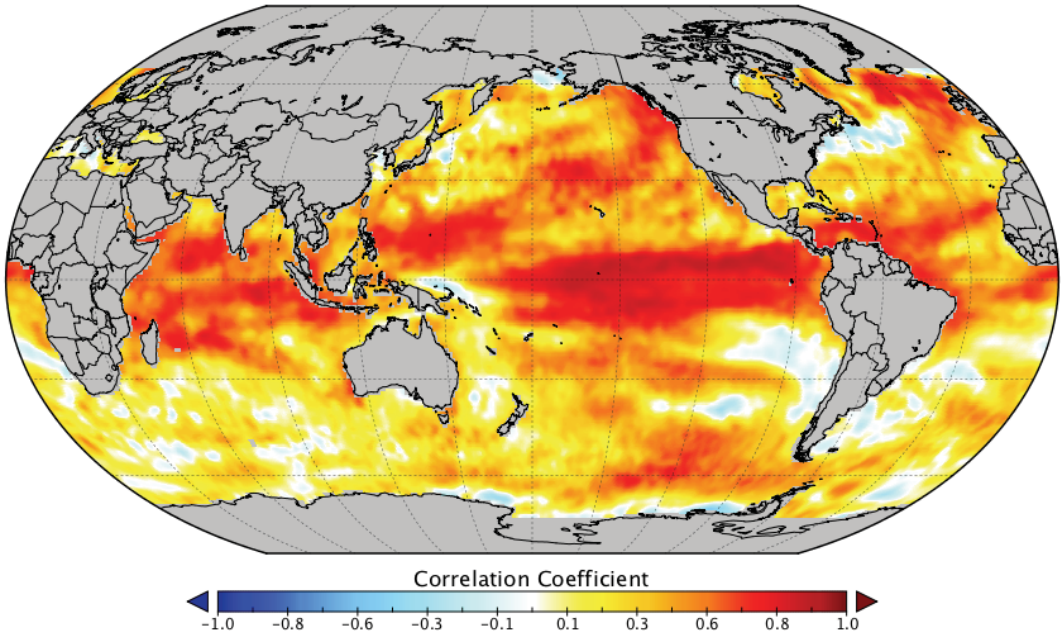

Fig. 3: SSTA correlation coefficient – each ensemble member weighted equally. Retrospective forecasts are initialized in August 1981-2010 and verifying in following February (i.e., 5.5 month lead).

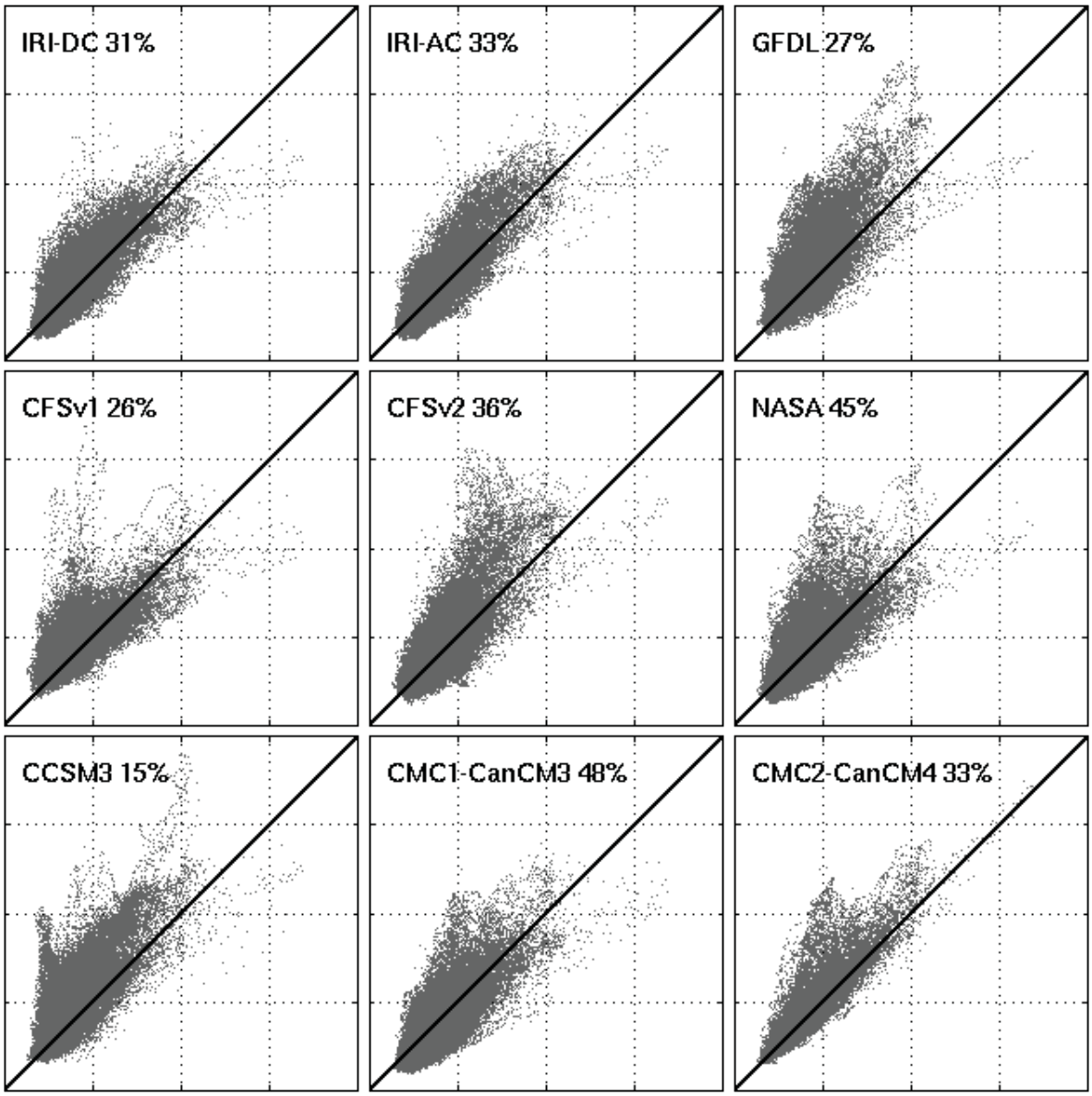

Fig. 4: SSTA Root Mean Squared Error (RMSE) 20S-20N for each individual model compared to the multi-model mean, September starts 1982-2009, leads 0.5-5.5. The x-axis ranges from 0 to 2°C and corresponds to the NMME RSME and the y-axis ranges from 0 to 2°C and corresponds to the individual model RMSE. Dots above the diagonal imply NMME has smaller RMSE. The percentage of points below the diagonal is noted in each panel.

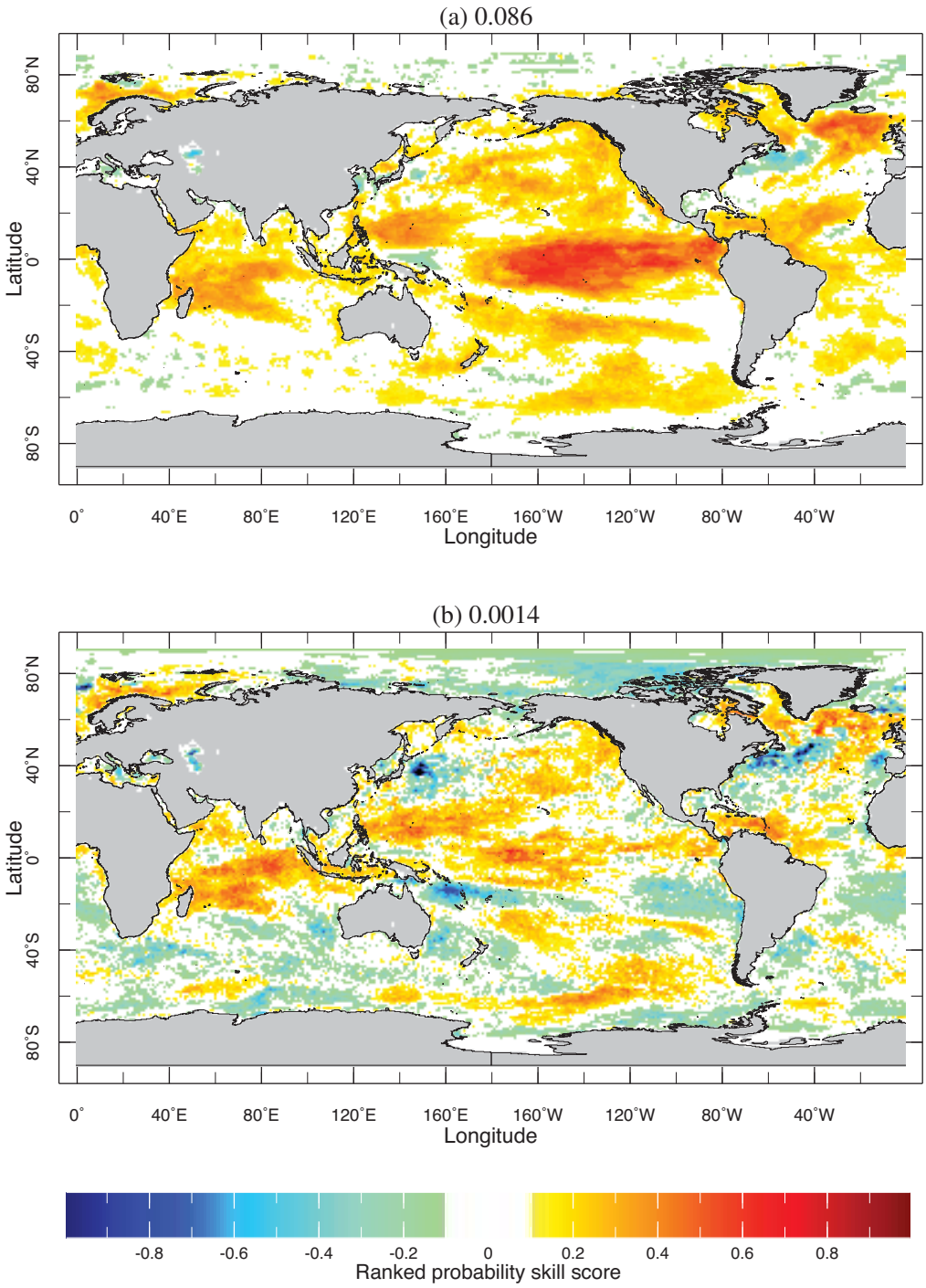

Figure 5: SSTA Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (top panel) and for CFSv2 (bottom panel). The skill is based on hindcasts initialized in July 1982-2009 and verifying the following DJF seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast. Global averaged RPSS is noted on the figure.

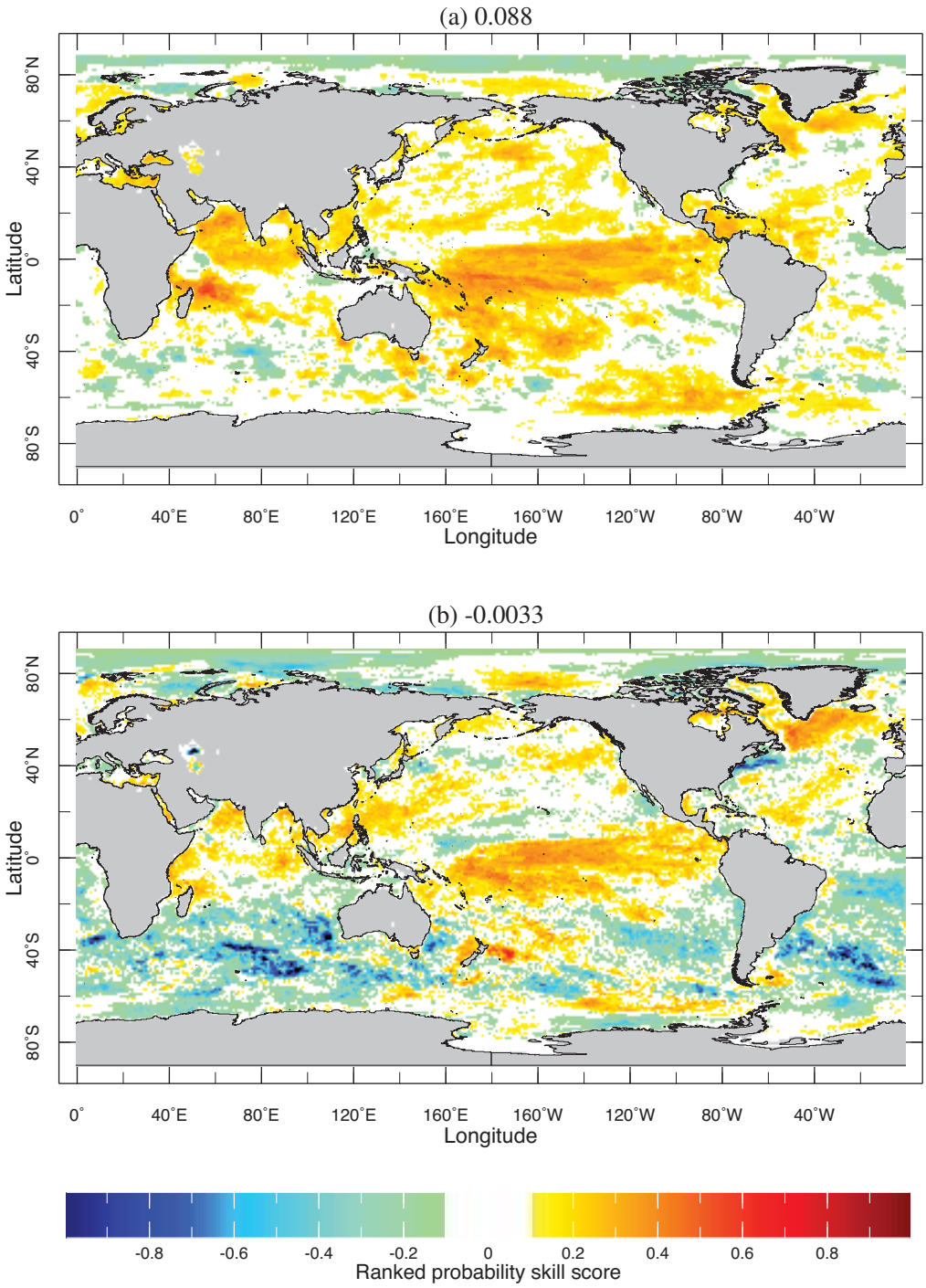

Figure 6: SSTA Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (top panel) and for CFSv2 (bottom panel). The skill is based on hindcasts initialized in January 1982-2009 and verifying the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast. Global averaged RPSS is noted on the figure..

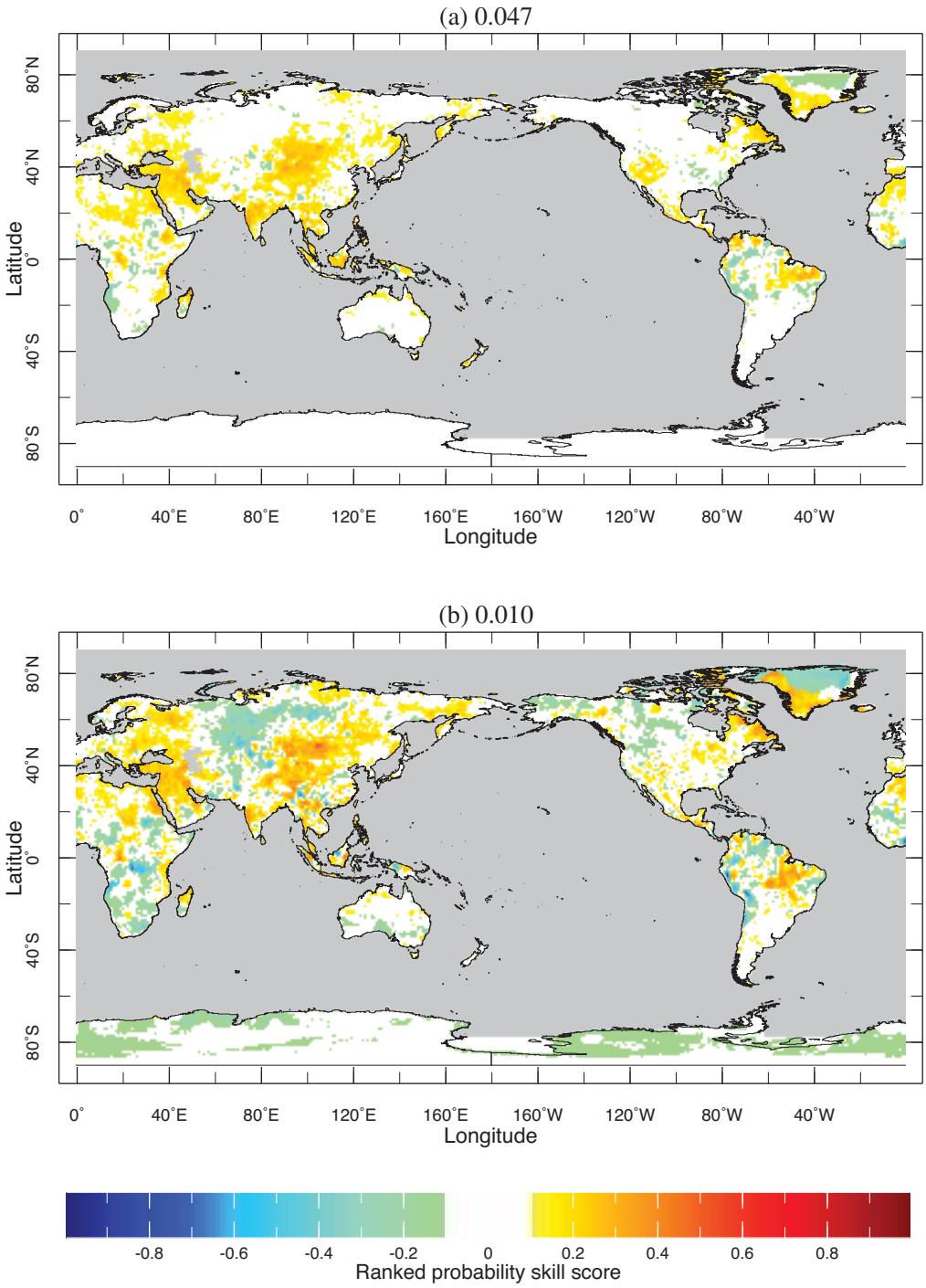

Figure 7: Surface atmospheric temperature (2 meter) Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (top panel) and for CFSv2 (bottom panel). The skill is based on hindcasts initialized in January1982-2009 and verifying the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast. Global averaged RPSS is noted on the figure.

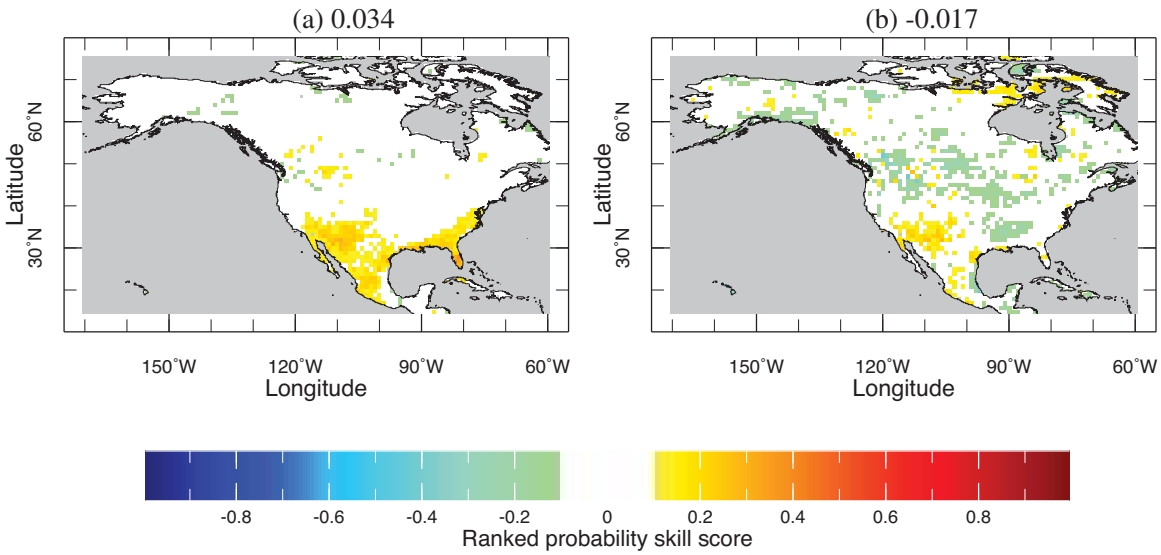

Figure 8: Precipitation forecast Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (left panel) and for CFSv2 (right panel). The skill is based on hindcasts initialized in July 1982-2010 and verifying the following DJF seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast.

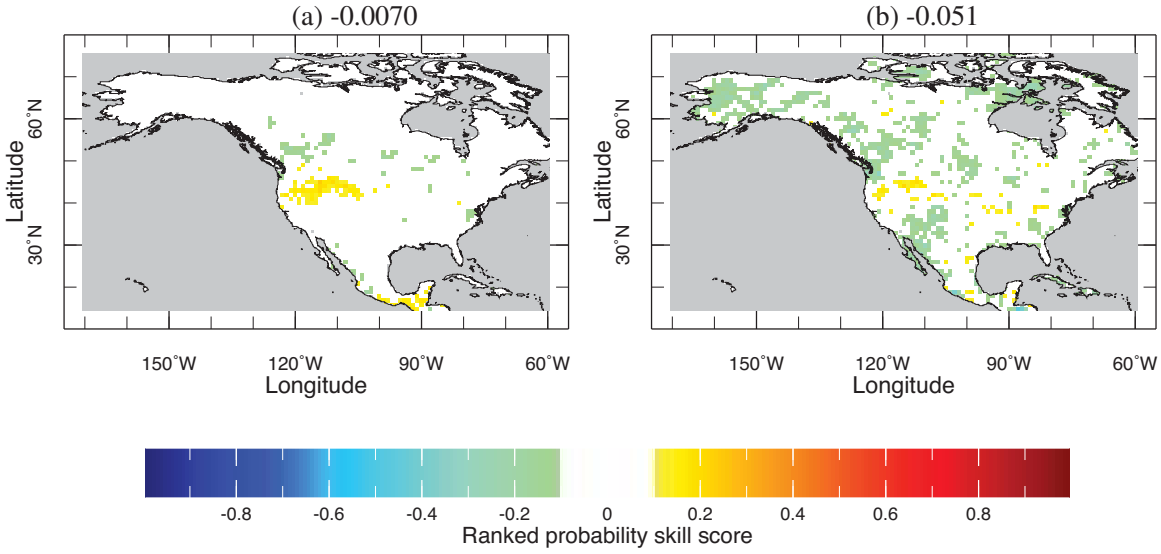

Figure 9: Precipitation forecast Rank Probability Skill Scores (RPSS) for the grand NMME multi-model ensemble (left panel) and for CFSv2 (right panel). The skill is based on hindcasts initialized in January 1982-2010 and verifying the following JJA seasonal mean for tercile forecasts. Positive values indicate probabilistic skill that is better then climatology and negative values indicate probabilistic skill that is worse than a climatological forecast.

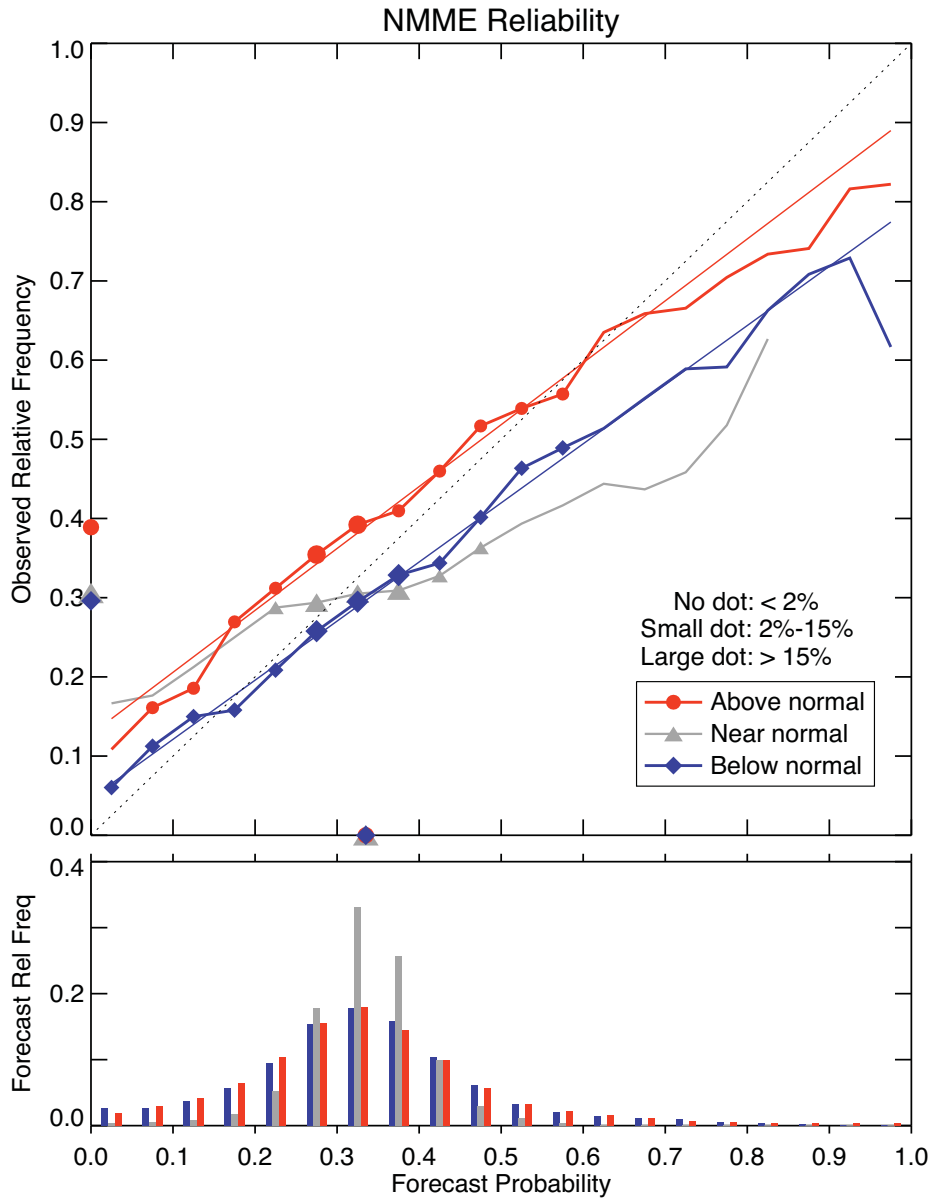

Figure 10: NMME reliability diagram for 2-meter temperature anomalies throughout the globe. The reliability corresponds to forecasts initialized in October 1982-2009 and verifying in following JFM season.

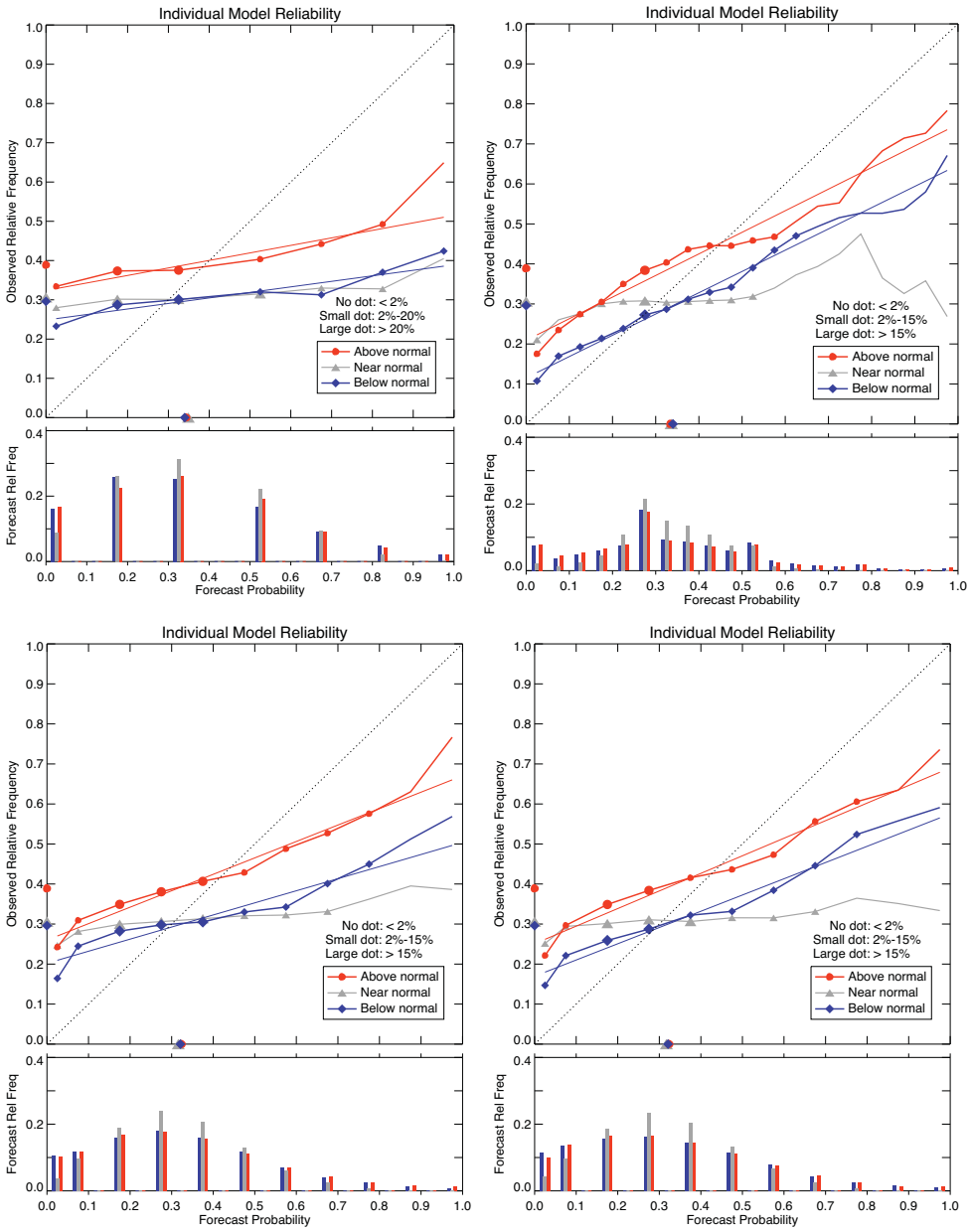

Figure 11: Reliability diagram for 2-meter temperature anomalies throughout the globe from a sample of individual models. The reliability corresponds to forecasts initialized in October 1982-2009 and verifying in following JFM season.

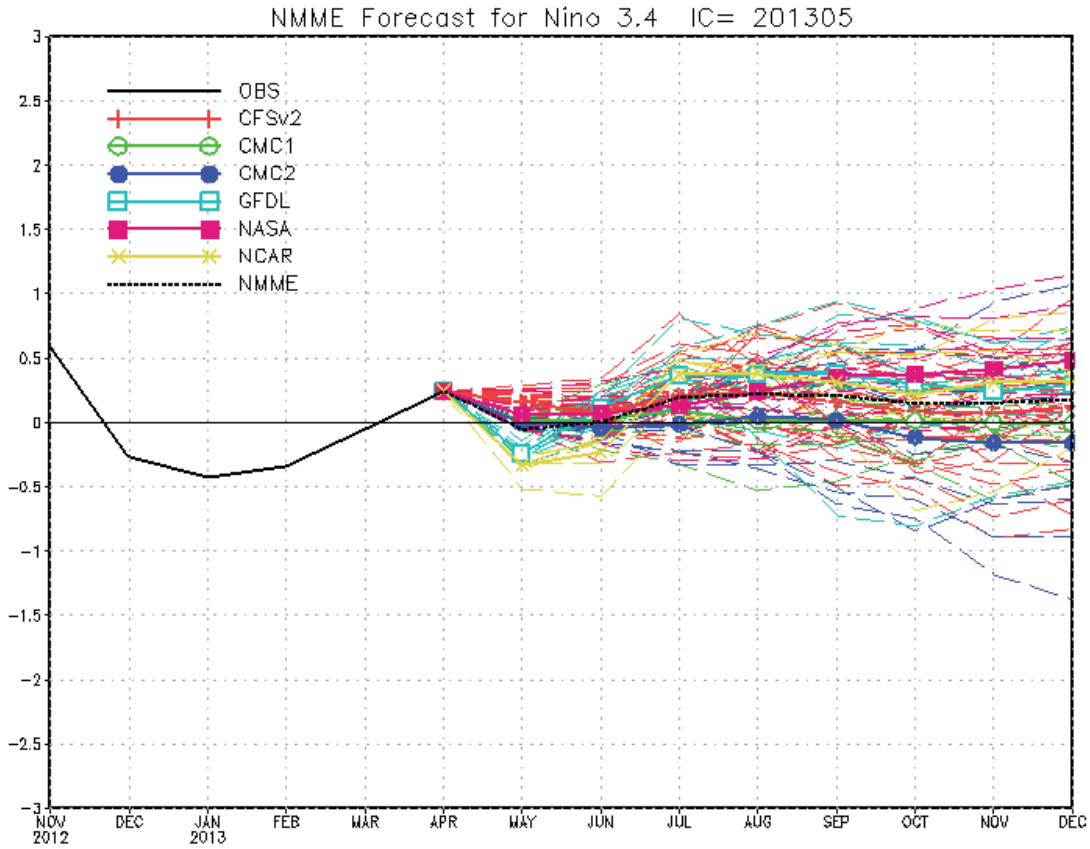

Figure 12: Real-time NINO3.4 predictions initialized in May 2013.

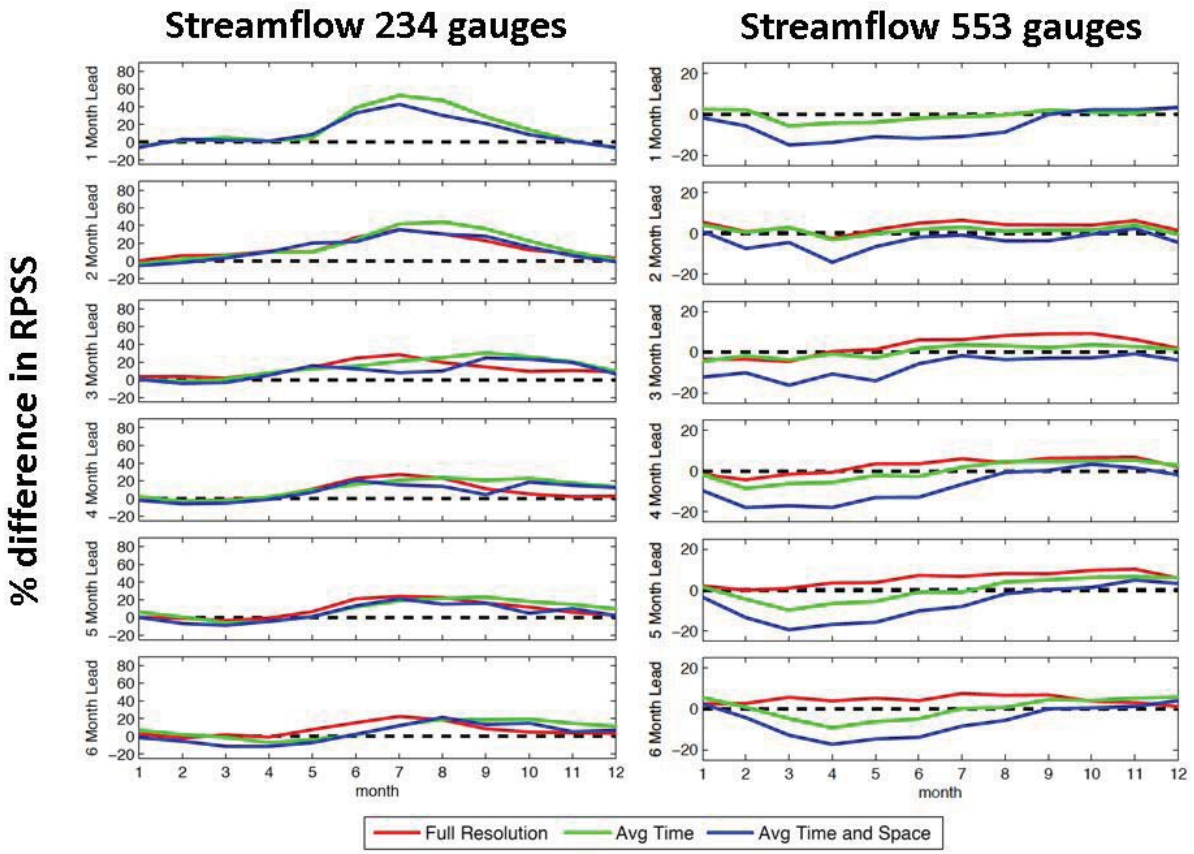

Figure 13: Percent difference in RPSS skill of streamflow forecasts over the Colorado NIDIS testbed (left panel) and SE US NIDIS testbed (right panel) with lead times out to 6 months. Skill differences above 0 indicates NMME forecasts are more skilful than ESP. "Full resolution" indicates using the downscaled 1/8th degree, daily seasonal climate model variables; "Avg Time" indicates the forecasts are averaged over the lead time; and "Avg Time and Space" indicates that the forecast are averaged over the lead times and domain.

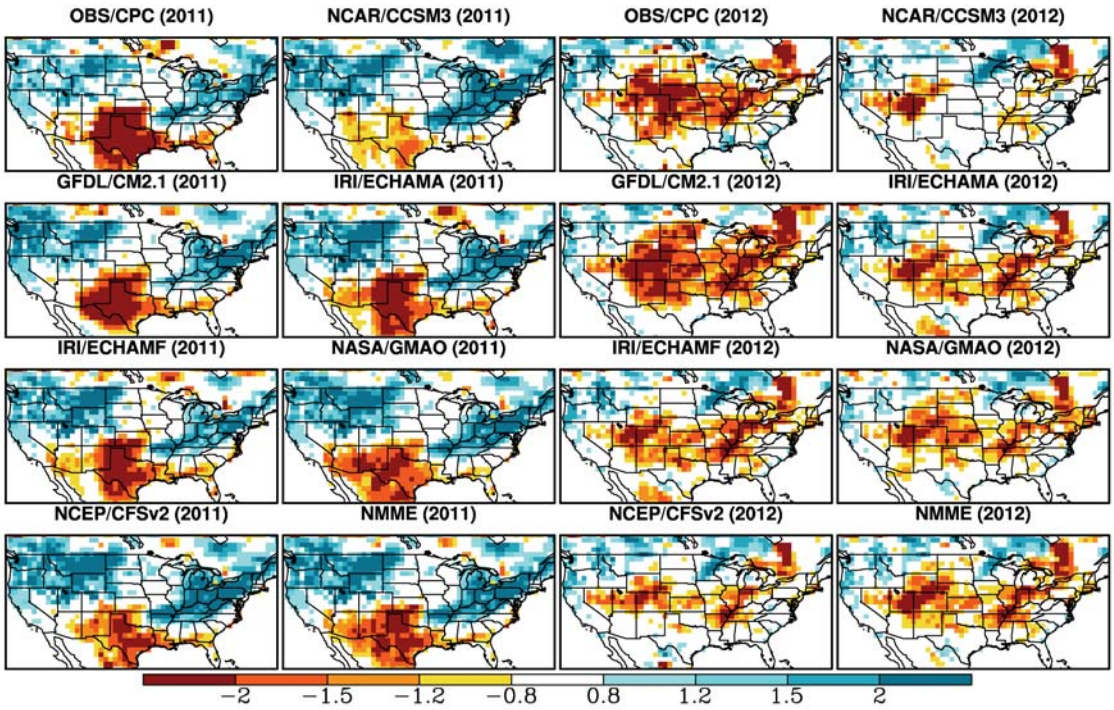

Figure 14: NMME SPI6 forecasts initialized June 1, 2011 and Jun1, 2012. Observed MAM precipitation is combined with JJA model ensemble mean forecast. The NMME forecast is the equally weighted ensemble model average.

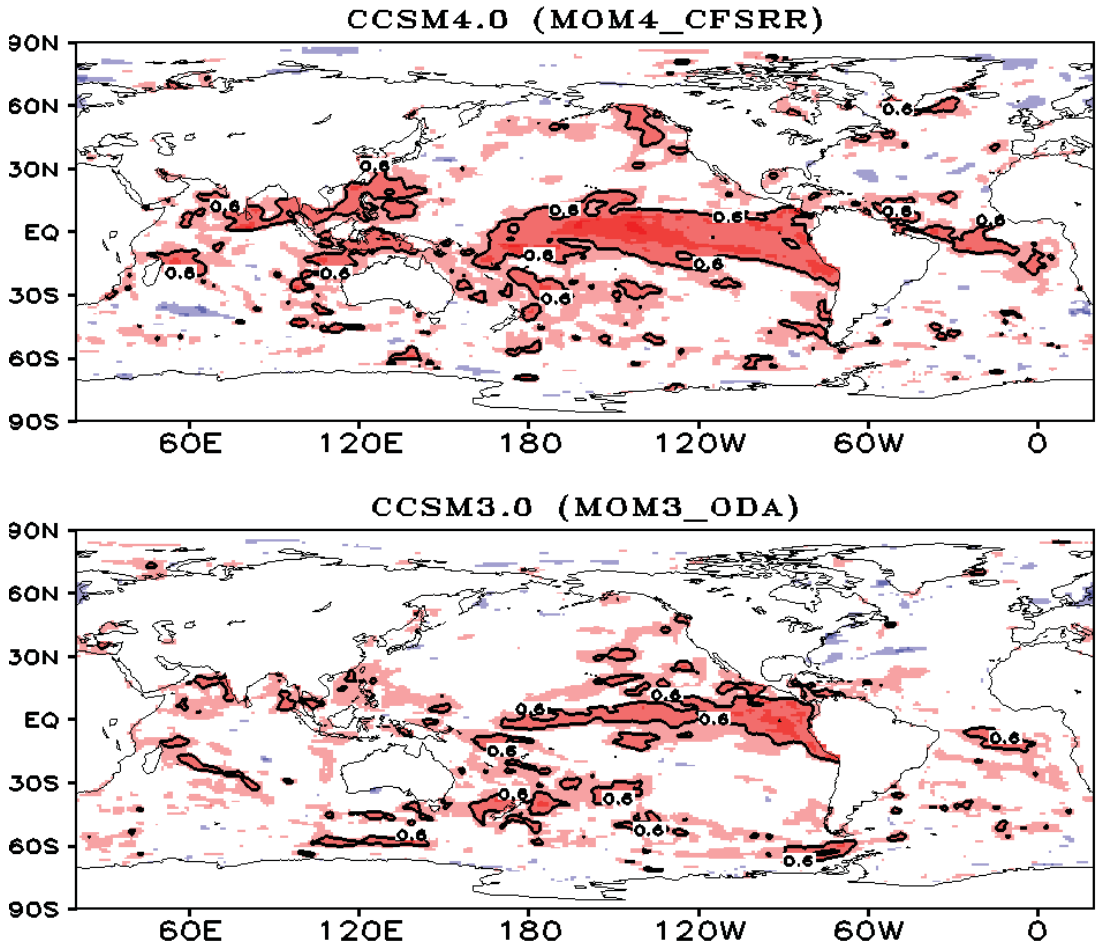

Figure 15: SSTA correlation coefficient for forecasts initialized in early January and verifying for May (1982-2000). The top panel shows results using CCSM4 and CFSR initial states for the ocean and the bottom panel shows results for CCSM3 using MOM3 ODA initial states.