Probabilistic Seasonal Forecasts in the North American Multimodel Ensemble: A Baseline Skill Assessment

EMILY BECKER AND HUUG VAN DEN DOOL

Climate Prediction Center, NCEP/NWS/NOAA, College Park, Maryland

(Manuscript received 23 December 2014, in final form 24 November 2015)

ABSTRACT

The North American Multimodel Ensemble (NMME) forecasting system has been continuously producing seasonal forecasts since August 2011. The NMME, with its suite of diverse models, provides a valuable opportunity for characterizing forecast confidence using probabilistic forecasts. The current experimental probabilistic forecast product (in map format) presents the most likely tercile for the seasonal mean value, chosen out of above normal, near normal, or below normal categories, using a nonparametric counting method to determine the probability of each class. The skill of the 3-month-mean probabilistic forecasts of 2-m surface temperature (T2m), precipitation rate, and sea surface temperature is assessed using forecasts from the 29-yr (1982–2010) NMME hindcast database. Three forecast configurations are considered: a full six-model NMME; a ''mini-NMME'' with 24 members, four each from six models; and the 24-member CFSv2 alone. Skill is assessed on the cross-validated hindcasts using the Brier skill score (BSS); forecast reliability and resolution are also assessed. This study provides a baseline skill assessment of the current method of creating probabilistic forecasts from the NMME system.

For forecasts in the above- and below-normal terciles for all variables and geographical regions examined in this study, BSS for NMME forecasts is higher than BSS for CFSv2 forecasts. Niño-3.4 forecasts from the full NMME and the mini-NMME receive nearly identical BSS that are higher than BSS for CFSv2 forecasts. Even systems with modest BSS, such as T2m in the Northern Hemisphere, have generally high reliability, as shown in reliability diagrams.

1. Introduction

Most official short-term climate forecasts from the National Oceanic and Atmospheric Administration (NOAA) and other forecasting centers are now issued in a probabilistic format, providing the end user with quantitative information about forecast uncertainty and allowing for more informed risk assessment and decision making (Tebaldi and Knutti 2007). A considerable advantage of multimodel ensembles (MMEs) over a single-model approach is thought to be the ability to sample forecast uncertainty due to model diversity (Palmer et al. 2004; Hagedorn et al. 2005; Doblas-Reyes et al. 2005; Kirtman and Min 2009). Performance improvements from one MME system to another have been noted as model resolution has increased and upgrades have been made to physics, coupling mechanisms, and data assimilation; an example of this is the progression from DEMETER to ENSEMBLES (Weisheimer et al. 2009; Alessandri et al. 2011).

The North American Multimodel Ensemble (NMME) has been producing real-time monthly-mean and seasonal anomaly forecasts regularly since August 2011 (Kirtman et al. 2014). The real-time forecasts, which are published by NOAA's Climate Prediction Center (CPC) by the ninth day of every month, include deterministic forecasts (forecasts for a specific anomaly) for each participating model and the MME, and probabilistic forecasts based on the entire ensemble of all models, all members. The NMME, with its large number of contributing models and ensemble members, presents a valuable opportunity for characterizing uncertainty due to both model diversity and uncertainty in initial conditions, and the probabilistic forecasts were added to the suite of real-time NMME forecast products in November 2012.

This study assesses the skill of the current version of NMME probabilistic forecasts. These forecasts are

Corresponding author address: Emily Becker, NOAA Center for Weather and Climate Prediction, 5830 University Research Court, College Park, MD 20740.
E-mail: emily.becker@noaa.gov

DOI: 10.1175/JCLI-D-14-00862.1

currently in use by operational forecasters and other users of climate outlooks, and have been targeted for further development and calibration. As such, an assessment of the current baseline skill is in order. As the NMME has both a large total number of ensemble members and the advantage that comes from the combination of several models with different physics, data assimilation, and initializations, we present statistical analysis of the skill of forecasts for three configurations: all ensemble members from the entire six-model NMME, a 24-member "mini-NMME" comprising four members from each of the six models, and the 24-member Climate Forecast System version 2 (CFSv2).

CFSv2 is the NMME participant model with the largest number of ensemble members, allowing a better chance for resolving forecast probabilities, and probabilistic forecasts from this model were already in use by forecasters at the CPC before NMME emerged. The mini-NMME is included to assess the skill of a multimodel system with an equal number of ensemble members per model; a combination with four members from each of the six models provides an ensemble with the same number of members as the CFSv2. The comparison of results from these three systems (NMME, mini-NMME, and CFSv2) also allows us to draw some preliminary conclusions about the relative advantages of increased ensemble size and model diversity. As the number of real-time forecasts is small as yet (November 2012 to the present), this paper focuses on the retrospective forecasts (also known as hindcasts) from 1982 to 2010.

2. Data and methods

a. The North American Multimodel Ensemble

The NMME is a forecasting system consisting of coupled models from U.S. and Canadian modeling centers (Kirtman et al. 2014). The NMME has been producing global monthly-mean and seasonal forecasts since August 2011 for 2-m surface temperature, sea surface temperature (SST), and precipitation rate; real-time and archived forecast graphics from August 2011 to the present are available online at www.cpc.ncep.noaa.gov/products/NMME/. Real-time probabilistic forecasting with the NMME started in November 2012.

To date, at various times, six centers have contributed 11 models to the NMME (see www.cpc.ncep.noaa.gov/products/NMME/Phase1models.png for details of all the models). All participating models are required to produce monthly-mean forecast data on a 1.0° longitude × 1.0° latitude grid (for a grid size of 360 × 181), with leads up to at least 7 months. Initialization, data assimilation, and model physics are left up to the modeling centers. Forecasts are produced monthly, for a total of 12 initial condition months. Retrospective runs of the forecast version of each model from 1982–2010 allow for model calibration, bias removal, and skill assessment.

As several of these models are recent additions, this study uses an NMME consisting of six models that have been in regular use in all of 2014–15 to approximate the skill of the current real-time forecasting system. All models have year-round hindcasts available over the period 1982–2010. These are the NCEP-CFSv2 (Saha et al. 2014), CMC-CanCM3 and CanCM4 (Merryfield et al. 2013), GFDL-CM2.1 (Zhang et al. 2007), NCAR-CCSM4 (B. Kirtman et al. 2015, unpublished manuscript), and NASA-GEOS5 (Vernieres et al. 2012). See Table 1 for more detail about the six models used in this study. The hindcasts from several older and newer models, not used for this study, are still available for research.

b. Verification fields

The observation verification field for 2-m temperature (T2m) is the station observation-based monthly mean surface air temperature dataset GHCN+CAMS, a combination of the Global Historical Climatology Network (GHCN) and the Climate Anomaly Monitoring System (CAMS), two station networks. The station reports are interpolated to a grid with native resolution of 0.5° latitude × 0.5° longitude (Fan and van den Dool 2008). The data were regridded to the 1.0° × 1.0° grid for NMME purposes. As the 7-month lead forecasts initialized in 2010 stretch into 2011, the observation period used in this study for all verification fields runs from January 1982 to July 2011.

Precipitation forecasts are verified using the CPC Merged Analysis of Precipitation (CMAP). This dataset merges rain gauge observations with precipitation estimates from several satellite-based algorithms (Xie and Arkin 1997). CMAP, which is produced on a 2.5° × 2.5° latitude/longitude grid, is rescaled via bilinear interpolation to 1.0° × 1.0° for this study.

The SST prediction was verified using the optimum interpolation version 2 (OI) analysis of Reynolds et al. (2002). This analysis, produced at NOAA, uses both satellite data and in situ records from ships and buoys. The native resolution of the Reynolds et al. (2002) SST is 1° latitude × 1° longitude.

c. Probabilistic forecast construction in real-time NMME

In a general sense, probabilistic forecasts from ensemble models are constructed by determining how

TABLE 1. North American Multimodel Ensemble (NMME) models used in this study. (Expansions of acronyms are available online at http://www.ametsoc.org/PubsAcronymList.)

Center/model	Hindcast period	No. of members	Arrangement of members	Lead (month)	Model resolution atmos.	Model resolution ocean	Reference
NCEP/CFSv2	1982–2010	24 (28 Nov)	4 members (0000, 0600, 1200, 1800 UTC)	0–9	T126L64	MOM4L40 0.25° equator (Eq)	Saha et al. (2014)
GFDL/CM2.1	1982–2010	10	All 1st of the month 0000 UTC	0–11	2 × 2.5° L24	MOM4L50 0.3° Eq	Zhang et al. (2007)
Environment Canada/CMC1-CanCM3	1981–2010	10	All 1st of the month 0000 UTC	0–11	CanAM3 T63L31	CanOM4L40 0.94° Eq	Merryfield et al. (2013)
Environment Canada/CMC1-CanCM4	1981–2010	10	All 1st of the month 0000 UTC	0–11	CanAM4 T63L35	CanOM4L40 0.94° Eq	Merryfield et al. (2013)
NCAR/CCSM4	1982–2010	10	All 1st of the month 0000 UTC	0–11	0.9 × 1.25° L26	POPL60 0.25° Eq	B. Kirtman et al. (2015, unpublished manuscript)
NASA/GEOS5	1981–2010	11	4 members every 5th days; 7 members on the last day of last month	0–9	1 × 1.25° L72	MOM4L40 0.25° Eq	Vernieres et al. (2012)

many ensemble members fall into the same predetermined category, indicating the model confidence in a particular outcome. A widely used format is to present probabilities for each of three equal tercile categories: above-normal, near-normal, and below-normal. NMME probabilistic forecasts take this format, as have CPC's operational monthly-mean and seasonal outlooks for many years.

The current NMME real-time probabilistic forecasts begin with the calculation of forecast anomalies. Each month, the hindcast ensemble mean climatology of each model is found using all members and all years (1982–2010) of the hindcasts. This climatology is then subtracted from each forecast ensemble member to create forecast anomalies. The use of each model's hindcast climatology when computing anomalies corrects for systematic bias in the mean (i.e., the systematic difference between the climatology of the model and the observations).

The next step is the determination of tercile thresholds, that is, the values that delimit the highest one-third of the forecast anomalies, the middle one-third of the forecast anomalies, and the lowest one-third. For each model and forecast field, a normal (Gaussian) distribution is assumed to fit the data (please see note in next paragraph about the Gaussian assumption). For each model individually, a normal (Gaussian) distribution is fit grid-point-wise to the 1982–2010 hindcasts (all members, fixed lead, all years for each initial month.) The standard deviation (std. dev.) of this Gaussian distribution is then found and used to define the tercile thresholds. With a normal distribution, forecast anomalies above +0.43 std. dev. are considered above normal (A), below −0.43 std. dev. are below normal (B), and any anomalies falling between −0.43 and +0.43 std. dev. are near normal (N). Systematic bias in the mean is removed when the model's own climatology is used to create the forecast anomalies. Systematic bias in the distribution (a model may have too much or too little spread compared to the observed field) is accounted for as the standard deviation fit to the model data is used to determine the tercile limits.

In the case of precipitation, a power transform of 1/4 is applied to the original hindcast and forecast values to bring the precipitation distribution closer to normal. Some testing of this method has been done by the authors, and it is found to be acceptably effective at correcting the skewed distribution. However, it is very likely that this technique, and the assumption of a Gaussian distribution for some other variables, can be improved upon, and different distributions and methods of determining appropriate tercile thresholds for a nonnormal

distribution will be tested in the course of the development of forecast calibrations. Areas where 0 mm of seasonal precipitation is recorded for more than a third of the years in the CMAP observed precipitation historical record (10 or more years in the 29-yr record used in this study; e.g., the eastern Sahara) have been masked out.

Probabilistic forecasts are then formed using the forecast anomalies by counting the number of ensemble members in each model that fall above, between, and below the tercile thresholds from its own hindcast. The numbers of ensemble members in each category for all the models are then added together, and divided by the total number of ensemble members in that month's NMME. The real-time NMME forecasts include over 110 members. Other methods of probabilistic forecast formation, including a parametric probability estimator applied to the forecast (in the current study, a parametric fit is only applied to the hindcast, to identify the tercile boundaries), have been found to result in higher skill scores for multimodel ensembles (Kharin et al. 2009). A thorough comparison of these and other techniques in the context of the NMME is left for a later study. The present study is a benchmark.

For this skill assessment, the above method is applied to the hindcasts, with the following modifications: for each of the 29 yr, one year is held out as the forecast to be verified, and the model climatology and tercile thresholds are calculated using the other 28 yr (cross validation). Skill is evaluated for each forecast year, and aggregated over the 29 yr (left out in turn). All results shown below are cross-validated this way.

The hindcast skills of three different systems are considered here. The "full NMME" comprises six models: NCEP's Climate Forecast System version 2 (CFSv2; Saha et al. 2014), CMC-CanCM3 and CanCM4 (Merryfield et al. 2013), GFDL-CM2.1 (Zhang et al. 2007), NCAR-CCSM4 (B. Kirtman et al. 2015, unpublished manuscript), and NASA-GEOS5 (Vernieres et al. 2012) (Table 1). CFSv2 has 24 members for each initial month (except for November, which has 28 members); the remaining models have 10 members, except for GEOS5, which has 11, for a total of 75 members for the full NMME. (For the purposes of this study, the last four members of the November initial conditions from CFSv2 are ignored.) The second combination considered in this study is a "mini-NMME" comprising four members from each of the six models, for a total of 24 members. The last system considered is forecasts from the 24-member CFSv2 alone. This model has the highest number of ensemble members of any of the NMME participant models, allowing for probabilistic forecasts with a relatively high resolution. The evaluation of the three different combinations provides some opportunity for comparing the relative skill of the multimodel system with that of a single model, and the effect of model diversity and ensemble size.

TABLE 2. Brier skill scores (BSS) for 3-month-mean seasons, averaged over all 12 initial conditions for each lead, for SST forecast–observation pairs area-aggregated in the Niño-3.4 region (5°S–5°N, 190°–240°E). Rows show BSS for forecasts in each of three terciles, above-normal (A), near-normal (N), and below-normal (B), for three forecast systems: the full, 6-model NMME with all members; a mini-NMME comprising 4 members from each of the 6 models for a total of 24 members; and the 24-member CFSv2 model. Six leads are shown; lead 0 is the 3-month-mean forecast that includes the initial month, e.g., January–March for January initial conditions, and so on. BSS for lead-1 forecasts are printed in bold for visibility, as the discussion focuses on these results. The upper and lower bounds of the 95% confidence interval, determined from 1000 bootstrap samples with replacement, are shown in brackets.

BSS: SST in Niño-3.4 region						
	Lead 0	**Lead 1**	Lead 2	Lead 3	Lead 4	Lead 5
		NMME				
A	0.68	**0.60 [0.603, 0.596]**	0.54	0.48	0.42	0.38
N	0.34	**0.23 [0.236, 0.228]**	0.18	0.14	0.12	0.09
B	0.65	**0.58 [0.580, 0.572]**	0.54	0.51	0.48	0.45
		mini-NMME				
A	0.68	**0.60 [0.607, 0.599]**	0.54	0.48	0.43	0.38
N	0.35	**0.24 [0.248, 0.239]**	0.18	0.15	0.13	0.10
B	0.66	**0.60 [0.599, 0.592]**	0.56	0.53	0.49	0.45
		CFSv2				
A	0.53	**0.45 [0.452, 0.441]**	0.38	0.32	0.27	0.24
N	0.10	**0.04 [0.043, 0.032]**	0.02	0.03	0.03	0.01
B	0.49	**0.43 [0.431, 0.420]**	0.39	0.37	0.36	0.34

In the probabilistic NMME forecast expression, the model forecasts are given equal weights in the consolidation: one member, one vote (probabilistic). Many previous studies have concluded that successful weighting schemes are difficult to find, due at least in part to the relatively short hindcast periods available (Hagedorn et al. 2005; Tebaldi and Knutti 2007) and the high colinearity of the forecasts (Peña and van den Dool 2008). Hence, so-called equal weights should be satisfactory. However, note that as each ensemble member is equally counted in the probabilistic forecast, models with more ensemble members contribute more to the probability forecast.

d. Skill measures

Skill of the retrospective probabilistic forecasts is assessed using the Brier score (BS), which provides a summary of the mean squared error of probability forecasts (Brier 1950; Wilks 2006). In the case of tercile forecasts, each category is assessed separately as a dichotomous event forecast, where the observation is either 1 (event occurred) or 0 (event did not occur), and

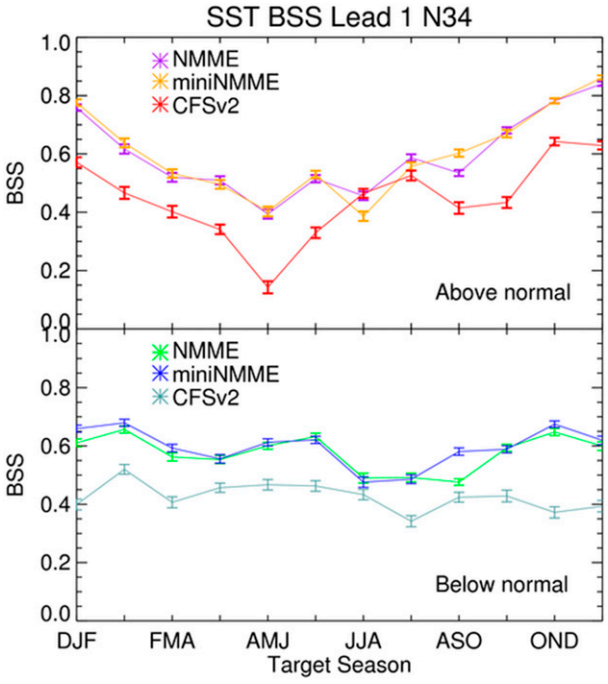

FIG. 1. Brier skill score (BSS) area-aggregated for all lead-1 seasonal probabilistic forecasts of SST in the Niño-3.4 region (5°S–5°N, 190°–240°E). Forecasts for the (top) above-normal and (bottom) below-normal terciles are shown for three forecasting systems: all members of the 6-model NMME (75 members), a mini-NMME comprising 4 members from each of the 6 models (24 members), and the CFSv2 alone (24 members). The lead-1 forecast for December–February (DJF) is made using November initial conditions, etc. Bars show the 95% confidence interval from 1000 bootstrap samples.

the forecast is a percent likelihood (e.g., 0.45). The Brier skill score (BSS) compares the Brier score of the model forecasts to the Brier score of a reference (climatological) forecast of 0.33 per tercile. A positive BSS indicates that the BS of the model forecast is lower than the BS of a climatological forecast.

The Brier score can be decomposed into three terms: reliability, resolution, and an uncertainty term [see Wilks (2006) for the algebraic decomposition], such that Brier score = reliability − resolution + uncertainty. In forecast verification, reliability represents the comparison of a forecast probability for an event to the observed frequency of that event. For example, for all forecasts of 60% probability of above normal T2m, above normal should be observed 60% of the time (for the forecast to be reliable). Resolution gives an indication of the use of different forecast probabilities, that is, the ability of the forecast system to assign probabilities different from the climatological probability (Wilks 2006). [The uncertainty term depends on the observed event frequency only, and so does not change in this evaluation among

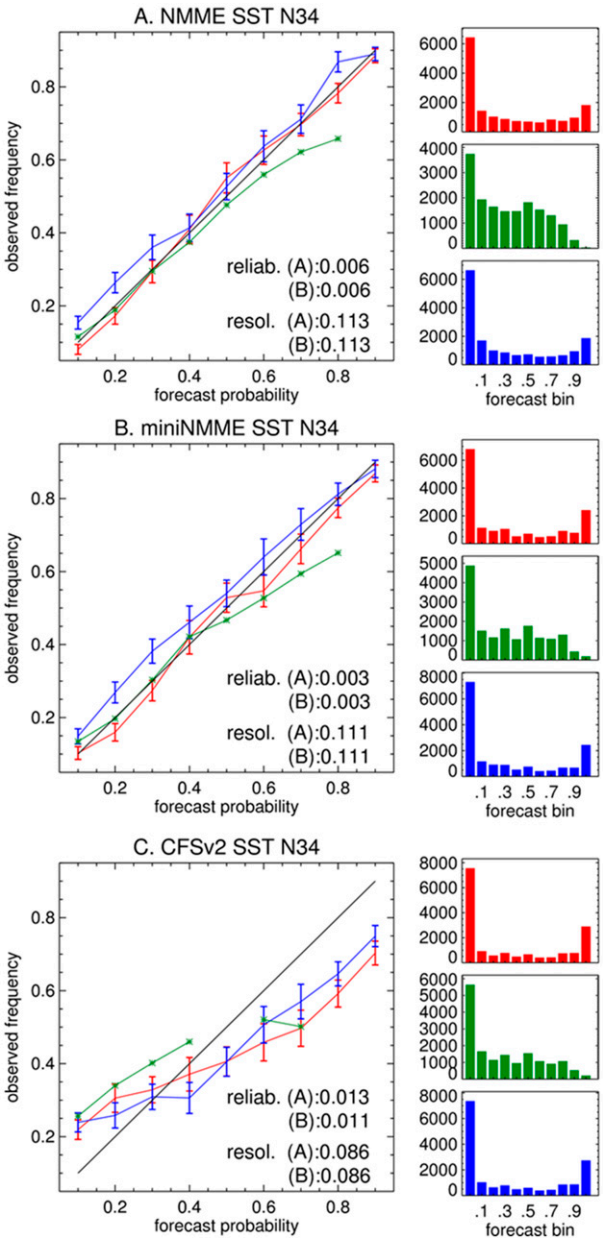

FIG. 2. Reliability diagrams for probabilistic forecasts of sea surface temperature (SST) in the Niño-3.4 region (5°S–5°N, 190°–240°E) for the (a) NMME, (b) mini-NMME, and (c) CFSv2, aggregated for the lead-1 season from all initial conditions. Red lines indicate forecasts in the above tercile, blue the below, and green the near-normal. Lines closer to the black diagonal mean the observed event frequency (y axis) is close to the forecast probability (x axis), and therefore the forecasts are more reliable. Histograms indicate how often each forecast bin is used; numbers on the y axis has been divided by 1000. All diagrams use 10 bins of size 0.1. Alphanumeric insets show the reliability and resolution terms of the Brier score for the above (A) and below (B) terciles. As Brier score = reliability − resolution + uncertainty, higher-resolution values and lower reliability values are desirable. Bars indicate the 95% confidence intervals derived from 1000 bootstrapping tests.

TABLE 3. As in Table 2, but for SST in the Northern Hemisphere (all ocean grid points 23°–75°N).

BSS: SST in Northern Hemisphere						
	Lead 0	**Lead 1**	Lead 2	Lead 3	Lead 4	Lead 5
			NMME			
A	0.30	**0.201 [0.202, 0.201]**	0.16	0.14	0.13	0.12
N	0.04	**−0.001 [0.0, −0.001]**	−0.01	−0.02	−0.02	−0.02
B	0.29	**0.187 [0.188, 0.186]**	0.15	0.13	0.11	0.10
			mini-NMME			
A	0.29	**0.182 [0.183, 0.181]**	0.14	0.11	0.10	0.09
N	0.03	**−0.02 [−0.017, −0.019]**	−0.03	−0.04	−0.04	−0.04
B	0.27	**0.161 [0.162, 0.160]**	0.12	0.10	0.09	0.08
			CFSv2			
A	0.25	**0.125 [0.126, 0.124]**	0.08	0.05	0.04	0.03
N	0.00	**−0.047 [-0.046, −0.048]**	−0.06	−0.06	−0.06	−0.06
B	0.25	**0.122 [0.123, 0.120]**	0.08	0.05	0.04	0.03

forecast strategies (CFSv2 only, NMME, etc.) and is not considered.]

The reliability and resolution can be illustrated by the standard reliability diagram and insets, which shows the full joint distribution of the forecasts and observations (Wilks 2006; Weisheimer and Palmer 2014). These diagrams allow for visual comparison of the conditional event frequency and the forecast probability; the associated ''sharpness diagrams'' indicate how often the forecast probabilities in each bin are issued (Wilks 2006; Jolliffe and Stephenson 2012). The ''event'' is an observation falling in a particular tercile. Probability forecasts are assigned to one of 10 bins (0–0.1, etc.). Reliability is indicated by the distance of the forecast line away from the diagonal, and resolution is assessed by the distance of the forecast line from the horizontal line along the observed climatological frequency (0.33, in the case of tercile categories; not shown.).

Confidence intervals for the statistics in this study have been determined using a bootstrapping approach (Wilks 2006; Doblas-Reyes et al. 2009). For each of the geographical regions defined below, the forecast–observation grid point pairs (which have already been subjected to cross validation) have been resampled with replacement 1000 times, and the BSS and reliability statistics have been computed for each of the resulting 1000 samples; it is assumed that the resulting distribution of scores represents the true underlying sampling distribution (Wilks 2006). The confidence intervals are determined by ranking the results of the 1000 bootstrapping tests and finding the 2.5th and 97.5th percentiles. The standard error in a Brier score is inversely proportional to the squared sample size (Ferro 2007). The sample sizes used in this study are deliberately selected to be very large, and it is expected that the confidence intervals will be relatively small for the large area-aggregated statistics herein.

3. Results

The Brier skill scores and other statistics used in this study are area-aggregated over the following regions: the Northern Hemisphere (23°–75°N; land-only T2m and precipitation; ocean-only SST), the tropics (23°S–23°N; ocean and land precipitation), and the Niño-3.4 region (5°S–5°N, 190°–240°E; SST). The substantial drawback to a limited hindcast dataset (in this case, 29 forecast/observation pairs per grid point, initial condition, and target season) is that it is difficult to achieve robustness in analyzed statistics. Area and other aggregation methods may obscure local and seasonal variations, but they allow for a large sample and robust statistics and therefore a meaningful comparison of results. As the current study is a baseline, the ability to confidently compare potentially small differences in skill from different forecasting systems is important.

We first examine the region and variable where the highest skill is expected based on previous work: SST in the Niño-3.4 region (Table 2). This table, and the subsequent ones, present area-aggregated Brier skill scores for the six 3-month-mean leads available, averaged over all initial conditions, showing the change in BSS with lead. Lead-0 forecasts are not issued in the real-time NMME, and so the focus of this discussion is the results for lead-1 (bold column in tables). The upper and lower bounds of the 95% confidence interval of the sampling distribution derived from 1000 bootstrap samples are shown in parentheses for the lead-1 BSS. The seasonal variation of BSS in lead-1 forecasts of SST in the Niño-3.4 region is illustrated in Fig. 1, along with the confidence intervals.

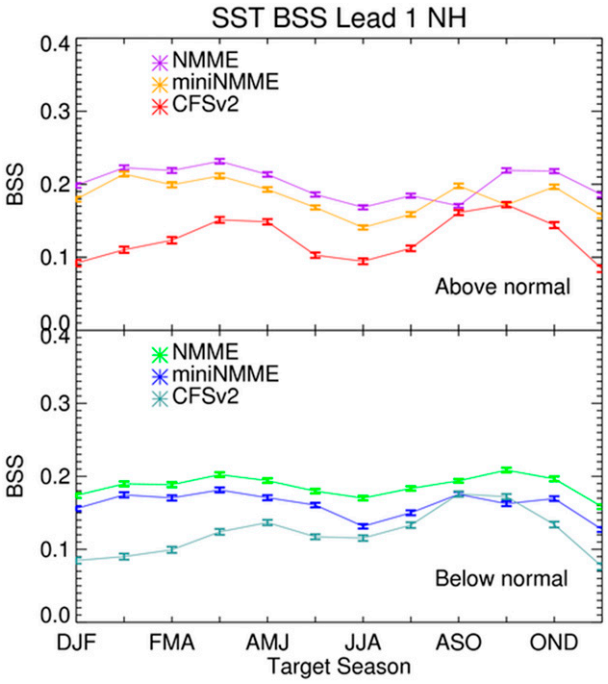

FIG. 3. As in Fig. 1, but for SST in the Northern Hemisphere (NH; all ocean grid points 23°–75°N).

We find expectations for relatively high skill in the Niño-3.4 region have been met, with BSS for the NMME lead-1 forecasts of 0.60 for above normal and 0.58 for below normal (Table 2). BSS generally ranges from 0 (skill no better than a climatological forecast) to 1 (a perfect forecast, where a probability of 100% was issued for the category that was observed). Negative BSS can occur, when the model forecast had less skill than a climatological forecast. BSS = 1 is likely impossible. A BSS of 0.6 represents a Brier score 60% better than the Brier score for the climatological forecast (always issuing a forecast probability of 0.33). BSS = 0.60 is quite high, indicating that confident probabilities were issued for the category that was observed. Even forecasts in the near-normal tercile, a notoriously difficult target (van den Dool and Toth 1991; Kharin and Zwiers 2003), earn a BSS of 0.23 at lead 1. The 24-member mini-NMME sees equivalent BSS to the full NMME, suggesting that the larger ensemble does not increase skill in this region of high predictability. BSS for forecasts from the 24-member CFSv2 are substantially lower for all terciles. This relationship is further illustrated when the seasonal cycle of the above-normal (Fig. 1, top) or below-normal (Fig. 1, bottom) category forecasts is considered, with the NMME and mini-NMME mostly overlapping, and CFSv2 often substantially lower.

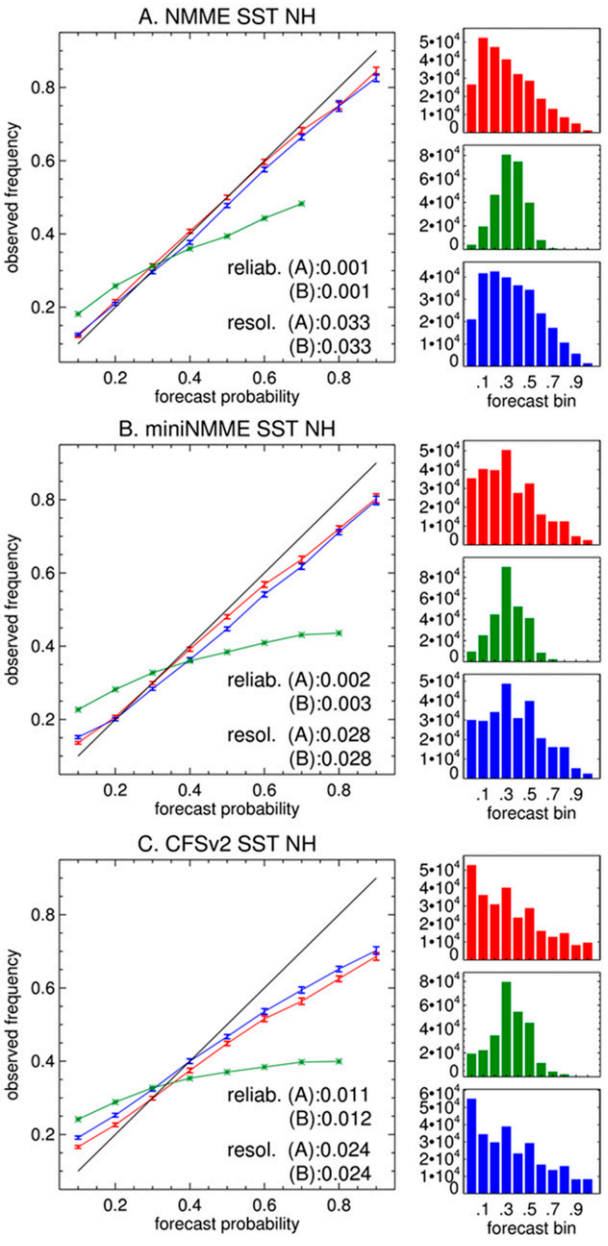

FIG. 4. As in Fig. 2, for SST in the NH (23°–75°N).

Brier skill scores are highest for forecasts for the above-normal category from all model combinations with target seasons in the boreal winter months. BSS for forecasts of the above-normal category clearly show the "spring barrier," including the March initial condition forecasts for April–June (Fig. 1, top). There is some difference in the BSS by target season between the above- and below-normal category forecasts, as the below-normal category forecasts (Fig. 1, bottom) do not show the spring barrier effect.

The advantage of the multimodel forecasts over the single model is further illustrated in the reliability

TABLE 4. As in Table 2, but for precipitation rate in the tropics (land and ocean, 23°S–23°N).

BSS: Precipitation rate in tropics						
	Lead 0	**Lead 1**	Lead 2	Lead 3	Lead 4	Lead 5
			NMME			
A	0.17	**0.119 [0.119, 0.118]**	0.10	0.08	0.06	0.05
N	0.02	**0.006 [0.006, 0.005]**	0.00	0.00	−0.01	−0.01
B	0.17	**0.118 [0.119, 0.117]**	0.10	0.08	0.06	0.05
			mini-NMME			
A	0.15	**0.095 [0.096, 0.094]**	0.07	0.05	0.04	0.03
N	0.00	**−0.02 [−0.018, −0.019]**	−0.02	−0.03	−0.03	−0.03
B	0.15	**0.097 [0.097, 0.096]**	0.07	0.05	0.04	0.03
			CFSv2			
A	0.09	**0.0038 [0.038, 0.037]**	0.02	0.00	−0.01	−0.01
N	−0.03	**−0.041 [−0.041, −0.042]**	−0.04	−0.04	−0.05	−0.04
B	0.08	**0.032 [0.032, 0.030]**	0.01	0.00	−0.01	−0.02

diagrams for SST in the Niño-3.4 region. Looking at NMME and mini-NMME reliability (Figs. 2a,b), the above and below lines (red and blue, respectively) are aligned with the diagonal (45° line), meaning these forecasts are generally highly reliable, with observed frequencies matching closely with predicted probabilities. Reliability and resolution terms from the Brier score, weighted by the frequency of use of each probability bin (insets in Figs. 2a,b), indicate that the reliability from the mini-NMME is slightly better than the NMME (smaller reliability term), and the resolution is similar but slightly higher in the full NMME.[1]

Reliability diagrams for the CFSv2 forecasts are flatter (Fig. 2c), indicating lower resolution and overconfidence; the resolution term is substantially lower than for the other two forecast configurations. Sharpness diagrams for all three forecasting systems indicate that the lowest probabilities (0–0.1) and very highest (0.9–1.0) are used much more often than others in this region; this is not unexpected, as the persistence of anomalies in this region would lead to a low likelihood that the opposite tercile would be achieved, and the shape of these diagrams is similar to that found for tropical surface temperature in DEMETER and ENSEMBLES (Alessandri et al. 2011). The CFSv2 forecasts use the highest and lowest probabilities more often than the NMME does, consistent with the overconfidence. As the points do not lie perfectly on the 45° line, and as the reliability term is not zero, even for the full NMME, we cannot say that the system is perfectly reliable (we would be surprised if it were).

Continuing with SST, but now in the extratropical Northern Hemisphere (Table 3), Brier skill scores for lead-1 probabilistic forecasts from the NMME are close to those of the mini-NMME: 0.2 versus 0.18 (A) and 0.19 versus 0.17 (B). BSS for these categories from CFSv2 forecasts are lower: 0.13 for both A and B. Forecasts for near-normal terciles are barely above zero for the NMME, and are negative for the mini-NMME and CFSv2, meaning these forecasts have too large probability anomalies (i.e., the difference between the forecast probability and the climatological probability of 0.33) regarding extratropical SST. Examining the BSS by target season (Fig. 3), we find that the NMME and mini-NMME have a greater advantage over the CFSv2 for seasons in the first half of the year for both above- and below-normal forecast categories. BSS in the late boreal summer and fall converge for the three systems, especially for forecasts of below normal. Compared to the results for the Niño-3.4 region, the importance of model diversity is slightly reduced.

The reliability diagram for forecasts of SST in the Northern Hemisphere from the NMME lies very close to the 45° diagonal for most probability bins (Fig. 4), and mini-NMME forecasts are only slightly overconfident (shallower slope). Assessment of CFSv2 forecasts reveals a substantially shallower slope and an increased reliability term (worse reliability), suggesting that the multimodel approach is important for increasing the quality of probabilistic forecasts for SST in the extratropics. However, forecasts for the highest probabilities (>80%) of both A and B categories for the full NMME do not represent the observed frequency as well (Fig. 4a), and so this may represent a limitation in the models.

Moving to a field with less predictability (and therefore expected lower skill), precipitation in the tropics has a maximum BSS for lead-1 forecasts in the above and below terciles of 0.12 for the NMME and 0.1 for the

[1] Brier score = reliability − resolution + uncertainty.

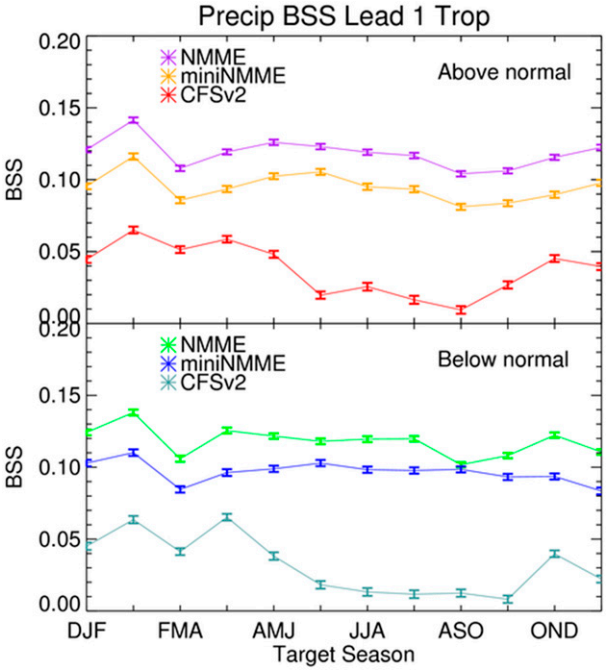

FIG. 5. As in Fig. 1, but for precipitation rate forecasts for all grid points in the tropics (23°S–23°N).

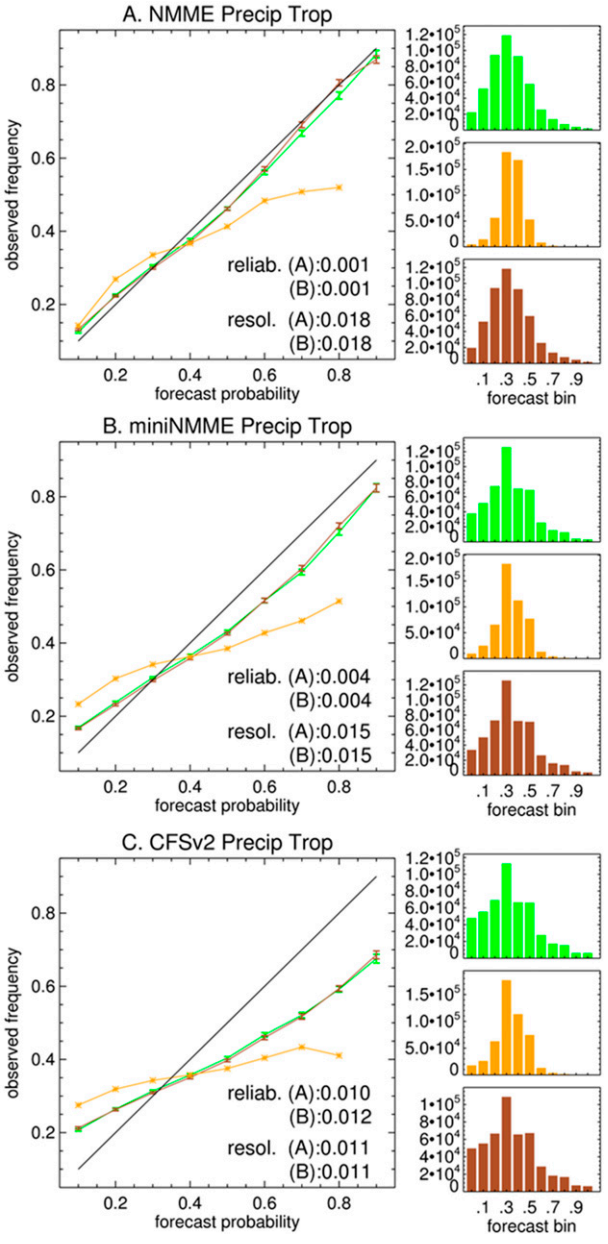

FIG. 6. As in Fig. 2, but for precipitation rate in the tropics (23°S–23°N). Green indicates forecasts in the above tercile, brown the below, and orange the near-normal.

mini-NMME (Table 4). The 24-member CFSv2 lead-1 forecasts achieve only BSS = 0.04 for the above tercile, and 0.03 for the below tercile. In both forecast configurations, the skill for forecasts in the above and below categories are nearly identical, implying that the 0.25 power transformation satisfactorily modifies the skewness of the precipitation distribution for this assessment (see section 2c for introduction of power transform procedure). Forecasts from the CFSv2 have particularly low BSS during the boreal summer and fall (Fig. 5), while less seasonal variation is seen for the two multimodel forecasts.

The 0.3 probability bin (i.e., climatological) is heavily favored by CFSv2 precipitation forecasts in the tropics, while NMME forecasts are distributed somewhat more across the 0.2–0.49 bins (Fig. 6, sharpness diagrams). The heavy use of near-climatological forecast bins indicates that while these forecasts have "good" reliability, they may not be particularly useful. Precipitation forecasts for the extratropical Northern Hemisphere land were also examined (not shown); BSS were uniformly negative for these forecasts, and NMME scores were not substantially better than the single-model performance.

Now, turning to a field with lower predictability, we find probability forecasts for 2-m surface temperature (T2m) in the Northern Hemisphere achieve BSS = 0.08 for the NMME lead-1 above and below categories (Table 5). Although modest, these scores are higher than those seen for forecasts from the mini-NMME (0.05 for A and B) and CFSv2 (0.03 for A and 0.04 for B). Both ensemble size and model diversity help for a low-skill variable. The near-normal tercile BSS is below zero from all three forecast systems. Both larger ensemble size and a multimodel ensemble have been noted to contribute value in low-skill areas such as Northern Hemisphere surface temperature forecasts (Déqué 1997; Palmer et al. 2004).

TABLE 5. As in Table 2, but for 2-m surface temperature in the Northern Hemisphere (all land grid points, 23°–75°N).

BSS: T2m in Northern Hemisphere						
	Lead 0	**Lead 1**	Lead 2	Lead 3	Lead 4	Lead 5
			NMME			
A	0.14	**0.070 [0.071, 0.069]**	0.06	0.06	0.05	0.05
N	−0.01	**−0.019 [−0.019, −0,019]**	−0.02	−0.02	−0.02	−0.02
B	0.14	**0.074 [0.075, 0.073]**	0.07	0.07	0.06	0.06
			mini-NMME			
A	0.12	**0.046 [0.046, 0,045]**	0.04	0.04	0.03	0.03
N	−0.02	**−0.037 [−0.036, −0.037]**	−0.04	−0.04	−0.04	−0.04
B	0.12	**0.047 [0.048, 0.046]**	0.04	0.04	0.03	0.03
			CFSv2			
A	0.10	**0.027 [0.028, 0,026]**	0.02	0.01	0.01	0.01
N	−0.03	**−0.042 [−0.042, −0.043]**	−0.04	−0.04	−0.04	−0.04
B	0.11	**0.034 [0.035, 0.034]**	0.03	0.03	0.02	0.02

Lower BSS is found for 2-m surface temperature forecasts of the boreal winter target seasons, and the mini-NMME and CFSv2 are near zero at times (Fig. 7). While all three forecasts score higher BSS for summer and fall seasons, CFSv2 closes the gap between it and the NMME, particularly for forecasts for the below-normal category.

Forecasts from the NMME and mini-NMME for T2m in the Northern Hemisphere have similar reliability: 0.001 and 0.002, respectively (Figs. 8a,b); the reliability term is larger for CFSv2 (Fig. 8c). Ensemble size appears to improve resolution slightly in T2m, as the two 24-member systems (mini-NMME and CFSv2) have similar resolutions, slightly lower than the full NMME. Differences between the three model systems for the uppermost forecast bins (80% and 90%) are not outside of the confidence intervals. It is remarkable that a low skill variable such as T2m can be predicted with very good reliability by the full NMME.

4. Summary and discussion

This study assessed the skill of the current method used to produce NMME probabilistic forecasts using 29 years of cross-validated hindcasts. Probabilistic forecasts from a six-model, 75-member NMME; a mini-NMME comprised of four members from each of the six models; and the 24-member CFSv2 were assessed using the Brier skill score (BSS). The hindcast assessment is employed due to the relatively short time that real-time probabilistic forecasts have been issued (about two years). The real-time forecasts have become an important tool for NOAA Climate Prediction Center seasonal forecasters, as well as many others. This study will serve as a baseline skill assessment before a series of improvements are attempted for the probability forecast construction method.

The use of large area aggregations in this study has the benefit of allowing a high degree of statistical significance, so that we may comfortably compare these relatively low scores between forecast systems. The drawback of pooling many geographical areas is that it is unlikely that the true reliability of forecasts for the southeastern United States is the same as that of forecasts for Siberia. A further study that marks specific regions with higher or lower reliability [such as the method of Weisheimer and Palmer (2014)] would be interesting, especially after the method of constructing

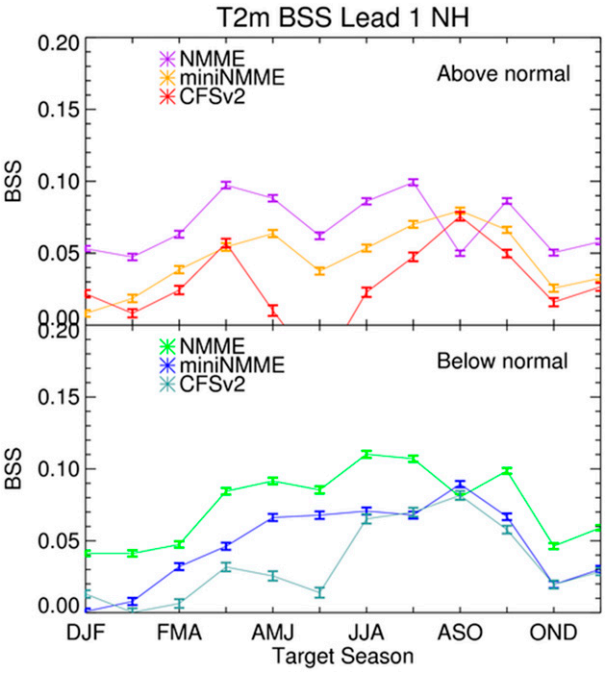

FIG. 7. As in Fig. 1, but for forecasts of 2-m temperature for all land grid points in the NH (23°–75°N).

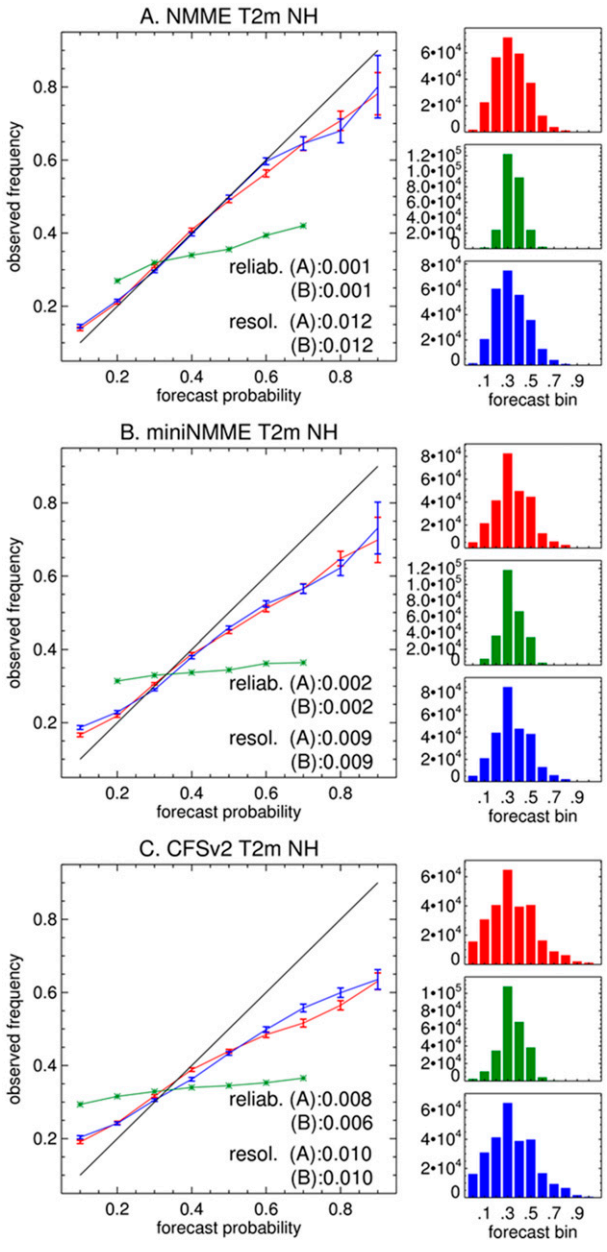

FIG. 8. As in Fig. 2, but for 2-m surface temperature over land in the NH.

probabilistic forecasts has been advanced. As well, a thorough comparison to the results of earlier MME systems, such as DEMETER and ENSEMBLES, while out of the scope of the current study, would potentially reveal the effects of higher horizontal resolution and other model improvements.

For all of the areas and fields (2-m land-only surface air temperature in the Northern Hemisphere, sea surface temperature in the Niño-3.4 region and the extratropical Northern Hemisphere, and precipitation in the tropics) the 95% confidence intervals for BSS for NMME forecasts are higher than those of CFSv2 forecasts. Skill scores for SST forecasts in both the Niño-3.4 region and Northern Hemisphere were very close or equivalent for the NMME and the mini-NMME, and higher than CFSv2, suggesting that the model diversity is an important source of increased skill in these forecasts. Results for precipitation rate in the tropics had a similar pattern, although skill scores were lower in general in this field. A lesser effect is seen in forecasts for Northern Hemisphere T2m, where the modest scores increased from CFSv2 to the mini-NMME, and again from the mini-NMME to the NMME. It appears that the importance of model diversity goes up as a function of the forecast skill, and the importance of ensemble size goes down. Forecasts in the near-normal tercile are near or below zero for all fields except sea surface temperature in the Niño-3.4 region; even there, BSS of the near-normal tercile is nevertheless much lower than the above and below categories.

This preliminary study lends confidence that the NMME probabilistic forecasts, even as is, provide value beyond that of the CFSv2 alone. The NMME benefits from both a higher number of ensemble members and model diversity. Further study is required to assess the sources of improved skill in the various fields and regions. While the beneficial effect of larger ensemble sizes is well known (e.g., Déqué 1997; Kumar et al. 2001; Richardson 2001), the results of this study suggest that the higher skill seen in the NMME is not merely the effect of a larger ensemble. Ensemble member solutions within a single model have a degree of correlation with each other, resulting in a reduced effective number of ensemble members. Because of this, 24 members from the CFSv2 are likely not perfectly equivalent to 24 members collected from six models, as the effective number of ensemble members would vary. This is something to note, but a topic for another study.

While skill is limited for some of the elements and domains studied, the visual inspection (via reliability diagrams) brings out a generally high reliability, even when skill is modest. An example of this is Northern Hemisphere extratropical T2m (Fig. 8). This encouraging result is at least in part a consequence of the large number of ensemble members. The further improvements in BSS that we will seek in future work may not come from improved reliability, but rather from fine-tuning the sharpness distribution shown in the histograms.

For probabilistic forecasts, we presently use the count method, which is not very precise, since it converts a single forecast (with a known error, that we ignore) into terciles with two ''zeros'' and one ''one.'' The imprecision of the count method is a drawback mainly for small

ensemble sizes, less so for larger ensembles like NMME (75 members). It is likely then that when we use better techniques to convert point forecasts to probabilities for three classes (perhaps using ensemble regression; Unger et al. 2008), CFSv2 may fare better in its BSS, and the improvement of NMME over CFSv2 would be smaller.

Acknowledgments. The authors thank the three anonymous reviewers whose comments greatly improved this manuscript. This study was supported by a MAPP program award. NMME project and data dissemination is supported by NOAA, NSF, NASA, and DOE. The NMME forecasts and data archive are created, updated, and maintained by NCEP, IRI, and NCAR personnel.

REFERENCES

Alessandri, A., A. Borrelli, A. Navarra, A. Arribas, M. Déqué, P. Rogel, and A. Weisheimer, 2011: Evaluation of probabilistic quality and value of the ENSEMBLES multimodel seasonal forecasts: Comparison with DEMETER. *Mon. Wea. Rev.*, **139**, 581–607, doi:10.1175/2010MWR3417.1.

Brier, G. W., 1950: Verification of forecasts expressed in terms of probability. *Mon. Wea. Rev.*, **78**, 1–3, doi:10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2.

Déqué, M., 1997: Ensemble size for numerical seasonal forecasts. *Tellus*, **49A**, 74–86, doi:10.1034/j.1600-0870.1997.00005.x.

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.

——, and Coauthors, 2009: Addressing model uncertainty in seasonal and annual dynamical ensemble forecasts. *Quart. J. Roy. Meteor. Soc.*, **135**, 1538–1559, doi:10.1002/qj.464.

Fan, Y., and H. van den Dool, 2008: A global monthly land surface air temperature analysis for 1948–present. *J. Geophys. Res.*, **113**, D01103, doi:10.1029/2007JD008470.

Ferro, C., 2007: Comparing probabilistic forecasting systems with the Brier score. *Wea. Forecasting*, **22**, 1076–1088, doi:10.1175/WAF1034.1.

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.

Jolliffe, I. T., and D. B. Stephenson, 2012: *Forecast Verification: A Practitioner's Guide in Atmospheric Science.* 2nd ed. John Wiley and Sons, 274 pp.

Kharin, V. V., and F. W. Zwiers, 2003: On the ROC score of probability forecasts. *J. Climate*, **16**, 4145–4150, doi:10.1175/1520-0442(2003)016<4145:OTRSOP>2.0.CO;2.

——, Q. Teng, F. W. Zwiers, G. J. Boer, J. Derome, and J. S. Fontecilla, 2009: Skill assessment of seasonal hindcasts from the Canadian Historical Forecast Project. *Atmos.–Ocean*, **47**, 204–223, doi:10.3137/AO1101.2009.

Kirtman, B. P., and D. Min, 2009: Multimodel ensemble ENSO prediction with CCSM and CFS. *Mon. Wea. Rev.*, **137**, 2908–2930, doi:10.1175/2009MWR2672.1.

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

Kumar, A., A. Barnston, and M. Hoerling, 2001: Seasonal predictions, probabilistic verifications, and ensemble size. *J. Climate*, **14**, 1671–1676, doi:10.1175/1520-0442(2001)014<1671:SPPVAE>2.0.CO;2.

Merryfield, W. J., and Coauthors, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

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, doi:10.1175/BAMS-85-6-853.

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

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

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, 2014: The NCEP Climate Forecast System version 2. *J. Climate*, **27**, 2185–2208, doi:10.1175/JCLI-D-12-00823.1.

Tebaldi, C., and R. Knutti, 2007: The use of the multi-model ensemble in probabilistic climate projections. *Philos. Trans. Roy. Meteor. Soc.*, **365**, 2053–2075, doi:10.1098/rsta.2007.2076.

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

van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for near normal often fail? *Wea. Forecasting*, **6**, 76–85, doi:10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

Vernieres, G., M. M. Rienecker, R. Kovach, and C. L. Keppenne, 2012: The GEOS-iODAS: Description and evaluation. NASA Tech. Rep. NASA/TM-2012-104606, Vol. 30, 61 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.]

Weisheimer, A., and T. Palmer, 2014: On the reliability of seasonal climate forecasts. *J. Roy. Soc. Interface*, **11**, 20131162, doi:10.1098/rsif.2013.1162.

——, and Coauthors, 2009: ENSEMBLES: A new multi-model ensemble for seasonal-to-annual predictions—Skill and progress beyond DEMETER in forecasting tropical Pacific SSTs. *Geophys. Res. Lett.*, **36**, L21711, doi:10.1029/2009GL040896.

Wilks, D., 2006: *Statistical Methods in the Atmospheric Sciences.* Academic Press, 627 pp.

Xie, P., and P. A. Arkin, 1997: Global precipitation: A 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. *Bull. Amer. Meteor. Soc.*, **78**, 2539–2558, doi:10.1175/1520-0477(1997)078<2539:GPAYMA>2.0.CO;2.

Zhang, S., M. J. Harrison, A. Rosati, and A. 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.