Predictability and Forecast Skill in NMME

EMILY BECKER, HUUG VAN DEN DOOL, AND QIN ZHANG

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

(Manuscript received 2 October 2013, in final form 21 April 2014)

ABSTRACT

Forecast skill and potential predictability of 2-m temperature, precipitation rate, and sea surface temperature are assessed using 29 yr of hindcast data from models included in phase 1 of the North American Multimodel Ensemble (NMME) project. Forecast skill is examined using the anomaly correlation (AC); skill of the bias-corrected ensemble means (EMs) of the individual models and of the NMME 7-model EM are verified against the observed value. Forecast skill is also assessed using the root-mean-square error. The models' representation of the size of forecast anomalies is also studied. Predictability was considered from two angles: homogeneous, where one model is verified against a single member from its own ensemble, and heterogeneous, where a model's EM is compared to a single member from another model. This study provides insight both into the physical predictability of the three fields and into the NMME and its contributing models.

Most of the models in the NMME have fairly realistic spread, as represented by the interannual variability. The NMME 7-model forecast skill, verified against observations, is equal to or higher than the individual models' forecast ACs. Two-meter temperature (T2m) skill matches the highest single-model skill, while precipitation rate and sea surface temperature NMME EM skill is higher than for any single model. Homogeneous predictability is higher than reported skill in all fields, suggesting there may be room for some improvement in model prediction, although there are many regional and seasonal variations. The estimate of potential predictability is not overly sensitive to the choice of model. In general, models with higher homogeneous predictability show higher forecast skill.

1. Introduction

In the process of making weather or climate forecasts and verifying them against observations, we cannot escape the question: ''How much better could we do?'' This is true when forecasts are in general very good—for example, extratropical Northern Hemisphere 500-hPa height 2 days out—but even more so when we are verifying modest-skill seasonal forecasts, perhaps of 2-m temperature over land. Is the low skill of a seasonal prediction 6 months out a problem we can overcome or is it always going to be this way, with only marginal improvements to be expected? This line of thinking leads to the notion of potential predictability, which roughly means the forecast skill level we would achieve if no impediments (e.g., a lack of computing power and/or observations) existed. Thus, predictability is a theoretical notion, a property of the fluid itself that exists regardless of the human endeavor.

Coming up with a definition that escapes our own lack of understanding may not be easy—it sounds contradictory—but in the late 1960s the first numerical weather prediction (NWP) model experiments were conducted to track the divergence of forecasts due to a small uncertainty in the initial condition (Lorenz 1969, 1982). This is referred to as studying the predictability of the first kind. The essence is that one model run is taken to be the forecast, and another run by the same model, starting from slightly different initial conditions, is the ''verification.'' This is called the perfect model assumption, because both the forecast and the proxy observation come from the same world, which is not the case when verifying against a real observation. The answer one gets still depends on the realism of the model being used. However, if the model is realistic, one can measure predictability this way.

The definition of predictability of the first kind is more or less accepted. It has been qualified by the disclaimer ''of the first kind'' because of a different kind of predictability forced into the interior of the fluid by anomalous boundary conditions. The latter has been called predictability of the second kind, to distinguish it from

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

DOI: 10.1175/JCLI-D-13-00597.1

predictability that is only limited by the growth of initially small errors. A strict definition of predictability of the second kind has been difficult to agree on. Considering the anomalous boundary condition as external to the system may have been a necessity early on, and it led to such things as Atmospheric Model Intercomparison Project (AMIP) runs (Gates 1992), where an atmospheric model was run N times for any number of years with prescribed observed global SST. In this case, we are acting as though we know the boundary condition perfectly ahead of time. This is too optimistic, likely yielding an overestimate of predictability of the second kind. This is even more so for runs with prescribed soil moisture anomalies, because soil moisture can change very quickly when forced by precipitation events. As well, the same AMIP runs may also be too pessimistic: with sea surface temperature (SST) anomalies forcing the atmosphere, but not the other way around, the achievable physical realism is reduced. The latter has now been overcome (albeit imperfectly) by fully coupled models. In a fully coupled model, the erstwhile boundary conditions (SST over the ocean and soil moisture over land) have become part of the initial condition. Thus, we are at liberty to apply the original intent and definition of predictability of the first kind to the seasonal forecast problem, a primary focus of this paper.

Recently, the North American Multimodel Ensemble (NMME) has been launched in the United States (Kirtman et al. 2014), with real-time experimental-operational forecasts out of National Oceanic and Atmospheric Administration (NOAA)/National Centers for Environmental Prediction (NCEP) starting in August 2011. Each of the participating models comes with an approximately 30-yr hindcast (intended for systematic error correction of the mean and calibration of probabilities in subsequent real-time forecasts), used in this study to gauge prediction skill as well as predictability. This allows us to compare each model's predictability estimate over the exact same years and hence to address several questions: Does predictability still depend heavily on the model (e.g., Rodwell and Doblas-Reyes 2006)? If not, the perfect model assumption may be more acceptable. Do forecast skill and predictability (over the same period, using the same metric) have a strong relationship? Since NMME has many models, one can distinguish homogeneous from heterogeneous predictability: that is, to verify a model's EM forecast against a single member of the same model (homogeneous) or a single member of another model (heterogeneous). In the past, expressions like identical and fraternal twin experiments have been used.

Some assessment of the models' representation of reality is made herein, by comparing the interannual variability of each model to the observations. Examining the interannual variability of each model versus observations addresses the question of systems being over- or underdispersive. Metrics for assessing dispersion have varied: Johnson and Bowler (2009) noted a tendency for the community designing seasonal prediction systems to judge over- or underdispersion by the interannual variance, while shorter-range ensemble weather prediction has most often been judged by comparing the spread of the members of an ensemble to the root-mean-square error of the forecast (i.e., an actual skill attribute). They go on to show how these two approaches are in fact related. We will comment regularly on both ways of judging whether the NMME models are under or over dispersive. The details are given in the appendix. An additional study of the models' realism is performed through the comparison of month-to-month persistence in models and observations.

The main focus in this paper is to study prediction skill and predictability, from various angles, using each of the models in the NMME. In the course of doing so, each model's own systematic error in the mean is removed. The topic of mean error is not further addressed in this paper; it is perhaps a subject for future studies. On the other hand, we do not correct for systematic errors in the standard deviation (SD), because the spread around the ensemble mean and the magnitude of the interannual variance are precisely the focus of investigation when discussing predictability. As this paper is research oriented, probability scores and how the forecast is presented to the user in real time are not addressed herein. The reader should also note that, while we are using NMME models, for the most part we calculate results for each model separately and then compare the outcomes among models, and the NMME as a multimodel system is not the focus.

Three variables were saved for NMME phase 1, on a global 1.0° latitude by 1.0° longitude grid: monthly means of 2-m temperature (T2m), precipitation rate (prate), and SST. We will apply the calculations outlined in section 2 to the three fields in each model, focusing on the Northern Hemisphere, with additional focus on the Niño-3.4 region for SST. The leads of these forecasts are at least 1 month: that is, we are focusing on short-term climate prediction, beyond the 1–2-week range of weather prediction. Results are presented in section 3, and a summary and discussion are in section 4.

2. Data and methods

a. *The North American Multimodel Ensemble project*

The NMME is a forecasting system consisting of coupled models from U.S. and Canadian modeling centers

TABLE 1. All models included in the North American Multimodel Ensemble project, phase 1.

Model	Model expansion	Organization	Hindcast period	Ensemble size	Lead times	Reference
CFSv1	Climate Forecast System, version 1	NCEP	1981–2009	15	0–8 months	Saha et al. 2006
CFSv2	Climate Forecast System, version 2	NCEP	1982–2010	24 (28)	0–9 months	Saha et al. 2014
GFDL CM2.1	Geophysical Fluid Dynamics Laboratory (GFDL) Climate Model, version 2.1	GFDL	1982–2010	10	0–11 months	Zhang et al. 2007
ECHAM4-a	—	IRI	1982–2010	12	0–7 months	DeWitt 2005
ECHAM4-f	—	IRI	1982–2010	12	0–7 months	DeWitt 2005
CanCM3	Third Generation Canadian Coupled Global Climate Model	Canadian Meteorological Centre (CMC)	1981–2010	10	0–11 months	Merryfield et al. 2013
CanCM4	Fourth Generation Canadian Coupled Global Climate Model	CMC	1981–2010	10	0–11 months	Merryfield et al. 2013
CCSM3	Community Climate System Model, version 3	National Center for Atmospheric Research (NCAR)	1982–2010	6	0–11 months	Kirtman and Min 2009
GEOS5	Goddard Earth Observing System Model, version 5	National Aeronautics and Space Administration (NASA)	1981–2010	10	0–9 months	Vernieres et al. 2012

(Kirtman et al. 2014). The multimodel ensemble approach has been shown to produce better prediction quality on average than any single model ensemble, motivating the NMME undertaking (Palmer et al. 2004; Hagedorn et al. 2005; Doblas-Reyes et al. 2005; Smith et al. 2013). The environmental variables included in the first 2 yr of phase 1 (August 2011–July 2013) are T2m, SST, and prate; real-time and archived forecast graphics from August 2011 to the present are available online (at http://www.cpc.ncep.noaa.gov/products/NMME). Other environmental variables were added in year 2, including soil moisture, maximum and minimum surface temperature, and 200-hPa geopotential height. Hindcast and forecast data are archived at the International Research Institute for Climate and Society (IRI) (accessible at http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/).

Table 1 lists the models included in NMME phase 1 (including their expanded names). The first column includes the center where each model was produced, and the name of the model. All model outputs have 1.0° latitude by 1.0° longitude horizontal resolution and forecast leads of at least 0–7 months. Each model was run retrospectively, and 29 yr of hindcasts (1982–2010) were available for all models except CFSv1 (28 yr: 1982–2009), for all 12 initial months. Year-1 models included in the real-time forecasts were CFSv1, CFSv2, ECHAM4-a, ECHAM4-f, GFDL CM2.1, CCSM3, and GEOS5. Year-2 real-time NMME forecasts comprise CFSv2, CanCM3, CanCM4, GFDL CM2.1, CCSM3, and GEOS5. GFDL CM2.1 underwent some modifications between years 1 and 2; hindcasts from the year-1 version are used in this study. The models have various ensemble sizes, ranging from 6 members to 24. Further details about the individual models can be found in their reference papers, listed in Table 1.

Some ambiguity is found in the literature regarding the definition of forecast lead. In this study, by "1-month lead," we mean a forecast made from initial conditions at the beginning of one month for the next. For example, the 1-month-lead forecast from June initial conditions (IC) is the forecast for July. (The forecast for June itself would be the "lead zero" forecast.) Following this, the seasonal lead-1 forecast is for the first complete 3-month period following the initial month: in the June IC example, the lead-1 seasonal forecast is for the July–September period.

b. *Verification fields*

The observation verification field for T2m is the station observation-based Global Historical Climatology Network + Climate Anomaly Monitoring System (GHCN+CAMS; Fan and Van den Dool 2008), a monthly mean surface air temperature dataset. GHCN+CAMS combines two large individual datasets of station observations. GHCN+CAMS has a native resolution of 0.5° latitude × 0.5° longitude and was regridded to 1.0° × 1.0° for NMME purposes. As the 7-month-lead forecasts initialized in 2010 stretch into 2011, so the observation

period used in this study for all verification fields runs from January 1982 to May 2011.

The Climate Prediction Center (CPC) global daily Unified Raingauge Database (URD) gauge analysis provides the verification field for precipitation rate. This global land-only dataset uses quality-controlled input from over 30 000 stations in the Global Telecommunication System (GTS) and many other national and international collections (P. Xie et al. 2010, personal communication). The URD is also available on a 0.5° latitude–longitude grid and was regridded to 1.0° × 1.0° for this study. The daily data were averaged into monthly means.

The sea surface temperature 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. Predictability and forecast skill

This study assesses forecast skill, homogeneous predictability, and heterogeneous predictability for the 29 yr of hindcasts for all models, primarily using the anomaly correlation (AC). The AC is a measure of the association between the anomalies of (usually) gridpoint forecast and verification values (Wilks 1995; Van den Dool 2007). By "forecast skill," we generally mean an assessment of how well each model's ensemble mean (EM) forecasts the observed value; however, verification attributes for single members of each model are included as well.

Two approaches are taken to considering predictability. Homogeneous predictability assesses one model's EM, based on $N - 1$ members, against the one member that is left out (the proxy observation). Heterogeneous predictability refers to one model's EM (based on all N members) versus one member of another model. Heterogeneous predictability is obviously part of NMME with its many models, and the heterogeneous situation resembles verification against observations in that the model is obviously not perfect, and a systematic error correction could be done.

The fields used in this study are monthly means and can be represented by $O(s, j, m)$ and $F(s, j, m, \tau)$ for the observation and forecast fields, respectively. Here, s is a spatial (grid point) index, j indicates the year (1982–2010, except for CFSv1, 1982–2009), and m is the target month. The argument τ, the forecast lead, is only a consideration for the model forecasts. Here, we primarily focus on the lead-1 season. The lead-1 seasonal forecast from January is February–April (FMA) and so on. In this study, the zero lead is not considered since it is mainly weather prediction.

The NMME models produce ensemble forecasts, and we can express the forecast as $F(s, j, m, n, \tau)$, where n indicates the ensemble member number, $n = 1, \ldots, N$, where N is the total number of ensemble members per model. The EM for each model is constructed as

$$F_{\text{ens}}(s,j,m,\tau) = \Sigma_n F(s,j,m,n,\tau)/N. \tag{1a}$$

An additional quantity, the NMME ensemble mean, is formed by averaging together the EMs of all the K models,

$$F_{\text{NMME}}(s,j,m,\tau) = \Sigma_k F_{\text{ens},k}(s,j,m,\tau)/K, \tag{1b}$$

where k is the model number, $k = 1, \ldots, K$.

For NMME phase 1 real-time forecasts at CPC, the model's EMs are equally weighted in the NMME EM, a practice that is used in this study on the hindcasts.

Anomalies are formed by subtracting climatological means: that is, for the observations,

$$O'(s,j,m) = O(s,j,m) - C_o(s,m), \tag{2a}$$

where $C_o(s, m)$ is the observed local climatology at grid point s, in this case calculated over the period 1982–2010.

Long-lead dynamical model forecasts often include a systematic bias, and it is desirable to correct for this bias (Smith and Livezey 1999). For this study, we present results that do not include cross validation (e.g., Becker et al. 2013; see note in section 4) and so employ a shortcut in calculating the anomalies for both the EM forecast F_{ens} and single member forecasts F_n, using the climatology of the model,

$$F'_{\text{ens}}(s,j,m,\tau) = F_{\text{ens}}(s,j,m,\tau) - \{F_{\text{ens}}(s,m,\tau)\} \tag{2b}$$

and

$$F'_n(s,j,m,n,\tau) = F_n(s,j,m,n,\tau) - \{F_n(s,m,n,\tau)\} \tag{2c}$$

where $\{\cdot\}$ is the mean over the 1982–2010 available forecast period.[1] Equations (2b) and (2c) effectively remove the mean biases model by model. The biases themselves are not studied explicitly in this paper.

The anomaly correlation can be written as

$$\mathrm{AC}(m,\tau) = \frac{\sum_s \sum_j \frac{w_s X'(s,j,m,t) Y'(s,j,m,\tau)}{W}}{\left[\sum_s \sum_j \frac{w_s X'(s,j,m,\tau)}{W}\right]^2 \left[\sum_s \sum_j \frac{w_s Y'(s,j,m,\tau)}{W}\right]^2}, \tag{3}$$

[1] There is a subtle and debatable point as to whether $\{F_{\text{ens}}(s, m, \tau)\}$ or $\{F_n(s, m, n, \tau)\}$ should be subtracted in Eq. (2c).

where $X'(s, j, m, \tau)$ is a prediction, and $Y'(s, j, m, \tau)$ is a verification field (Wilks 1995; Van den Dool 2007). The double summation in Eq. (3) is over all years ($j = 29$; 28 if CFSv1 is either X or Y) and space. A weight w_s is included to account for the area represented by each grid point; W is the sum of w_s over all grid points and time steps. The numerator is a covariance, and the two terms in the denominator are standard deviations; the double summation is performed on these three terms, and then the multiplication and division is carried out. In the case of assessing forecast skill, X is a model EM and Y is an observation. In the case of homogeneous predictability (AC_{hom}), X is the EM of one model's $N - 1$ ensemble members and Y is the remaining member. In the case of heterogeneous predictability (AC_{het}), X is the EM of one model (all members) and Y is a single member from another model. The anomaly correlation coefficient AC is a number between -1 and $+1$, where $+1$ refers to a perfect forecast and 0 refers to random forecasts.

The root-mean-square error (RMSE) is

$$\mathrm{RMSE}(m,\tau) = \left(\sum_s \sum_j \frac{w_s [X(s,j,m,\tau) - Y(s,j,m)]^2}{W} \right)^{1/2}, \tag{4}$$

with latitude weighting where appropriate. In this case, bias-corrected forecasts are used: X and Y are from Eqs. (2b) or (2c) and (2a), respectively. Similar to the descriptions of homogeneous and heterogeneous predictability as per AC in Eq. (3), the X and Y in Eq. (4) can be drawn from a single model, different models, single members, the ensemble mean, or the observed values. The appendix lays out in some detail the eight RMS differences one can contemplate, five of which are calculated (boldface entries in Table A1).

The ensemble spread, defined as the SD of the forecasts of the members of an individual model, is an important indicator of a model's representation of reality. If the forecast X is the EM of one model's $N - 1$ members and Y is the remaining member [i.e., ensemble mean homogeneous RMSE (EM_RMSE_{hom})], the ensemble spread is related to the EM_RMSE_{hom} such that

$$\mathrm{Spread} = \mathrm{EM\_RMSE}_{\mathrm{hom}}. \tag{5}$$

Spread can also be expressed approximately as a function of the SD of the EM and the SD of an individual member,

$$\mathrm{Singmem\_SD}^2 = \mathrm{EM\_SD}^2 + \mathrm{Spread}^2. \tag{6}$$

We will also judge dispersion by comparing Singmem_SD to Obs_SD. Johnson and Bowler (2009) discussed both possibilities. The reader is referred to the appendix for more details, nomenclature, and notation.

It is expected that AC_{hom} will be higher than the AC: that is, we expect that predictability is greater than our current skill.[2] By a different measure but following the same logic, we expect the RMS difference between the $N - 1$ EM and a single member will be lower than the RMSE against observations: that is, $RMSE_{hom} < RMSE$, both with EM or Singmem as prefix.

For six of the seven NMME models, 29 yr of hindcasts are available; CFSv1 has 28 yr (1982–2009). When a multimodel ensemble (MME) average was used in this study, it comprised the following models: CFSv1, CFSv2, CanCM3, CanCM4, GFDL CM2.1, CCSM3, and GEOS5. As CFSv1 hindcasts are not available for 2010, the last year of the MME is composed of six models [i.e., $K = 6$ in Eq. (1b)]. All heterogeneous predictability and forecast skill experiments that use CFSv1 are performed over the period of 1982–2009.

3. Results

A number of area-aggregated anomaly correlations are presented herein in tabular format. In Figs. 1 and 4, the areas for 2-m temperature and precipitation rate are land-only Northern Hemisphere, all grid points between 23° and 75°N. Figure 7 (shown later) for sea surface temperature, aggregates scores over Northern Hemisphere ocean, between 23° and 75°N. Figure 9a (shown later) presents sea surface temperature in the Niño-3.4 region (Barnston et al. 1997): that is, the aggregated results for all grid points in the box 5°S–5°N, 170°–120°W. Figure 9b contains results for the Niño-3.4 index: that is, the area-average SST is taken before the AC calculation and the spatial-average step in the AC is ignored.

The table in Fig. 1 includes homogeneous potential predictability (AC_{hom}; on the diagonal yellow highlight); heterogeneous potential predictability (AC_{het}); forecast skill (AC), RMSE, and SD of the EM; and forecast skill, RMSE, and SD of a single member. The SD of the observations is the eighth value in the bottom row. Each individual number is the value for "season 1" (e.g., the forecast from January initial conditions for February–April average, etc.), averaged over all 12 initial conditions, area aggregated over the areas specified above. One further result (a single number) is included: the forecast skill of the NMME 7-model EM, labeled "NMME."

[2] Because of sampling, one may encounter $AC_{hom} < AC$ at certain locations. However, for models that pass sanity checks regarding overall variance and temporal persistence this should not happen very often. See Kumar (2009) for more discussion.

TMP2m Northern Hemisphere Season 1										
	cfsv1	cfsv2	cmc1	cmc2	gfdl	nasa	ncar	obs (EM AC)	EM RMSE (C)	*EM SD*
cfsv1 EM	29	12	12	11	10	16	8	12	1.41	*0.59*
cfsv2 EM	14	38	16	24	26	28	1	29	1.32	*0.62*
cmc1 EM	11	14	30	19	18	21	5	17	1.38	*0.60*
cmc2 EM	10	23	21	38	27	27	3	27	1.36	*0.72*
gfdl EM	9	23	16	25	36	26	0	25	1.38	*0.77*
nasa EM	12	25	19	24	25	39	2	23	1.37	*0.69*
ncar EM	6	1	4	6	3	3	19	0	1.62	*0.84*
singmem AC	3	13	7	16	13	11	0	NMME=29		
singmem RMSE	1.91	1.79	1.77	1.76	1.90	1.79	1.99			
singmem & obs SD	*1.41*	*1.35*	*1.26*	*1.38*	*1.54*	*1.34*	*1.46*	*1.38*		

FIG. 1. Area-aggregated results for 2-m temperature, land-only Northern Hemisphere, 23°–75°N, averaged over the 12 lead-1 seasons. The seven rows and columns in black are results for predictability. The orange highlighted diagonal shows the homogeneous potential predictability anomaly correlation (AC). The black, off-diagonal elements show the heterogeneous potential predictability AC. As the EM is the prediction and the single member is the verification, each row reads horizontally. For example, the row labeled "cmc1 EM" shows AC_{het} for the CanCM3's EM prediction of a single member of CFsv1, CFSv2, AC_{hom} for itself, AC_{het} for CanCM4, AC_{het} for GFDL CM2.1, and so on. Blue column: forecast skill (AC) of the model EMs verified against observations and, in the 8th row, of the 7-model NMME. Green column: RMSE of the model EMs. Black column: SD of the model EMs. Blue row: skill (AC) of a single model member, verified against observations. Green row: RMSE of single model members. Black row: SD of the single model member and (8th column) of the observations. All anomaly correlations presented in the tables have been multiplied by 100 for visibility.

To arrive at a single AC for each cell in the tables, the following steps were taken. First, the prediction and verification fields were chosen. In the case of AC_{hom}, prediction is EM_{N-1} (N = number of ensemble members) and verification is the remaining member. In the case of AC_{het}, prediction is EM_N of one model, verification is a single member of another model. In the case of forecast skill, the prediction is either the EM or a single member of the model and verification is the observation. In the case of the forecast skill of the NMME EM, the prediction is the average of the seven EMs. Second, the monthly means at leads 1–3 were averaged to form season 1 and anomalies were formed as in Eqs. (2a)–(2c). Third, the anomaly correlation was generated using Eq. (3), including the double summation over the common years and designated spatial domain (with latitude weights) and the multiplication and division of the three summations. This was done for each of the 12 initial months, and those 12 resulting season-1 values were averaged together. The RMSE was arrived at via a similar method [see Eq. (4)]. SD averages are the aggregate over the spatial domain of the gridpoint SD for each EM or a single member of the model and are naturally tilted toward higher variance areas.

Examining the results for 2-m temperature over the Northern Hemisphere in detail (Fig. 1), we first focus on the size of the anomalies produced by the models. The "single member" SD, bottom row of all models (i.e., from individual model runs), agrees very well with the observations; all are in the region of 1.26°–1.54°C, versus the observed 1.38°C. This is encouraging, and different from earlier impressions [mainly from Development of a European Multimodel Ensemble System for Seasonal-to-Interannual Prediction (DEMETER); Palmer et al. 2004] that models are underdispersive. The SD of the model EM (far right column) ranges from about 0.59° to 0.84°C, which is appropriately smaller than the SD of either the individual ensemble members or the observations. This decrease in SD follows from damping of the noise, leaving mainly the signal in the EM. This relationship can be expressed as

$$\sigma_{\mathrm{EM}} = \sigma_i \sqrt{\frac{1}{N} + \rho \frac{N-1}{N}}, \tag{7}$$

where N is the number of ensemble members and ρ is the correlation between the ensemble members of a single model (M. Peña 2013, personal communication). Calculations using the EM SD, single-member SD, and the intramodel member correlation (not shown) confirm this rate of noise damping.

T2m prediction skill for the model EMs, measured here by the AC (Fig. 1, blue column), varies from 0.0 to 0.29. These are modest numbers but, aside from the 0.00 AC for one model, highly significant, because they are based on a huge sample: the uncertainty in a correlation (for a small correlation) is 1/sqrt($N_{\mathrm{eff}} - 2$), where N_{eff} is the effective number of cases (Van den Dool and Chervin 1986; Van den Dool 2007). Aggregating over large spatial domains and all seasons leads to larger N_{eff} and therefore greater statistical significance (Saha et al. 2006). For 12 seasons, 30 yr, and large domains (20–50

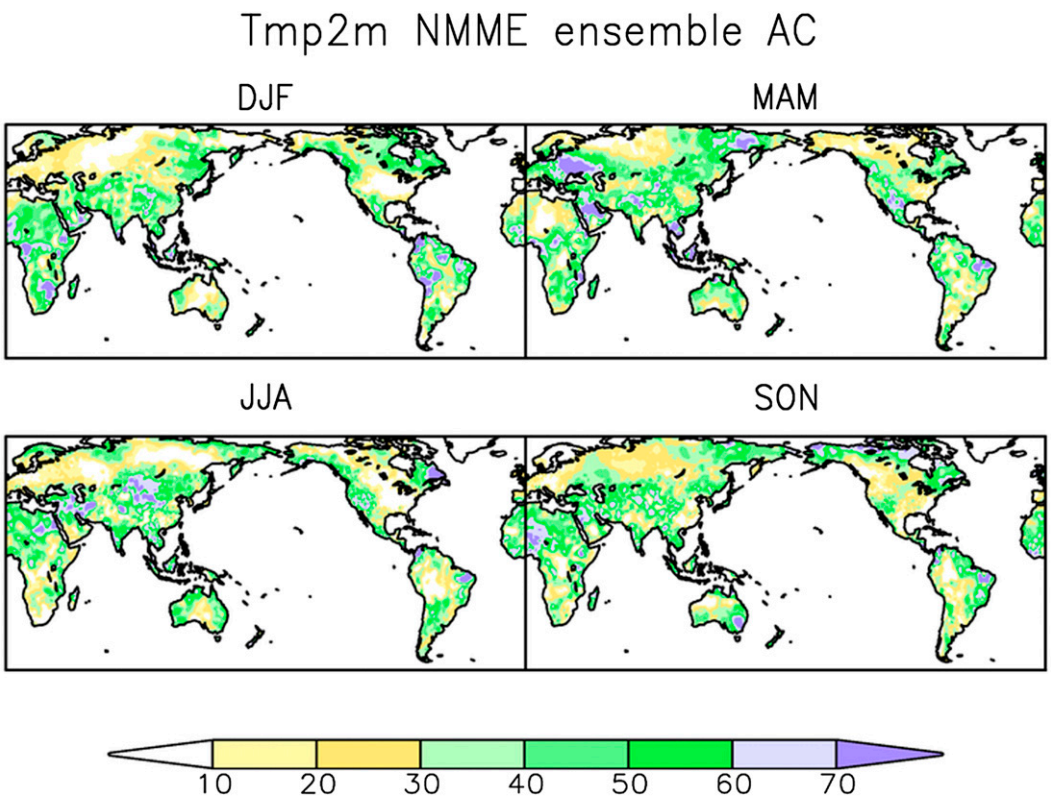

FIG. 2. Forecast skill measured by the anomaly correlation for NMME 7-model EM prediction of 2-m temperature. Four seasons are shown: lead-1 DJF, MAM, JJA, and SON. ACs are multiplied by 100. Note that the domain is near global: the NMME score of AC = 0.29 quoted in Fig. 1 corresponds to the extratropical Northern Hemisphere portion of the above maps.

spatial degrees of freedom) N_{eff} would be well over 1000, so +0.05 would be significantly different from 0 for SST and T2m. For precipitation a 0.03 correlation would be enough for statistical significance because that field has even more degrees of freedom. However, this is not to say that a 0.03 or 0.05 correlation is practically useful.

Prediction skill of 2-m temperature is low (zero) at present for the National Center for Atmospheric Research (NCAR) CCSM3, in part because this model has only ocean initialization: that is, the atmosphere and land initial state is random. The other models attempt to have a realistic atmosphere and, except GFDL CM2.1, also a land initial state, in addition to a realistic initial ocean. Forecast skill for the single ensemble members is, unsurprisingly, substantially lower than EM skill, ranging from 0.0 to 0.16. The models with highest EM skill also have the highest single-member skill. The range of RMSE results for the EMs (green column) is small (1.32–1.62), where lower AC is clearly related to higher RMSE.

The overall prediction skill AC for the NMME 7-model average is 0.29. As this AC is an area average, if we are curious about what areas show higher skill, we can examine this AC for the NMME multimodel average with the spatial and multiseasonal aggregation/averaging in Eq. (3) suppressed for the canonical lead-1 seasons (Fig. 2). To be informative about NMME, the area shown in Fig. 2 is global. This example, for December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON), shows a range of AC from under 0.1 to greater than 0.7. Northern Asia features consistently low AC through the four seasons, while southern and western Asia, as well as some areas of western North America, are generally above 0.3. ACs in the NH during DJF are the lowest overall for the four seasons shown. This supports the recent findings of Feng et al. (2013) that winter mean surface temperature over North America is not significantly predictable, but this is a big change from earlier work that suggested winter T2m was in fact more predictable than other seasons (Madden and Shea 1978). This point may not be settled, since Arribas et al. (2011) do report skill for North American winter with a modern model.

Returning to Fig. 1, the T2m homogeneous predictability AC_{hom} (orange highlighted diagonal) ranges from 0.19 to 0.39. This is higher than the reported skill

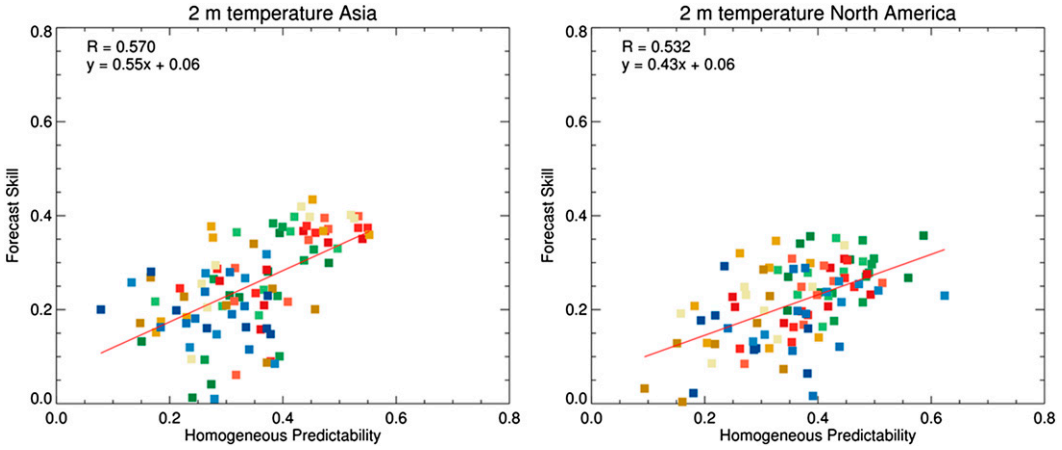

FIG. 3. The 2-m temperature homogeneous predictability AC (*x* axis) vs forecast skill AC (*y* axis). Colors indicate season: boreal winter seasons (November–January, DJF, and January–March) are blue, spring seasons (FMA, MAM, and April–June) are green, summer seasons [May–July (MJJ), JJA, and July–September (JAS)] are red, and autumn seasons (August–October, SON, and October–December) are orange. The shades of each color vary from lighter (first 3-month period; e.g., MJJ) to darker (third 3-month period; e.g., JAS). Each individual season has seven points, one for each model. The linear fit is depicted by the red line. (left) Southeastern Asia (5°–50°N, 70°–145°E) and (right) extratropical North America (north of 23°N; Greenland not included).

(≤0.29), which suggests that our forecasts might eventually improve, but not by huge margins. The heterogeneous predictability AC_{het} (Fig. 1, black off-diagonal element) ranges from 0.0 to 0.28, almost exactly the range of skill already achieved. Heterogeneous predictability and the actual skill suffer equally from a mismatch in climate between model and verification: only homogeneous predictability estimates are based with justification on a perfect model assumption. In summary, all models predict themselves better than they predict other models (or reality).

To study the relationship between homogeneous predictability and forecast skill in greater detail, the 2-m temperature AC_{hom} and forecast skill for each model and each season are shown in Fig. 3. Only lead-1 seasonal forecasts are shown, as defined above. Area aggregates are shown for southeastern Asia (5°N–50°N, 70°E–145°E) and extratropical North America (north of 23°N; Greenland not included). In a few cases, the forecast skill is higher than the AC_{hom}, but in general the predictability is the greater: that is, most points fall below the 1:1 line, as we would expect. (The sample size N_{eff} is reduced when considering individual models and seasons, over smaller geographical areas, leading to higher uncertainty in the results.) Over Asia (left), most models show both higher AC_{hom} and higher skill during the spring, summer, and early fall, although there are several exceptions to this, and spring seasons in particular are widely scattered. Forecast skill and predictability tend to increase together. North America (right) has the highest predictability and forecast skill during spring and summer, when soil moisture anomalies in one month can lead surface temperature anomalies of the opposite sign during subsequent months (Huang and Van den Dool 1993), and the lowest predictability and forecast skill during autumn and winter.

Regarding heterogeneous predictability (black off-diagonal elements in Fig. 1), we note that the 7 × 7 matrix is largely symmetric. This means that "to predict" and "to be predicted" is similar: that is, if model A (EM) can predict model B (single member), then the reverse is also true. Curiously, CCSM3 is exceptional in a way: it has a hard time predicting T2m anomalies in the other models or being predicted by the other models. While orthogonal behavior may in general be desirable, since CCSM3 also poorly verifies against the observations, its orthogonality has no discernible benefit so far. The models with the highest skill against observations, GFDL CM2.1, GEOS5, CanCM4, and CFSv2, correlate the most to each other, and all are within the AC_{het} range of 0.23–0.28. The CFSv1 and CFSv2 have a shared pedigree, but these two NCEP models do not predict each other very well: nor do the two models from Environment Canada, CanCM3 and CanCM4.

Beyond examining the models' interannual variability, another factor we can test is the persistence of anomalies in the models and how it compares to reality. Potential predictability could appear artificially high if the models are too persistent. We examined persistence in the model forecasts in all three fields, in the form of the AC of the lead-1 month forecast with the 2-month-lead forecast (not shown). When compared to the

PRATE Northern Hemisphere Season 1										
	cfsv1	cfsv2	cmc1	cmc2	gfdl	nasa	ncar	obs (EM AC)	EM RMSE	*EM SD*
cfsv1 EM	24	11	8	12	11	10	4	10	0.41	*0.21*
cfsv2 EM	13	20	7	11	9	10	4	12	0.38	*0.15*
cmc1 EM	6	6	16	14	9	7	6	9	0.40	*0.16*
cmc2 EM	10	6	13	25	13	10	5	11	0.40	*0.18*
gfdl EM	10	8	9	14	22	11	5	12	0.40	*0.20*
nasa EM	9	8	6	10	10	18	4	9	0.40	*0.19*
ncar EM	4	3	5	6	5	3	12	4	0.45	*0.25*
singmem AC	3	4	3	6	5	4	2	NMME=16		
singmem RMSE	0.67	0.61	0.56	0.56	0.60	0.61	0.65			
singmem & obs SD	*0.57*	*0.50*	*0.42*	*0.43*	*0.48*	*0.48*	*0.54*	*0.37*		

FIG. 4. As in Fig. 1, but for precipitation rate over Northern Hemisphere land. RMSE is in millimeters per day.

persistence in the observed fields, the models were found to be very similar to reality in all three fields.

Having discussed the table in Fig. 1 in great detail, we can summarize the results in the remaining tables in Figs. 4, 7, and 9 (Figs. 7 and 9 shown later) more quickly. Season-1 precipitation rate (Fig. 4) shows a range in EM prediction from AC = 0.04 to AC = 0.12. Six of the models are within 0.09–0.12, with the CCSM3 = 0.04. While it is unsurprising that skill for precipitation is lower than for T2m, the NMME 7-model-average AC of 0.16 is an improvement over the individual models and more clearly so than for T2m. The global map view (Fig. 5) shows large areas with AC < 0.1, with exceptions being some areas of western North America throughout the four seasons and the southern tier of North America in DJF, a pattern resembling an ENSO composite. The SDs of single model runs (Fig. 4, bottom) are all slightly higher than the observed SD of 0.37, indicating a small tendency toward overly dispersive forecasts.

The homogeneous potential predictability for prate (Fig. 4, yellow diagonal element) suggests there is still room for improvement in precipitation forecasting, with AC_{hom} for most models ranging from 0.16 to 0.25 (CCSM3 again is the outlier, at 0.12). While these are

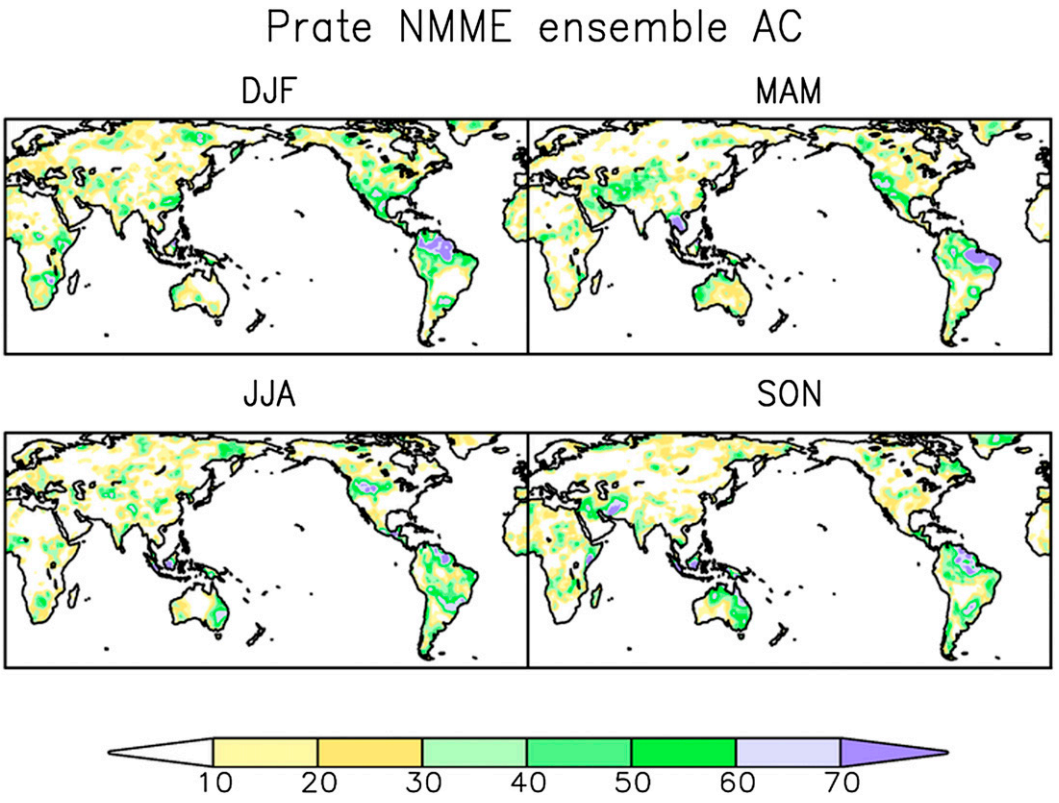

FIG. 5. As in Fig. 2, but for precipitation rate. The NMME skill of AC = 0.16 quoted in Fig. 4 refers to the extratropical NH portion of the above maps.

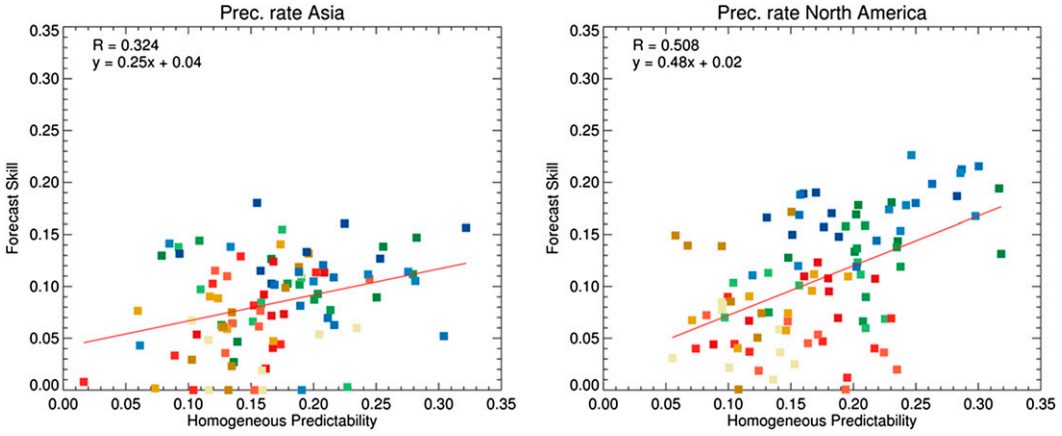

FIG. 6. As in Fig. 3, but for precipitation rate.

not particularly impressive numbers, they do represent a doubling of the EM AC skill for most models. Heterogeneous potential predictability for precipitation, like for T2m, shows that the models predict each other about as well as they predict anomalies in nature.

The relationship between precipitation AC_{hom} and forecast skill for prate, shown in Fig. 6, features a readily apparent seasonality in both Asia (Fig. 6, left) and North America (right). Forecast skill is highest during winter (blue) and spring (green) seasons, and these seasons tend to have higher potential predictability. The scores are overall very low, however: note the axes range from 0.0 to 0.35.

Turning to extratropical SST and jumping upward in both potential and realized skill, the range of ACs for the area-aggregate Northern Hemisphere extratropics (23°–75°N) is 0.29–0.46 for six of the seven individual models; the NMME skill is higher, at 0.5 (Fig. 7). NH extratropical forecasting skill is highest in the eastern Pacific, particularly during DJF and MAM, and areas of the northern Atlantic (Fig. 8). The GEOS5 and CFSv2 models have somewhat higher SDs than the observed (0.71 and 0.75, respectively, compared to 0.62 for the observation), while the other models are close to the observation. With the exception of CFSv1, potential predictability is fairly high (0.68–0.82) and substantially higher than the forecast skill, providing some hope for improved extratropical SST prediction. The AC_{hom} from CCSM3, highest of the seven models at 0.81, is astoundingly higher than the CCSM3 EM skill of 0.15. Heterogeneous predictability for three of the models with the highest skill, CFSv2, GFDL CM2.1, and GEOS5, within the range 0.32–0.47, is similar to the achieved range. The two CMC models are an exception when predicting each other, resulting in AC_{het} of 0.62–0.64.

Prediction skill for sea surface temperature in the Niño-3.4 region (5°S–5°N, 170°–120°W) is known to be especially high, particularly at such a relatively short lead of one season. ACs in this region, as for the other physical variables presented in this study, were calculated on the grid points in the region, not on the area-average Niño-3.4 index. This study confirms the expected high skill, with ACs for forecast skill of 0.80–0.88 and even the single ensemble member ACs all greater than 0.75 (Fig. 9a). However, the AC_{hom} ranging from 0.93 to 0.98 suggests even yet some room for improvement in forecasting in this area. The AC_{het} scores, like in the

SST Northern Hemisphere Season 1										
	cfsv1	cfsv2	cmc1	cmc2	gfdl	nasa	ncar	obs (EM AC)	EM RMSE (C)	*EM SD*
cfsv1 EM	49	28	38	37	33	27	23	29	0.63	*0.39*
cfsv2 EM	23	68	38	40	38	47	14	41	0.62	*0.52*
cmc1 EM	26	34	71	64	44	35	23	44	0.55	*0.40*
cmc2 EM	26	35	62	74	45	36	21	46	0.55	*0.45*
gfdl EM	22	32	42	43	78	34	17	42	0.59	*0.46*
nasa EM	19	43	33	36	36	74	12	35	0.66	*0.55*
ncar EM	16	11	21	19	16	12	81	15	0.76	*0.54*
singmem AC	15	27	35	36	35	27	12	NMME=50		
singmem RMSE	0.83	0.82	0.65	0.66	0.69	0.81	0.82			
singmem & obs SD	*0.64*	*0.75*	*0.54*	*0.58*	*0.57*	*0.71*	*0.63*	*0.62*		

FIG. 7. As in Fig. 1, but for sea surface temperature aggregate scores over Northern Hemisphere ocean, between 23° and 75°N.

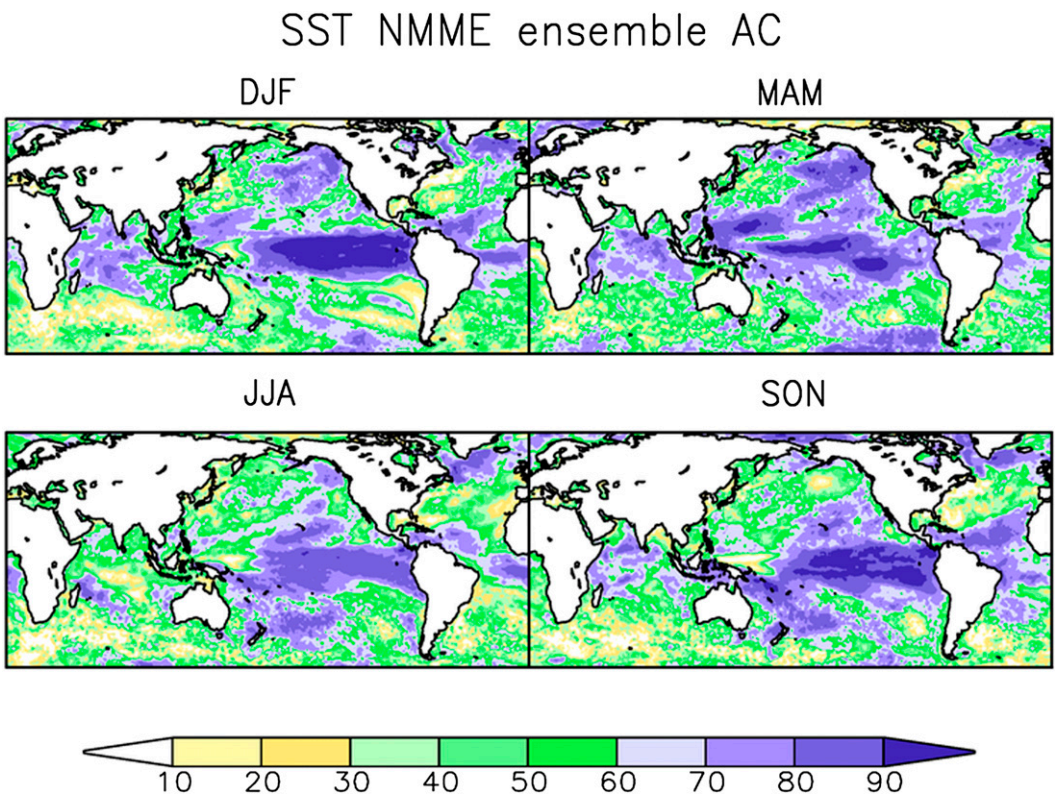

FIG. 8. As in Fig. 2, but for sea surface temperature. The NMME skill of AC = 0.50 quoted in Fig. 7 refers to the extratropical NH portion of the above maps.

other fields we have seen, are similar to those of the already-achieved forecast AC. SDs of all the models but the CanCM3 are higher than the observed 0.89°C in the Niño-3.4 region. Figure 9b shows the same but now for the Niño-3.4 index. By taking a spatial mean of SST first, models are forgiven for placement errors within the region, and all AC numbers go up and most RMSE and SD go down. However, the conclusions do not change.

GEOS5's prediction of SST in the Niño-3.4 region may be an example of an area where a model is possibly underdispersive (Fig. 9a). GEOS5 has a 0.98 correlation between the ensemble mean and the members (AC_{hom}), which means that the members cluster closely around the ensemble mean, more so than, for example, CFSv2, which features a 0.93 correlation between members and ensemble mean. The GEOS5 EM_SD is 1.17, and the single member SD is 1.22. As per Eq. (6), this results in Spread = 0.35. That may be a bit low, compared to 0.56, the RMSE for GEOS5's ensemble mean. For CFSv2, we note a Spread of 0.43, which is also somewhat low but not to the same degree as GEOS5.

Three subregions were considered in Fig. 10: SST in the extratropical Pacific, 23°– 75°N, the northern Atlantic (45°–75°N), and the Niño-3.4 region. The Niño-3.4 region (Fig. 10, bottom), as expected, has both high AC_{hom} and forecast skill (note the axes range from 0.70 to 1.0 in this panel), and higher forecast skill is found in seasons with higher predictability. Boreal winter AC_{hom}–AC pairs are distinctly higher, and summer is lowest. On the other hand, although the North Atlantic (top right) has higher potential predictability during the winter and lower during the summer, the realized forecast skill does not reflect this. In this region, AC_{hom} in general (~0.5–0.8) is substantially higher than the skill (less than ~0.6), suggesting we may hope for future improvements in forecasting SST for this area, which is a potentially important predictor for the North Atlantic Oscillation and European climate. The Pacific (top left) has higher AC_{hom} and forecast skill during the spring, but other AC_{hom}–AC pairs are more widely spread.

4. Summary and discussion

The National Multimodel Ensemble (NMME) forecast project presents a great opportunity to study the potential predictability of 2-m surface temperature, precipitation rate, and sea surface temperature, using the 29 yr of hindcasts available. In this study, we assess the potential predictability "of the first kind": that is, we make use of the perfect model assumption to assess the

A. SST Nino3.4 Region Season 1

	cfsv1	cfsv2	cmc1	cmc2	gfdl	nasa	ncar	obs (EM AC)	EM RMSE (C)	*EM SD*
cfsv1 EM	94	74	85	86	78	87	82	82	0.61	*1.12*
cfsv2 EM	69	93	78	76	80	86	80	82	0.58	*1.03*
cmc1 EM	84	76	96	93	82	90	83	87	0.43	*0.83*
cmc2 EM	83	73	92	97	80	88	82	85	0.57	*1.13*
gfdl EM	76	78	83	82	94	87	80	80	0.65	*1.13*
nasa EM	83	84	88	87	84	98	85	88	0.56	*1.17*
ncar EM	79	78	82	82	79	85	95	80	0.67	*1.14*
singmem AC	76	76	84	83	75	85	78	NMME=89		
singmem RMSE	0.74	0.70	0.46	0.63	0.73	0.62	0.70			
singmem & obs SD	*1.19*	*1.12*	*0.85*	*1.16*	*1.17*	*1.22*	*1.15*	*0.89*		

B. SST Nino3.4 Index Season 1

	cfsv1	cfsv2	cmc1	cmc2	gfdl	nasa	ncar	obs (EM AC)	EM RMSE (C)	*EM SD*
cfsv1 EM	95	76	89	91	81	89	86	86	0.52	*1.09*
cfsv2 EM	70	94	82	80	85	89	84	86	0.49	*0.99*
cmc1 EM	88	81	97	96	87	94	90	92	0.31	*0.79*
cmc2 EM	87	77	95	98	86	91	88	91	0.44	*1.09*
gfdl EM	79	84	88	87	95	92	85	86	0.52	*1.08*
nasa EM	85	87	92	91	89	98	89	92	0.48	*1.14*
ncar EM	83	83	89	88	85	89	96	87	0.52	*1.08*
singmem AC	80	82	90	89	82	90	85	NMME=93		
singmem RMSE	0.66	0.59	0.33	0.49	0.59	0.52	0.54			
singmem & obs SD	*1.15*	*1.06*	*0.80*	*1.11*	*1.11*	*1.17*	*1.08*	*0.83*		

FIG. 9. As in Fig. 1, but for sea surface temperature in (a) the Niño-3.4 region (aggregated results for all grid points in the area 5°S–5°N, 170°–120°W) and (b) the Niño-3.4 index (area-averaged SST).

predictability limited only by small errors in the initial conditions, which in a coupled model include SST and land surface. Since the NMME comprises several independent models, we can study predictability as both homogeneous—how well a single model predicts itself—and heterogeneous: how well one model predicts another. The forecast skill (how well each model, both its EM and individual members, and the multimodel ensemble verify against observations) is also analyzed, as is the variability in the forecasts as their representation of reality.

Homogeneous predictability is higher than reported skill in all fields, suggesting there may be room for some improvement in model prediction. In many cases, the margins are not especially great; for example, if we choose the highest potential for T2m forecasting, AC = 0.39 as reported by GEOS5, while the NMME reported skill of 2-m temperature forecasting is 0.29. Northern Hemisphere precipitation shows the potential for a doubling of its modest forecast skill. Sea surface temperature in the Northern Hemisphere extratropics, with skill generally on the order of AC = 0.5, has homogeneous predictability ACs up to 0.82. Even in the Niño-3.4 region, where SST forecast skill is already very high, there may be some room for improvement, if we accept this type of analysis.

Heterogeneous predictability is generally lower than homogeneous predictability, and very close to actual forecast skill. The heterogeneous estimates from various models are rather similar in this regard. This similarity among models does not imply in any way that models make the same errors (for a discussion on that aspect in the context of atmospheric blocking, see Scaife et al. 2010). The lower heterogeneous predictability gives one pause before applying predictability estimates of the first kind to the NMME as one big system, which would be dominated by heterogeneous predictability. While NMME usually has higher prediction skill than most or all individual models, NMME clearly does not have higher predictability by the method used here.

In general, models with higher homogeneous predictability show higher forecast skill. However, both potential predictability and realized forecast skill vary depending on geographical region and season. In North America, the skill of precipitation forecasts, especially during winter, is close to its potential. Sea surface temperature in the North Atlantic has higher predictability during the winter but does not show increased forecasting skill during the winter.

The estimate of predictability of the first kind is only mildly dependent on the model we use. This appears different from Rodwell and Doblas-Reyes (2006), who reported that ''potential predictability estimates are sensitive to the coupled model used'' (p. 6025). Perhaps models have improved since 2006: our results support the perfect model assumption underlying the first-kind estimate in general, unless all models have the same errors (e.g., missing physics.)

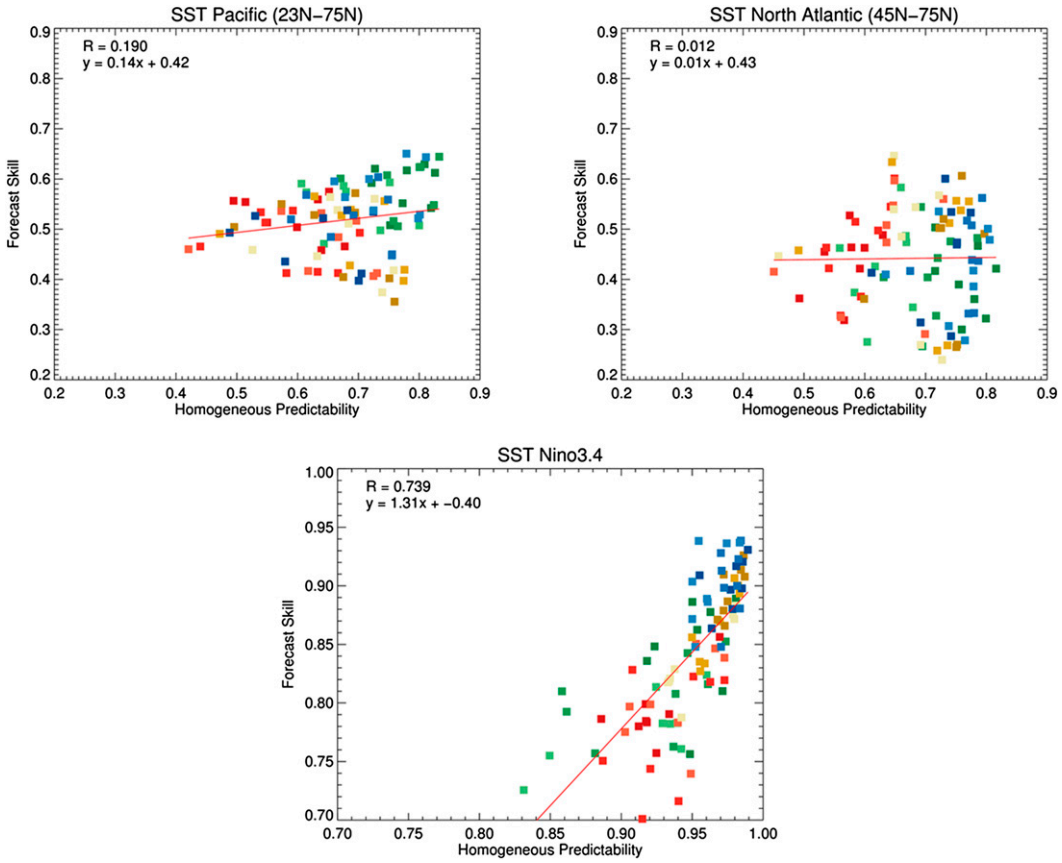

FIG. 10. As in Fig. 3, but for sea surface temperature: (top left) Pacific Ocean (23°–75°N); (top right) Atlantic Ocean (45°–75°N); and (bottom) Niño-3.4 region.

NMME 7-model forecast skill, verified against observations, is equal to or higher than the individual models' forecast ACs for all the environmental fields examined. In the case of precipitation, while all scores are admittedly low, the NMME EM is a clear improvement over the individual models. Sea surface temperature forecast skill from the NMME is also higher than the best individual model.

It does not appear that the present set of models suffers greatly from underdispersion, especially not for T2m and prate. Some of the models' prediction of SST in the Niño-3.4 region show SD less than that of observations, and a spread less than the RMSE of the ensemble mean, including GEOS5. So is GEOS5's system underdispersive or does it show high potential predictability? Ultimately, we need to strive for spread and RMSE being equal, but this should be achieved by lowering RMSE the normal way (i.e., by improving any or all models individually), not so much by increasing spread by artificial means, even if the latter helps us in making probability scores look better (although not everyone may agree with this).

There remains the question as to whether we are justified to calculate predictability by assuming a perfect model and following the logic of predictability of the first kind. A common objection is that the answer may depend too much on the model. We have shown here with seven models that the answer is not drastically different for different models. In addition, two ''sanity checks'' were applied upfront to all models. The first is about having enough overall interannual variance, and we feel all models pass this test for the variables studied. The second is about having the correct persistence of anomalies. For example, it was noted by Wu et al. (2009) that the CFSv1 persists the sign of the anomaly in the Niño-3.4 index in winter too much across the spring barrier, when nature often changes sign, leading to an overestimate of predictability for that part of the year. This problem is much reduced in CFSv2 (Saha et al. 2014). Month-to-month persistence was calculated for the three variables for all models and compared to that in the observations. Persistence in the models was found to be close to that in the observed fields. It is not clear how many more sanity checks are required before a model can be considered

a decent enough replica of nature to allow predictability estimates to be credible. One more check, comparing the EOFs of each model with those in nature (not done here), has shown great improvements over the years.

Regarding cross validation (CV) for this application, the aspect to be cross validated assumes that we know the model's climate mean. Even with 30 yr of data, this is far from perfectly true. Luckily, CV is unnecessary for assessing homogeneous predictability, since the prediction and the verification are both from the same world, so we are entitled to assuming they have the same mean. Regarding heterogeneous predictability, as well as actual forecast skill, the risk of overestimation of skill is present and CV such as "CV3RE" (Barnston and Van den Dool 1993; Van den Dool 2009; Becker et al. 2013) would be appropriate. Earlier studies have found that applying cross validation reduces low AC substantially while only slightly affecting higher correlations (Van den Dool 2007; Becker et al. 2013). In this study, the relationships between the resulting AC and AC_{hom} are of as much interest as the scores themselves, and so results are shown with no CV applied. When predictability and forecast skill was examined for cross-validated anomalies (not shown) the expected effect was observed: AC below approximately 0.1 was reduced to zero, while AC above approximately 0.4 were only slightly affected.

Acknowledgments. The authors are very grateful to three anonymous reviewers, whose detailed comments greatly improved this manuscript. The phase-1 NMME project was supported by the NOAA/MAPP program, and the phase-2 NMME project is support by NOAA/MAPP, NSF, NASA, and the DOE.

APPENDIX

Ensemble Spread and RMS Differences

We repeat Eq. (4) immediately below to show in some detail for which combination of fields one can conceivably take mean-square differences. Here we also want to tie together otherwise somewhat vague references to "dispersion," "spread," interannual standard deviation, etc.,

$$\mathrm{RMSD}(m,\tau) = \left(\sum_s \sum_j \frac{w_s [X - Y]^2}{W} \right)^{1/2},$$

where W is the sum of weights w_s over all time and space points. Here, X and Y are generic datasets, where the number of arguments of X and Y (at most 5 arguments: s, j, m, n, and τ) varies, where s, j, m, n, and τ if they apply stand for space, year, month, ensemble member, and lead, respectively. The systematic error (if the notion applies) has already been removed from X and Y before executing Eq. (4). We use the letter D in RMSD above because for some choices of X and Y the notion "error" as in RMSE does not apply. At least the following eight combinations of X and Y shown in Table A1 can be considered.

The boldfaced options 1, 2, 5, 7, and 8 (Table A1) and associated descriptors correspond exactly to what is given in Figs. 1, 4, 7, and 9. Not used explicitly in these figures are spread, option 6, and two versions of RMSE_{hom}, options 3 and 4. However, for Gaussian distributions it can be shown that

$$\sqrt{2}(\mathrm{Spread}) = \mathrm{Singmem\_RMSE}_{hom} \qquad \text{(A1)}$$

and

$$\mathrm{Spread} = \mathrm{EM\_RMSE}_{hom}, \qquad \text{(A2)}$$

while

$$(\mathrm{Singmem\_SD})^2 = (\mathrm{EM\_SD})^2 + \mathrm{Spread}^2. \qquad \text{(A3)}$$

Not included in Table A1 are heterogeneous estimates, which would be two more rows like options 3 and 4, leading to RMSE_{het}, with the added condition that Y is from a different model. For the anomaly correlation the equivalent of options 1–4 apply, but options 5–8 are unique to the RMS attribute.

Equation (A3) is only strictly valid when spread as function of time (j) does not depend on the time-dependent ensemble mean. Equation (A3) decomposes Singmem_SD, which can be compared directly to the observable Obs_SD, into two quantities, EM_SD and spread: neither of which has a counterpart in the observations but they can be calculated from a model ensemble. Obviously, when Singmem_SD is too large at least one of the two unobservable components may be too large also. Comparing Singmem_SD to Obs_SD is one popular way of judging whether a system has the right dispersion. The other popular way is to compare spread to EM_RMSE, which is the same as comparing RMSE_{hom} to RMSE, whether it has EM or Singmem as prefix.

It is expected that $\mathrm{RMSE}_{hom} < \mathrm{RMSE}$. The RMSE_{hom} is a measure for the spread of the members around the ensemble mean, so in this paper we judge there to be potential predictability over and beyond the already realized prediction skill when $\mathrm{RMSE}_{hom} < \mathrm{RMSE}$ (or $AC_{hom} > AC$; these two criteria should almost always agree). We should point out though that in many papers frequent note has been made of $\mathrm{RMSE}_{hom} < \mathrm{RMSE}$ as a sign of underdispersion, a sign of a too-narrow pdf

TABLE A1. Listing of RMS differences taken between fields X and Y, as in Eq (4). Symbol $\{\cdot\}$ means an average over 1982–2010. Arguments s, j, m, n, and τ stand for space, year, month, ensemble member, and lead, respectively. The significance of boldface entries described in text.

Option	Descriptor used in the tables in Figs. 1, 4, 7, and 9	X	Y
1*	**Singmem_RMSE**	$\boldsymbol{F(s, j, m, n, \tau)}$	$\boldsymbol{O(s, j, m)}$
2	**EM_RMSE**	$\boldsymbol{F_{\mathrm{ens}}(s, j, m, \tau)}$	$\boldsymbol{O(s, j, m)}$
3**	$\mathrm{Singmem\_RMSE_{hom}}$	$F(s, j, m, n_1, \tau)$	$F(s, j, m, n_2, \tau)$
4*	$\mathrm{EM\_RMSE_{hom}}$	$F_{\mathrm{ens}}(s, j, m, \tau)$	$F(s, j, m, n, \tau)$
5	**EM_SD**	$\boldsymbol{F_{\mathrm{ens}}(s, j, m, \tau)}$	$\boldsymbol{\{F_{\mathrm{ens}}(s, m, \tau)\}}$
6*	Spread	$F(s, j, m, n, \tau)$	$F_{\mathrm{ens}}(s, j, m, \tau)$
7*	**Singmem_SD**	$\boldsymbol{F(s, j, m, n, \tau)}$	$\boldsymbol{\{F_{\mathrm{ens}}(s, m, \tau)\}}$
8	**Obs_SD**	$\boldsymbol{O(s, j, m)}$	$\boldsymbol{\{O(s, m)\}}$

* The calculation is for single member n but, since the answer should not materially depend on n, we aggregate over all choices of n.
** The calculation is for single member n but, since the answer should not materially depend on n, we aggregate over all choices of $n_1 \neq n_2$.

needing a remedy, such as an added stochastic component (for a single model), postprocessing, or indeed the multimodel approach. So while underdispersion has been often mentioned as a bad thing, in this paper we take a positive view and appreciate potential predictability exceeding prediction skill.

REFERENCES

Arribas, A., and Coauthors, 2011: The GloSea4 ensemble prediction system for seasonal forecasting. *Mon. Wea. Rev.,* **139,** 1891–1910, doi:10.1175/2010MWR3615.1.

Barnston, A. G., and H. M. Van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate,* **6,** 963–977, doi:10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

——, M. Chelliah, and S. B. Goldenberg, 1997: Documentation of a highly ENSO-related SST region in the equatorial Pacific. *Atmos.–Ocean,* **35,** 367–383, doi:10.1080/07055900.1997.9649597.

Becker, E. J., H. M. Van den Dool, and M. Peña, 2013: Short-term climate extremes: Prediction skill and predictability. *J. Climate,* **26,** 512–531, doi:10.1175/JCLI-D-12-00177.1.

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, doi:10.1175/MWR3016.1.

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.

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.

Feng, X., T. DelSole, and P. Houser, 2013: Comparison of statistical estimates of potential seasonal predictability. *J. Geophys. Res. Atmos.,* **118,** 6002–6016, doi:10.1002/jgrd.50498.

Gates, W. L., 1992: AMIP: The Atmospheric Model Intercomparison Project. *Bull. Amer. Meteor. Soc.,* **73,** 1962–1970. doi:10.1175/1520-0477(1992)073<1962:ATAMIP>2.0.CO;2.

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.

Huang, J., and H. M. Van den Dool, 1993: Monthly precipitation–temperature relations and temperature prediction over the United States. *J. Climate,* **6,** 1111–1132, doi:10.1175/1520-0442(1993)006<1111:MPTRAT>2.0.CO;2.

Johnson, C., and N. Bowler, 2009: On the reliability and calibration of ensemble forecasts. *Mon. Wea. Rev.,* **137,** 1717–1720, doi:10.1175/2009MWR2715.1.

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 Multi-Model Ensemble (NMME): Phase-1 seasonal to interannual prediction, phase-2 toward developing intra-seasonal prediction. *Bull. Amer. Meteor. Soc.,* **95,** 585–601, doi:10.1175/BAMS-D-12-00050.1.

Kumar, A., 2009: Finite samples and uncertainty estimates for skill measures for seasonal prediction. *Mon. Wea. Rev.,* **137,** 2622–2631, doi:10.1175/2009MWR2814.1.

Lorenz, E. N., 1969: Three approaches to atmospheric predictability. *Bull. Amer. Meteor. Soc.,* **50,** 345–349.

——, 1982: Atmospheric predictability experiments with a large numerical model. *Tellus,* **34A,** 505–513, doi:10.1111/j.2153-3490.1982.tb01839.x.

Madden, R. A., and D. J. Shea, 1978: Estimates of the natural variability of time-averaged temperatures over the United States. *Mon. Wea. Rev.,* **106,** 1695–1703, doi:10.1175/1520-0493(1978)106<1695:EOTNVO>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.

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.

Rodwell, M. J., and F. J. Doblas-Reyes, 2006: Medium-range, monthly, and seasonal prediction for Europe and the use of forecast information. *J. Climate,* **19,** 6025–6046, doi:10.1175/JCLI3944.1.

Saha, S., and Coauthors, 2006: The NCEP Climate Forecast System. *J. Climate,* **19,** 3483–3517, doi:10.1175/JCLI3812.1.

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

Scaife, A. A., T. Woollings, J. R. Knight, G. Martin, and T. Hinton, 2010: Atmospheric blocking and mean biases in climate models. *J. Climate,* **23,** 6143–6152, doi:10.1175/2010JCLI3728.1.

Smith, D. M., and Coauthors, 2013: Real-time multi-model decadal climate predictions. *Climate Dyn.,* **41,** 2875–2888, doi:10.1007/s00382-012-1600-0.

Smith, T. M., and R. E. Livezey, 1999: GCM systematic error correction and specification of the seasonal mean Pacific–North America region atmosphere from global SSTs. *J. Climate,* **12,** 273–288, doi:10.1175/1520-0442-12.1.273.

Van den Dool, H. M., 2007: *Empirical Methods in Short-Term Climate Prediction.* Oxford University Press, 215 pp.

——, 2009: Methods of multi-model consolidation, with emphasis on the recommended three-year-out cross validation approach. *Extended Abstracts, NOAA CTB Joint Seminar Series,* Camp Springs, MD, NOAA, 75–77. [Available online at http://www.nws.noaa.gov/ost/climate/STIP/fy09jsctb.htm.]

——, and R. M. Chervin, 1986: A comparison of month-to-month persistence of anomalies in a general circulation model and in the earth's atmosphere. *J. Atmos. Sci.,* **43,** 1454–1466, doi:10.1175/1520-0469(1986)043<1454:ACOMTM>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.]

Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction.* Academic Press, 467 pp.

Wu, R., B. P. Kirtman, and H. M. Van den Dool, 2009: An analysis of ENSO prediction skill in the CFS retrospective forecasts. *J. Climate,* **22,** 1801–1818, doi:10.1175/2008JCLI2565.1.

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.