Southeastern U.S. Rainfall Prediction in the North American Multi-Model Ensemble

JOHANNA M. INFANTI AND BEN P. KIRTMAN

Department of Meteorology and Physical Oceanography, Rosenstiel School of Marine and Atmospheric Science, University of Miami, Miami, Florida

(Manuscript received 30 April 2013, in final form 5 September 2013)

ABSTRACT

The present study investigates the predictive skill of the North American Multi-Model Ensemble (NMME) system for intraseasonal-to-interannual (ISI) prediction with focus on southeastern U.S. precipitation. The southeastern United States is of particular interest because of the typically short-lived nature of above- and below-normal extended rainfall events allowing for focus on seasonal prediction, as well as the tendency for more predictability in the winter months. Included in this study is analysis of the forecast quality of the NMME system when predicting above- and below-normal rainfall and individual rainfall events, with particular emphasis on results from the 2007 dry period. Both deterministic and probabilistic measures of skill are utilized in order to gain a more complete understanding of how accurately the system predicts precipitation at both short and long lead times and to investigate the multimodel aspect of the system as compared to using an individual predictive model. The NMME system consistently shows low systematic error and relatively high skill in predicting precipitation, particularly in winter months as compared to individual model results.

1. Introduction

Precipitation variability throughout the United States is a subject of recent research focus as extended above- or below-normal rainfall events have the potential to incur significant economic and environmental effects on a given region. Thus, accurate prediction of precipitation variability is extremely important. Here we focus on southeastern U.S. precipitation, which is a challenging and interesting region from a prediction standpoint. Extended dry and wet periods are common in this region, but these events are shorter lived when compared to other U.S. regions, though they are still damaging economically and environmentally (Manuel 2008; Seager et al. 2009). More generally, extreme dry events or drought are one of the costliest natural disasters, and extreme wet events can cause flooding, mudslides, etc., affecting both human welfare and economic resources (Pielke and Downton 2000; Andreadis et al. 2005). Given the potential for economic and environmental damage, it is important to understand the current prediction capability within the region.

It is now fairly well understood that a multimodel approach to prediction is an imperfect but still pragmatic method to estimating forecast uncertainty (Krishnamurti et al. 1999, 2000; Doblas-Reyes et al. 2000; Palmer et al. 2004; Hagedorn et al. 2005; Weigel et al. 2008; Kirtman and Min 2009). In this paper we utilize phase-1 data from the North American Multi-Model Ensemble (NMME) system, a newly formed multi-institutional, multimodel ensemble system for intraseasonal-to-interannual (ISI) prediction, which includes models from nine institutional partners (Kirtman et al. 2014). The choice of this particular set of models is motivated by the availability of phase-1 data and the potential for inclusion of additional fields and improvements to be made with later phases. The NMME system is used in real time by the National Oceanic and Atmospheric Administration (NOAA)/Climate Prediction Center (CPC) for seasonal drought prediction in the United States (www.cpc.ncep.noaa.gov/products/expert_assessment/sdo_summary.html) as well as on an experimental basis (Kirtman et al. 2014), and the expectation is that the system will continue to be used in the future as more fields and improvements are made. Here we intend to examine how the multimodel ensemble compares to individual model(s) in terms of forecast skill.

The southeast U.S. lends itself particularly well to experiments concerning seasonal prediction because of the
nature of events in the region, but the Southeast has other properties that offer an interesting problem in the context of prediction. The evidence of a linkage with ENSO has been used to test the predictive skill within various models as the Southeast tends to show more predictability in winter months because of a stronger connection with ENSO (e.g., Cocke et al. 2007). Mo and Schemm (2008a) showed that the relationship between ENSO and the Southeast is seasonally dependent where cold (warm) ENSO favors dryness (wetness) in the winter, but the opposite is true during summer seasons. This highlights the seasonality of dry spells and rainfall in general over the Southeast and thus has motivated our choice in considering seasonal prediction capabilities within the NMME system.

Previous studies have examined the link to sea surface temperature anomalies (SSTAs) in the tropical Pacific or ENSO, which typically shows the strongest link in winter months (Ropelewski and Halpert 1986, 1987; Mo and Schemm 2008a,b; Seager et al. 2009). In contrast to Mo and Schemm (2008a), Seager et al. (2009) concluded that rainfall is more closely related to internal atmospheric variability, particularly in summer. Noise or internal atmospheric variability, for the purposes of this paper, is unpredictable, that is, we cannot relate it to specific forcing or feedback (ocean–atmosphere or atmosphere–land). This unpredictable variability is due to internal atmospheric dynamics and may be large relative to a forced signal leading to difficulties from a predictive standpoint (Schubert et al. 2009; Seager et al. 2009). This is not necessarily a phenomenological definition; however, we recognize that specific phenomena (sea breezes, thunderstorms, etc.) are important to prediction in this region and cannot be overlooked, though these phenomena are difficult to resolve without a high-resolution model (Stefanova et al. 2012a).

Additionally, for summertime precipitation, results from L. Li et al. (2011) and W. Li et al. (2011) show that the North Atlantic subtropical high (NASH) has affected precipitation variation in the southeast, and that summer precipitation variability has been enhanced in the past few decades due in part to the NASH. The Atlantic multidecadal oscillation (AMO) may also affect summertime precipitation, leading to increased summer rainfall during warm phases of the AMO (Hu et al. 2011). There is also evidence that tropical storms may affect precipitation over the Southeast during the summer months (Mo and Schemm 2008a). Though there may exist some controversy in the literature about the amount of possible predictive skill this region may have, particularly in summer, there has been some evidence that there may be hindcast skill when considering winter season precipitation in the Southeast in global climate models (GCMs), regardless of the type of SST forcing considered within each model (freely evolving SST versus prescribed; Stefanova et al. 2012b).

There is a need for accurate predictions within the region given the potential for damaging events, though internal atmospheric variability may overwhelm any forced signal, particularly in summer months (Seager et al. 2009). While we accept this as a potential hindrance and expect that our results will show higher prediction skill in winter seasons versus summer seasons, we hope that through this analysis we will gain a better understanding of what is truly achievable operationally in the context of seasonal prediction, as well as offering an examination of whether the multimodel aspect offers better skill than when considering individual models by themselves.

We offer an assessment of the skill of the NMME system when considering above- and below-normal precipitation in the southeastern United States. We use both deterministic and probabilistic methods, including systematic error measure, spatial RMSE, variance evolution (signal and internal variance), anomaly correlation, and relative operating characteristic analysis, as well as a case study analysis of the skill of the NMME system in hindcasting seasonal precipitation during the 2006–07 drought. In many cases, we compare the results of the NMME ensemble mean system to that of the individual models or randomly chosen ensemble members in order to determine if the multimodel aspect offers superior skill (e.g., Peng et al. 2002; Hagedorn et al. 2005). We focus on this problem from a seasonal standpoint given the nature of events within the region. Given the above information, we would expect to see more skill in the winter months and less skill in summer months.

Our results show that the NMME system shows potential in predicting southeastern U.S. prediction anomalies, mainly at short lead times in winter months, and offers more consistent results than when considering an individual prediction model. Individual models tend to show varying results, whereas NMME tends to show results near to or surpassing “best” model results. This result is similar to Kirtman and Min (2009), though they focused on multimodel ENSO prediction. Seasonality of the events in the region seems to be of particular importance because of the typical length of continually above- and below-normal events, and skill varies seasonally. When considering individual events, NMME has some difficulties at longer lead times, which is not necessarily the case if we consider the overall skill analysis. We notice that during periods of incorrect hindcasts, the NMME system also tends to incorrectly forecast the tropical Pacific SST, though additional
analysis needs to be carried out to determine the physical reasoning for incorrect hindcasts.

The organization of the text is as follows. Section 2 describes the datasets and general methodology used in our analysis. Section 3 contains an assessment of the mean error of the NMME system. Section 4 contains deterministic and probabilistic hindcast verification of the NMME system. Section 5 provides hindcast verification of an individual event. Section 6 provides discussion and conclusions.

2. Datasets and methodology

a. Observations

The CPC Merged Analysis of Precipitation (CMAP) is used for analysis of observational rainfall (Xie and Arkin 1997). CMAP precipitation data include monthly averaged precipitation rate from satellite estimates in millimeters per day on a 1.0° latitude × 1.0° longitude grid and have been provided by the International Research Institute/Lamont-Doherty Earth Observatory (IRI/LDEO) collection of climate data for verification of the NMME system. Data are available from January 1982 to October 2010, and we calculate anomalies (monthly precipitation minus climatology for that month) based on this reference period, applying a 3-month sliding average in time to create seasonal anomalies unless otherwise noted. We recognize that all datasets have biases, and precipitation biases may exist over land and ocean regions for different types of precipitation; however, this dataset shows good agreement with other observational datasets, particularly over land (Xie and Arkin 1997). Though we acknowledge that winter and summer Southeast precipitation have differences, we consider overall seasonal precipitation in this analysis and the blended nature of the CMAP dataset minimizes biases related to precipitation type. Observed sea surface temperature is from the National Climate Data Center (NCDC) Optimum Interpolation Monthly Sea Surface Temperature Analysis (OISST; Reynolds et al. 2002) in degrees Celsius on a 1.0° latitude × 1.0° longitude grid, also provided by IRI/LDEO for verification of the NMME system. Data are available from January 1982 to December 2010, with anomalies calculated similarly to observed precipitation.

b. NMME

Multimodel ensemble data considered in this analysis consist of the NMME system for ISI prediction. The NMME system is a multi-institutional intraseasonal-to-interannual climate prediction tool that provides real-time forecasts that adhere to the CPC operational schedule as well as supplying hindcast data. Model information and references are given in Table 1, and all models are dynamical global general circulation climate prediction models. Additional information on the NMME project is given in Kirtman et al. (2014). In this analysis we have considered precipitation and SST data from phase 1 (NMME-1) of the NMME project. Phase 2 (NMME-2) will bring additional fields and lead times as well added models and participation. The hindcast period for NMME-1 is about 30 yr, typically from 1982 to 2010, and 2-m temperature (T2m), precipitation rate, and SST data with lead times up to 11 months have been provided. Model configuration (resolution, version, physical parameterizations, initialization strategies, ensemble generation strategies, and number of ensemble members) is left to forecast providers (Kirtman et al. 2014). The resolution of the models varies, but all have been regridded to 1.0° latitude × 1.0° longitude.

Hindcast and real-time forecast data for NMME-1 are currently available at the IRI/LDEO data archive, and the CPC has agreed to evaluate hindcasts and perform verification (www.cpc.ncep.noaa.gov/products/NMME/), which primarily serves the real-time needs of the forecast system. The IRI/LDEO has also provided analysis tools (http://iridl.ldeo.columbia.edu/home/.tippett/.NMME/.Verification/), which primarily serves the hindcast and research needs of the project.

To analyze data from the NMME system, we consider the period defined by the hindcast initialization month (i.e., January, February, March, etc.) from 1982 to 2009, as this is the longest period in which all models have available hindcast data. We define hindcast lead time to be consistent with typical seasonal forecast centers (e.g., www.wмолc.org). For example, a hindcast initialized in February verifying in the February–April (FMA) season is a short-lead (SL) hindcast (or lead 0 hindcast). A hindcast initialized in September verifying in the FMA season is a long-lead (LL) hindcast (or lead 5 hindcast).

We focus on the above hindcast periods because SL hindcasts are the closest possible verification season to initialization month and LL hindcasts are the last possible verification season that includes all ensemble members given the differences in lead times for each model. We have chosen verification seasons of November–January (NDJ), December–February (DJF), and FMA (late fall, winter, and early spring seasons) and May–July (MJJ), June–August (JJA), and August–October (ASO; late spring, summer, and early fall seasons). The choice of seasons NDJ, FMA, MJJ, and ASO was motivated by case study results of the 2006–07 southeastern U.S. drought (section 5) that showed larger anomalies in FMA and ASO, and we have chosen the surrounding seasons of NDJ and MJJ for completeness. Previous
studies indicated that there is potential for more predictability in winter seasons and less in summer (Mo and Schemm 2008a; Seager et al. 2009; Stefanova et al. 2012a,b), and we have thus included typical definitions of winter (DJF) and summer (JJA) seasons to facilitate a more complete analysis. Short-lead hindcasts verifying in NDJ, DJF, FMA, MJJ, JJA, and ASO are initialized in November, December, February, May, June, and August, respectively, and long-lead hindcasts verifying in NDJ, DJF, FMA, MJJ, JJA, and ASO are initialized in June, July, September, December, January, and March, respectively.

NMME ensemble mean anomaly is computed using a “pooling” approach in which all ensemble members from all participating models are pooled into a single sample with equal weights; thus, models with more ensemble members are effectively weighted more (see appendix). Repetition of anomaly correlation experiment with the same amount of ensemble members from each model (six) yielded similar results. Depending on the availability of lead times for each model, the number of ensemble members ranges from 109 to 36; however, since we are considering only short and long leads as defined above, we are able to use the full 109 ensemble members at both lead times. At the time of analysis for this paper, hindcasts initialized in June and July were not available in the National Aeronautics and Space Administration (NASA) Global Modeling and Assimilation Office (GMAO) hindcasts; thus, this model was not considered in the full NMME ensemble mean for these two hindcast periods (missing hindcasts are reflected in ensemble mean). If considering an individual model, we consider only the ensemble members available for that model (e.g., six ensemble members for CCSM3). Anomalies are defined similarly to observations for each initialization month, and the ensemble mean of the NMME system or individual models acts to represent the most probable outcome.

3. Mean error

We begin this analysis by showing the agreement between NMME and observations at short and long leads using a measurement of systematic error (observed precipitation minus NMME precipitation; see appendix) in Fig. 1. The region of interest [30°–38°N, 268°–285°E, similar to that used in Seager et al. (2009)] is shown outlined in black in the top left panel of Fig. 1. We find that the NMME system shows overall low systematic error in the southeastern United States in winter

<table>
<thead>
<tr>
<th>Model</th>
<th>Hindcast period</th>
<th>Ensemble size</th>
<th>Lead times</th>
<th>Arrangement of ensemble members</th>
<th>Contact and reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Climate Forecast System, version 1 (CFSv1)</td>
<td>1981–2009</td>
<td>15</td>
<td>0.5–8.5 months</td>
<td>First, twenty-first, and eleventh days of the month, all at 0000 UTC ±2 days</td>
<td>Saha (Saha et al. 2006)</td>
</tr>
<tr>
<td>Climate Forecast System, version 2 (CFSv2)</td>
<td>1982–2010</td>
<td>24 (28)</td>
<td>0.5–9.5 months</td>
<td>Four members (0000, 0600, 1200, 1800 UTC) every fifth day</td>
<td>Saha (Saha et al. 2010)</td>
</tr>
<tr>
<td>Geophysical Fluid Dynamics Laboratory Climate Model, version 2.2 (GFDL CM2.2)</td>
<td>1982–2010</td>
<td>10</td>
<td>0.5–11.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>Rosati (Zhang et al. 2007)</td>
</tr>
<tr>
<td>IRI ECHAM4-f</td>
<td>1982–2010</td>
<td>12</td>
<td>0.5–7.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>DeWitt (DeWitt 2005)</td>
</tr>
<tr>
<td>IRI ECHAM4-a2</td>
<td>1982–2010</td>
<td>12</td>
<td>0.5–7.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>DeWitt (DeWitt 2005)</td>
</tr>
<tr>
<td>Community Climate System Model, version 3 (CCSM3)</td>
<td>1982–2010</td>
<td>6</td>
<td>0.5–11.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>Kirtman (Kirtman and Min 2009)</td>
</tr>
<tr>
<td>Goddard Earth Observing System Model, version 5 (GEOS-5)</td>
<td>1981–2010</td>
<td>11</td>
<td>0.5–9.5 months</td>
<td>One member every fifth day</td>
<td>Schubert (Vernieres et al. 2012)</td>
</tr>
<tr>
<td>Canadian Meteorological Centre 1–Third Generation Canadian Coupled Global Climate Model (CMC1-CanCM3)</td>
<td>1981–2010</td>
<td>10</td>
<td>0.5–11.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>Merryfield (Merryfield et al. 2013)</td>
</tr>
<tr>
<td>Canadian Meteorological Centre 2–Fourth Generation Canadian Coupled Global Climate Model (CMC2-CanCM4)</td>
<td>1981–2010</td>
<td>10</td>
<td>0.5–11.5 months</td>
<td>All first of the month at 0000 UTC</td>
<td>Merryfield (Merryfield et al. 2013)</td>
</tr>
</tbody>
</table>
months (NDJ, DJF, and FMA) when compared to, for instance, the northwestern United States, and the error increases only slightly (if at all) in summer months (MJJ, JJA, and ASO). NMME also shows low error when compared to individual model results (Fig. 2). The NMME ensemble mean shows similar magnitude and structure when considering short versus long leads in each season, and as such there is very little change in systematic error as lead time increases; thus, any error structure sets in quickly. Though the error is very similar across lead times, it is not exactly the same. We cannot fully blame degradation of forecast skill in the Southeast with lead time on an increase in systematic error.

Figure 2 shows a similar calculation, but for the individual model within the NMME system with the highest (lowest) systematic error. Calculation is based on the RMSE (see appendix) in an area-averaged southeastern U.S. region of 30°–38°N, 268°–285°E (see black outline in Figs. 1 and 2) and the model with the highest (lowest) systematic error for each season and lead time (see Table 2 for RMSE values). It becomes immediately apparent that the individual model with the highest (lowest) systematic error varies with season. When considering the NMME system as a whole (i.e., Fig. 1), the NMME systematic error is closest to that shown by the individual model with the smallest systematic error. The comparatively low systematic error in the NMME ensemble mean indicates that there is large error cancellation. This result can be most clearly seen in Table 2 when comparing the RMSE of individual models to the NMME ensemble mean RMSE. NMME has RMSE typically close to or lower than the individual model with the lowest RMSE.

The observed seasonal precipitation variance over the United States for all seasons (Fig. 3a) versus the NMME internal variance at short and long leads (Fig. 3b) and the NMME signal variance at short and long leads (Fig. 3c) is shown in Fig. 3 (see appendix for calculation of internal variance versus signal variance for the NMME system). Variance measures are similar for NDJ, DJF, and FMA and for MJJ, JJA, and ASO, and we only show FMA and ASO as these seasons relate to case study results presented in section 5. Observations (Fig. 3a) show that there is relatively large variance in the Southeast in both seasons when compared to other U.S. regions, with the most variance in ASO. NMME internal variance (Fig. 3b) is calculated similarly to Stefanova et al. (2012b) and is large when compared to the signal variance (Fig. 3c). Figure 3b shows that there is a large amount of variance internal to the NMME system regardless of season. The NMME internal variance is typically too large when compared to observed variance, but we note much of the
variance is located over the oceans. Discussion of the larger variance measures over the oceans is beyond the scope of this paper. The signal variance (Fig. 3c) is small in comparison, but we see the most signal variance in FMA. When considering the differences between short and long lead times, we find that the internal variance remains relatively constant with lead time, but the signal variance tends to diminish with lead time in FMA.

To further assess the variance measure in the southeastern United States specifically, we have plotted the

Fig. 2. Systematic error for individual NMME model (calculation similar to Fig. 1, but for individual model mean). Emphasis is on the individual model with the lowest systematic error vs the model with the highest systematic error. The choice of the model for lowest and highest error based on RMSE values is in Table 2. Lowest (highest) errors are found in columns 1 and 2 (3 and 4). Hindcasts verify in NDJ, DJF, FMA, MJJ, JJA, and ASO at (columns 1 and 3) short and (columns 2 and 4) long leads. Black rectangle in top left panel shows region of interest.
variance evolution of observations and a randomly chosen ensemble member from each model, as well as the NMME signal and total variance at each of the given hindcast initialization months in Fig. 4. We show forecast initializations of December, February, June, and August (similar results found for remaining initialization months). All values have been area averaged monthly in the Southeast region. Lead time for each initialization month is plotted along the $x$ axis and is used synonymously with the verification month. Individual model realizations appear to have variance that is similar (but typically larger) in amplitude to observations in the Southeast (gray lines in Fig. 4), while the signal variance (blue line in Fig. 4) is drastically decreased in comparison. Total variance (green line in Fig. 4) is close to observed variance with a tendency to slightly overestimate. The overestimation of the total variance suggests that the models overestimate noise. We find that individual model realizations have considerable variation among ensemble member variance measures. We also note that the signal variance is too small compared to the total variance, which highlights the difficulties in making skillful predictions.

4. Deterministic and probabilistic skill assessment

This section provides both deterministic and probabilistic skill assessment of the NMME system. We utilize both methods of assessment as there is evidence that the two approaches together provide a more complete representation of skill (Kirtman 2003). The metrics we consider for deterministic analysis include evolution of range of ensemble members about the ensemble mean and anomaly correlation. For probabilistic assessment, we use relative operating characteristic (ROC) analysis.

a. Deterministic skill

We first examine the hindcast skill from a deterministic standpoint in both the full NMME system as well as individual models. We consider the hindcast systematic evolution for the NMME ensemble mean and individual models for February and September hindcast initialization periods (similar results are found for the remaining initialization months) and anomaly correlation for all verification seasons.

The systematic or climatological evolution of hindcasts for February and September hindcast initializations is shown in Figs. 5a and 5b, respectively. We have
FIG. 3. (a) Observed precipitation variance (mm$^2$ day$^{-2}$) vs (b) NMME total precipitation variance and (c) NMME signal precipitation variance. Color scale is the same for all panels. Observations in (a) are plotted for (left) FMA and (right) ASO; (b),(c) are plotted verifying in (left) FMA and (right) ASO for short (long) leads is in the top (bottom) panels.
plotted each individual model as well as the full NMME ensemble mean for the area-averaged Southeast region (monthly values). This figure shows the systematic hindcast evolution of the ensemble mean (blue line for each model and NMME ensemble mean), observed climatology (red line), and the range of ensemble members about the ensemble mean (gray lines). Lead time is plotted on the $x$ axis similarly to Fig. 4. Some of the error noted earlier can also be detected in Figs. 5a and 5b, but we can easily see that the individual models are quite variable. For example, for February hindcast initialization (Fig. 5a), the NMME ensemble mean has some differences with observations, but overall tracks well compared to any individual model. Model mean 6 (Mod6), for example, does not follow observations well. The range of ensemble members about the ensemble mean for individual models is small (as compared to the full NMME system) and stays constant as lead time increases. The range of ensemble members about the NMME ensemble mean is, perhaps expectedly, much larger than the range for individual models. This indicates that model formulation differences, not initial condition differences, lead to a larger spread.

We have plotted the anomaly correlation between NMME and CMAP precipitation (Fig. 6, shading) for all seasons at short and long leads. To test the significance
of this correlation, we convert the anomaly correlation coefficient to a $Z$ score using Fisher’s $R$-to-$Z$ transformation (see appendix) and perform a Student’s $t$ test for significance on the transformed values [based on a one-tailed Student’s $t$ test with $N = 28$ seasons and degrees of freedom (df) = 26]. The 95% and 99% significance is contoured based on this method; however, as a general rule, an anomaly correlation of 0.6 is regarded as skillful (Wilks 2011). The winter months (NDJ, DJF, and FMA; left two columns) show significant correlation throughout both lead times in the coastal regions. Winter months have mostly constant correlation as lead time increases. Summer seasons (MJJ, JJA, and ASO; right two columns) show weaker correlation overall and no significance, with larger degradation of anomaly correlation with lead time. The weakening correlation in the summer seasons may be due to the spring prediction barrier in the tropical Pacific (e.g., Zheng and Zhu 2010), as these hindcasts would be initialized near spring, weaker forcing from the tropical Pacific (i.e., weaker signal), or it may be due to the spatially small forcing or internal atmospheric variability in summer, as mentioned in the introduction.

We are also interested in the difference between anomaly correlations when considering the NMME ensemble mean versus individual models. We again use Fisher’s $R$-to-$Z$ transformation to convert the (spatially averaged in the Southeast) anomaly correlation of both the NMME ensemble mean and individual model means into $Z$ scores, and then we take the difference to assess the significance of the difference between correlations (see appendix). Table 3 shows this score for the difference between NMME and each individual model, and results are significant for values greater than 0.856 for ($P < 0.20$; based on a one-tailed Student’s $t$ test with df = 26). Positive values indicate that the NMME ensemble
mean anomaly correlation is higher than the individual model, and we note that for most seasons and lead times, the values in the table are positive and sometimes significantly higher. The bold text with two asterisks indicates the largest increase in correlation for the NMME ensemble mean. The NMME system shows the highest increase in correlation in DJF at short and long leads and the smallest in the summer seasons.

b. Probabilistic skill assessment

Probabilistic skill assessment offers additional information that can be used in conjunction with deterministic analysis, with the added advantage of allowing analysis of above- and below-normal events separately. We consider relative operating characteristic analysis, which allows us to assign a score and metric to the ability of the NMME system in predicting above- and below-normal events. The World Meteorological Organization (WMO) standardized verification system (SVS) for long-range forecasts (see www.metoffice.gov.uk/media/pdf/j/6/SVSLRF.pdf) recommends this method as a summary statistic representing the skill of the forecast system. In our analysis, we follow the guidelines outlined in the WMO SVS for probabilistic forecasts in order to create the ROC curves. Above- and below-normal events are defined based on percent or fraction of ensemble members out of the total based on Gaussian fitting estimation of tercile-based categories (Kirtman 2003; Kirtman and Min 2009; Wilks 2011). Lower (x_{bn}) and upper (x_{an}) tercile boundaries are defined similarly to Min et al. (2009):

$$x_{bn} = -0.43\sigma_n$$ and $$x_{an} = +0.43\sigma_n$$, (1)

where $\sigma_n$ is the standard deviation of each individual ensemble member $n$ and $x_{bn}$ and $x_{an}$ (below and above normal) are defined for each ensemble member $n$. These are then converted into binary events. If an ensemble member falls into either category, it is counted as a 1; otherwise, it is a 0. We then sum over all ensemble members in each month to compute the number of ensemble members falling into each category, and divide by the total number of ensemble members (109) to give the percentage or fraction.

To compute the ROC curve, we consider southeastern U.S. precipitation for each individual month and grid point and compute the percent of ensemble members falling into each category as demonstrated above. A “warning” is issued when the forecast probability exceeds a certain threshold (say, at least 80% confidence or 80% of ensemble members predict above-normal conditions), and these are used to define hit rates and false alarm rates [discussion of thresholds is found in the in relevant section of Mason and Graham (1999), and equations are found in WMO SVS]. These values are then aggregated seasonally within the region.

The analysis technique is defined in Mason and Graham (1999) and Kirtman (2003). An ideal system would show large hit rates and small false alarm rates, with the points on the curve clustering in the upper left of the diagram. For a poor forecast system, the points of the ROC curve would lie close to or on the diagonal. We have also calculated the ROC score (area under the curve) using the trapezoidal rule, allowing us to assign a numerical value to the ROC curve. A perfect forecast system has an area of 1, and a forecast system with no useful information lies on the diagonal with an area of 0.5. Additional information and analysis methods can be found in (Wilks 2011).

ROC curves for above- and below-normal tercile events at short and long leads for hindcasts verifying in FMA and ASO are shown in Fig. 7. We have chosen to show these two seasons as they relate to the case study results in section 5, and we found similar results for other seasons. We show ROC scores in Table 4 for all seasons and lead times for the model mean with the highest ROC score, lowest ROC score, and NMME. We find that the skill of the NMME system and individual models diminishes when considering ASO versus FMA at short lead times, and the most skill is shown in FMA at short lead, above-normal events. Below-normal events at the same season and lead time are similar, but higher false alarm rates are found for this season. The skill diminishes as lead time increases in both seasons. When the NMME ROC curve is well separated from the diagonal, the NMME ensemble mean indicates largest skill, and when it is not well separated, the NMME ensemble mean indicates lowest skill. There is some variation among models overall, but typically the NMME ensemble system shows better skill in comparison to individual models. When considering the ROC scores (Table 4), the NMME system does not always show the highest ROC score, but typically shows a score near the top throughout all seasons. The exception to this is JJA at a long lead for below-normal events, where NMME ensemble mean shows the lowest score.

5. Hindcasting individual events

We have chosen to focus on the 2006–07 dry period in order to examine the skill of the NMME system in hindcasting individual events. This period involved below-normal rainfall (phase 1), followed by a brief reprieve (phase 2) before a stronger period of below-normal rainfall (phase 3), thus highlighting the need for seasonality within an event. This period was also studied by Seager et al. (2009), who found that the dry period was not
accurately represented by a Global Ocean Global Atmosphere (GOGA) model using 6-month averages, and the model failed to create a continuous drought. We believe that it is important to examine the seasonality of this drought as it highlights some of the prediction challenges.

We have plotted southeastern U.S. precipitation plumes for the 2006–07 dry period in Figs. 8a (February starts) and 8b (September starts). Figure 8a shows hindcasts initialized in (left to right) February 2005, 2006, and 2007, with lead time increasing along the x axis. Figure 8b shows hindcasts initialized in (left to right) September 2005, 2006, and 2007. The blue line indicates the NMME ensemble mean anomaly, the red line indicates observed precipitation anomaly, and the gray lines indicate individual ensemble member anomalies. All anomalies are standardized (divided by standard deviation). Blue
shading refers to phases 1, 2, and 3 of this drought. Anomalies are 3-month running means. It becomes apparent that observations show a brief below-normal period in early 2006 (phase 1), followed by a slight reprieve in mid-2006 (phase 2), then a stronger below-normal period in early 2007 (phase 3).

The plumes in Figs. 8a and 8b show constant range of ensemble members throughout all lead times and during the two hindcast initialization months; thus, the uncertainty at all lead times remains constant. We find good agreement between ensemble mean and observations during the first phase of the dry period when considering February starts (Fig. 8a, middle), and the ensemble mean stays neutral or weakly dry through phases 2 and 3. When looking at September hindcast initialization (Fig. 8b), there is less agreement between
observations and the ensemble mean overall. September hindcast initialization shows some drying during phase 1 of the drought (Fig. 8b, left) and captures the precipitation increase during phase 2 (Fig. 8b, middle), but remains wet throughout phase 3. There are ensemble members in all phases that appear close to observed, and the observations were not unpredictable by all ensemble members.

We focus on the season closest to the observed driest or wettest season during each phase. This corresponds to FMA2006 (phase 1), ASO2006 (phase 2), and FMA2007 (phase 3). Figure 9 shows the area-averaged precipitation from all ensemble members verifying in the above seasons at short and long leads versus observed precipitation. The precipitation anomaly in millimeters per day is given on the y-axis. On the x-axis, we have binned ensemble members, that is, bin 1 houses the first ensemble member from each model (nine ensemble members, one from each model), bin 2 houses the second ensemble member from each model, and so on until bin 24, which houses the twenty-fourth ensemble member from each model (only one model has 24 ensemble members available). Note that the number of ensemble members per bin decreases with higher bin numbers because of the varying amount of ensemble members available in each model. We have masked out ensemble members showing neutral rainfall and thus only focus on rainfall in the upper and lower terciles. The number of ensemble members out of 109 falling into upper and lower categories is noted in the bottom right of each panel. The red dotted line indicates the value of the observed precipitation anomaly.

We find the best agreement during a short lead verifying in FMA2006 (Fig. 9, top left), with less agreement as lead time increases. The worst agreement occurs at a long lead verifying in FMA2007 (Fig. 9, bottom right), in which the NMME system shows predominantly wet anomalies in the Southeast region. We also find that the ensemble members that are able to capture below-normal rainfall during FMA2007 at a long lead do not capture the magnitude of the observed precipitation anomaly very well, with only a few exceptions. The majority of ensemble members show below-normal rainfall in ASO2006 at both lead times (Fig. 9, middle panels), though this period is only weakly above normal in observations, and its possible the NMME system is not capturing this accurately as the observed anomaly is small.

We have plotted the NMME ensemble mean anomaly (most probable outcome) for FMA2006, ASO2006, and FMA2007 at short and long leads versus observations in Figs. 10a, 10b, and 10c, respectively. Figure 10a (FMA2006) shows similar results to previous figures, with good

| Table 3. Individual model Z-score difference (defined as the difference between NMME anomaly correlation and individual model anomaly correlation) for a given season and lead time in the area-averaged Southeast region: (a) SL and (b) LL. Bold text highlights the model with the highest and lowest Z-score difference. Bold with one asterisk indicates the model with the lowest Z difference, and bold with two asterisks indicates the model with the highest Z difference. Results are significant for values > 0.856 for (P < 0.20) (based on a one-tailed Student’s t test with df = 26).

<table>
<thead>
<tr>
<th></th>
<th>SL</th>
<th>DJF</th>
<th>FMA</th>
<th>MJJ</th>
<th>JJA</th>
<th>ASO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.239</td>
<td>0.799</td>
<td><strong>1.279</strong></td>
<td>0.334</td>
<td>0.362</td>
<td>−0.560</td>
</tr>
<tr>
<td>Model 2</td>
<td>−<strong>0.007</strong></td>
<td>0.903</td>
<td>0.937</td>
<td>0.743</td>
<td>0.709</td>
<td>0.062</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.938</td>
<td>0.591</td>
<td>0.624</td>
<td>−0.141</td>
<td>−0.155</td>
<td>−0.130</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.715</td>
<td>0.523</td>
<td>0.118</td>
<td>0.436</td>
<td>0.613</td>
<td>0.146</td>
</tr>
<tr>
<td>Model 5</td>
<td>0.083</td>
<td>0.757</td>
<td>−0.371</td>
<td>0.921</td>
<td>0.054</td>
<td><strong>0.788</strong></td>
</tr>
<tr>
<td>Model 6</td>
<td>0.872</td>
<td><strong>1.741</strong></td>
<td>0.537</td>
<td>0.548</td>
<td>N/A</td>
<td>−<strong>0.703</strong></td>
</tr>
<tr>
<td>Model 7</td>
<td><strong>1.090</strong></td>
<td>0.121</td>
<td>0.604</td>
<td>0.224</td>
<td>0.016</td>
<td>−0.355</td>
</tr>
<tr>
<td>Model 8</td>
<td>0.127</td>
<td>−<strong>0.281</strong></td>
<td>−<strong>0.497</strong></td>
<td>−<strong>0.232</strong></td>
<td>−<strong>0.409</strong></td>
<td>−0.135</td>
</tr>
<tr>
<td>Model 9</td>
<td>0.793</td>
<td>0.480</td>
<td>0.851</td>
<td><strong>0.938</strong></td>
<td><strong>1.030</strong></td>
<td>0.359</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>LL</th>
<th>DJF</th>
<th>FMA</th>
<th>MJJ</th>
<th>JJA</th>
<th>ASO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model 1</td>
<td>0.935</td>
<td>0.935</td>
<td>0.470</td>
<td>−0.101</td>
<td><strong>0.454</strong></td>
<td>−0.224</td>
</tr>
<tr>
<td>Model 2</td>
<td>−<strong>0.024</strong></td>
<td>0.541</td>
<td>0.346</td>
<td>−0.013</td>
<td>−0.257</td>
<td>−0.050</td>
</tr>
<tr>
<td>Model 3</td>
<td>0.733</td>
<td>0.566</td>
<td>−0.182</td>
<td>−0.198</td>
<td>−0.518</td>
<td>0.100</td>
</tr>
<tr>
<td>Model 4</td>
<td>0.912</td>
<td>0.250</td>
<td>0.525</td>
<td><strong>0.371</strong></td>
<td>−0.387</td>
<td>−<strong>1.079</strong></td>
</tr>
<tr>
<td>Model 5</td>
<td>0.813</td>
<td>0.306</td>
<td>0.594</td>
<td>−0.224</td>
<td>0.432</td>
<td>−0.262</td>
</tr>
<tr>
<td>Model 6</td>
<td>N/A</td>
<td>N/A</td>
<td>0.957</td>
<td>−<strong>1.189</strong></td>
<td>−0.328</td>
<td>0.847</td>
</tr>
<tr>
<td>Model 7</td>
<td>0.255</td>
<td>−<strong>0.130</strong></td>
<td>−<strong>1.017</strong></td>
<td>−0.118</td>
<td>−<strong>0.965</strong></td>
<td><strong>1.066</strong></td>
</tr>
<tr>
<td>Model 8</td>
<td>0.720</td>
<td><strong>1.602</strong></td>
<td><strong>1.393</strong></td>
<td>−0.051</td>
<td>−0.519</td>
<td>−0.756</td>
</tr>
<tr>
<td>Model 9</td>
<td><strong>0.984</strong></td>
<td>0.011</td>
<td>0.251</td>
<td>0.254</td>
<td>0.244</td>
<td>0.323</td>
</tr>
</tbody>
</table>
FIG. 7. ROC analysis for Southeast precipitation in the NMME system. Hindcast verification seasons are FMA and ASO with (left) short and (right) long leads. Top four panels show FMA hindcast verification for (top) above- and (bottom) below-normal precipitation. Bottom four panels show ASO hindcast verification for (top) above- and (bottom) below-normal precipitation. Blue line indicates NMME ensemble mean ROC curve, black dashed lines indicate individual model mean ROC curve, and black dotted line is the diagonal shown for clarity.
Table 4. ROC scores (calculated using the trapezoid rule) for individual model means and NMME full ensemble. The model with the highest and lowest ROC score is given. Table 4a shows SL for above-normal rainfall; Table 4b shows LL for above-normal rainfall; Table 4c shows SL for below-normal rainfall, and Table 4d shows LL for below-normal rainfall.

<table>
<thead>
<tr>
<th></th>
<th>SL above</th>
<th>LL above</th>
<th>SL below</th>
<th>LL below</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>NDJ</td>
<td>DJF</td>
<td>FMA</td>
<td>MJJ</td>
</tr>
<tr>
<td>High</td>
<td>0.64</td>
<td>0.663</td>
<td>0.66</td>
<td>0.553</td>
</tr>
<tr>
<td>Low</td>
<td>0.569</td>
<td>0.589</td>
<td>0.589</td>
<td>0.509</td>
</tr>
<tr>
<td>NMME</td>
<td>0.653</td>
<td>0.675</td>
<td>0.664</td>
<td>0.555</td>
</tr>
</tbody>
</table>

a)

b)

c)

d)

We find that during FMA2006, ASO2006, and FMA2007 (Figs. 11a–c) there is good agreement during the short-lead periods in the tropical Pacific, but during the long-lead hindcast, there are neutral SSTAs in FMA2006 and warm SSTAs in FMA2007; thus, the SSTA hindcast during these two time periods was incorrect. It is not apparent that the incorrect SSTA was causing the incorrect precipitation forecast or was in fact coincidental. Additional research must be carried out to understand the complete reasoning behind the incorrect hindcast, which can be due, for example, resolution of phenomena (e.g., Stefanova et al. 2012a), overall model skill or error, or parameterization of relevant physical processes, because prediction does not rely exclusively on these teleconnections.

6. Discussion and conclusions

A multimodel ensemble approach to precipitation prediction within the Southeast region provides low systematic error and high anomaly correlation, particularly in winter months. In particular, the NMME ensemble mean shows systematic error that is typically close to or surpassing the individual model with the lowest systematic error (e.g., Table 2, Fig. 2). Total variance of the NMME system is high compared to observations when considering August, September, November, and December initializations; thus, while there is high error cancellation, the NMME ensemble mean may overestimate the atmospheric noise in this region. When considering deterministic measures of forecast skill the most significant anomaly correlation results are found in the winter months along the coastal regions of the southeast (result consistent across lead times), and the NMME ensemble mean correlation is higher than individual models in most cases (e.g., Table 3, Fig. 6). Probabilistic measurement of skill (ROC analysis) shows similar results: typically, the NMME ensemble mean shows the highest, or close to the highest, ROC scores (see Table 4) for both above- and below-normal events in the winter seasons. Though the case study analysis showed slightly discouraging results, particularly at long lead times in FMA, the overall seasonal skill analysis was quite promising for the NMME ensemble system since it is as skillful or surpasses the skill of the best model. As such, utilizing the multimodel ensemble may provide a better precipitation prediction versus using an individual model, particularly when considering short lead winter seasons.

The NMME system has relatively good forecast quality (as compared to individual models) given the results from this analysis as well as analysis presented in Kirtman et al. (2014). We would expect some predictability in the southeastern United States during winter months, with predictability lessening in the summer.
FIG. 8. NMME hindcast plumes for the period between February 2005 thru August 2008 for hindcasts initialized in (a) February and (b) September 2005. Hindcast is initialized at the beginning of each panel with lead time increasing along the x axis. Red lines show observed standardized (divided by standard deviation) precipitation anomaly, blue lines show standardized NMME ensemble mean precipitation anomaly, and gray lines indicate standardized individual ensemble member precipitation anomalies. Units are mm day$^{-1}$ per standard deviation. Vertical shading represents phases 1, 2, and 3 of the 2006–07 southeastern U.S. drought (see description in section 5).
FIG. 9. Area-averaged ensemble members vs area-averaged observed precipitation anomaly grouped by ensemble member. Only ensemble members with above- and below-normal rainfall are plotted; neutral ensemble members are masked out. Blue bars indicate precipitation anomaly for each ensemble member, binned according to ensemble member. Bin 1 is the first ensemble member from each model, bin 2 is the second ensemble member, and so on to bin 24, the twenty-fourth ensemble member from each model (only one model has 24 ensemble members available). Horizontal red dashed lines indicate the approximate observed precipitation anomaly. Units are mm day$^{-1}$ ranging from −3.0 to 3.0 on the y axis. The designation letters A and B indicate the number of ensemble members predicting above- or below-normal rainfall. For example, for FMA2006 at a short lead time, $A = 7/109$ indicates that 7/109 ensemble members predict above-normal rainfall, and $B = 66/109$ that 66/109 ensemble members predict below-normal rainfall.
because of internal atmospheric variability. NMME showed the most skill when considering winter months, with a lack of skill in summer months, as might have been expected. We also have some capability to predict short-lead winter seasons (i.e., 3-month means), and as displayed in our case study analysis and results from Mo and Schemm (2008a, 2008b), seasonality is important for this region.

Though the NMME system may have a lack of skill at longer lead times, particularly in FMA when considering the 2006–07 dry period precipitation, it is interesting that the SSTA hindcast was also poorly forecasted. It is important to remember that the tropical Pacific is not the only region that may affect precipitation in the southeast, and other studies have shown that the largest precipitation responses over the United States typically occur when the Pacific and Atlantic have anomalies of opposite signs (Schubert et al. 2009). Still other studies have included the North Atlantic Oscillation (NAO), which has been shown to have pronounced effects on temperature as well as precipitation over the eastern part of the United States during the winter season (Durkee et al. 2007).

Though we have concluded that the NMME ensemble mean allows us to forego the problems with individual model variation in skill, we are left with some unanswered questions concerning precipitation prediction in the Southeast. We find a degree of skill seasonally in the region in winter, and this is in contrast to Seager et al. (2009), who used 6-month means. The authors discuss the possibility of a stronger ENSO–southeastern U.S. link during the winter that is related to Pacific decadal oscillation (PDO) phase. There is a possibility that the period considered is during a time when there is increased predictability. In the summer months, there are a multitude of climate systems and synoptic events that may affect southeastern U.S. precipitation (see introduction), which may affect predictability. These issues beg the question of what is truly controlling southeastern U.S. precipitation in winter and summer and the degree

![Fig. 10](image_url)
of relation of these controls. Nevertheless, we see overall promising results as compared to individual models, particularly during winter season short-lead prediction, when considering prediction using the NMME system.

Acknowledgments. The authors thank the NMME program partners, the CPC and the IRI, for support and data for the NMME project. We would also like to thank 4 anonymous reviewers for comments and suggestions that improved this manuscript. The authors acknowledge support from NSF ATM0754341 and AGS1137911 and NOAA NA10OAR4320143, NA11OAR4310155 and NA12OAR4310089.

APPENDIX

Methods of Calculation

Observed climatology will be referred to as $p$, and the observed anomaly will be referred to as $p'$. Similarly for each ensemble member ($m_i$), ensemble member climatology will be referred to as $\bar{m}_i$ and ensemble member anomaly will be referred to as $m'_i$.

We first define a climatology for each month for each ensemble member, where $\bar{m}_i$ represents climatology at each ensemble member, and $\bar{M}(M')$ represents the model mean, or ensemble mean, climatology (anomaly), with $N = 6$ through 109 ensemble members depending on model. The ensemble mean hindcast for any given month or season is then

$$\bar{M} = \frac{1}{N} \sum_{i=1}^{N} \bar{m}_i.$$  

We calculate the ensemble mean anomaly by deseasonalizing each ensemble member in order to bias correct the hindcasts. The ensemble mean anomaly is then

$$M' = \frac{1}{N} \sum_{i=1}^{N} (m'_i - \bar{m}_i).$$

FIG. 11. As in Fig. 10, but for the SSTA. Color scale and contours for the observed and NMME ensemble mean SSTA ranges for observed from $-3.0^\circ$ to $3.0^\circ$C with intervals of $0.3^\circ$, $0.5^\circ$, $1.0^\circ$, $1.5^\circ$, $2.0^\circ$, $2.5^\circ$, and $3.0^\circ$C for NMME from $0.1^\circ$, $0.2^\circ$, $0.3^\circ$, $0.5^\circ$, $1.0^\circ$, $1.5^\circ$, $2.0^\circ$, $2.5^\circ$, and $3.0^\circ$C. The $y$ axes run from 90$^\circ$S to 90$^\circ$N in increments of 30$^\circ$ and the $x$ axes run from 0$^\circ$ to 180$^\circ$E to 0$^\circ$ in increments of 60$^\circ$. 

Unauthenticated | Downloaded 10/17/22 04:13 AM UTC
We can then calculate the systematic error for the NMME system and individual models as follows,

\[ \text{Error} = p - \bar{M}. \]

RMSE is calculated over region using the following equation, where \( L \) used in the summation indicates the average over the region overall (30°–38°N, 268°–285°E):

\[ \text{RMSE} = \sqrt{\frac{1}{L \sum_{i=1}^{L} (p_{L,i} - M_{L,i})^2}.} \]

Internal variance as seen in Fig. 3b is defined as

\[ \sigma_I^2 = \frac{1}{N} \sum_{i=1}^{N} (m_{i} - \bar{m}_{i})^2. \]

Signal variance is similarly defined but depends on the ensemble mean anomaly:

\[ \text{Difference} = Z_{\text{NMME}} - Z_{\text{MOD1,2...9}} \sqrt{\frac{1}{N_{\text{NMME}} - 3} + \frac{1}{N_{\text{MOD1,2...9}} - 3}}. \]

Fisher’s \( R \)-to-\( Z \) transform is calculated using the anomaly correlation from NMME or individual model (where \( R \) refers to Pearson’s correlation coefficient for the anomaly correlation between NMME ensemble mean precipitation or individual model ensemble mean precipitation and observed CMAP precipitation anomaly), and \( Z \)-score difference follows from this transformation:

\[ Z'_{\text{NMME}} = 0.5 \ln \frac{(1 + R_{\text{NMME}})}{(1 - R_{\text{NMME}})}; \]

\[ Z'_{\text{MOD1,2...9}} = 0.5 \ln \frac{(1 + R_{\text{MOD1,2...9}})}{(1 - R_{\text{MOD1,2...9}})} \]

and

REFERENCES


