Clim Dyn
DOI 10.1007/s00382-010-0988-7

Hindcast skill and predictability for precipitation and two-meter air temperature anomalies in global circulation models over the Southeast United States

Lydia Stefanova · Vasubandhu Misra ·
James J. O'Brien · Eric P. Chassignet ·
Saji Hameed

Received: 25 June 2010 / Accepted: 23 December 2010

Abstract This paper presents an assessment of the seasonal prediction skill of current global circulation models, with a focus on the two-meter air temperature and precipitation over the Southeast United States. The model seasonal hindcasts are analyzed using measures of potential predictability, anomaly correlation, Brier skill score, and Gerrity skill score. The systematic differences in prediction skill of coupled ocean–atmosphere models versus models using prescribed (either observed or predicted) sea surface temperatures (SSTs) are documented. It is found that the predictability and the hindcast skill of the models vary seasonally and spatially. The largest potential predictability (signal-to-noise ratio) of precipitation anywhere in the United States is found in the Southeast in the spring and winter seasons. The maxima in the potential predictability of two-meter air temperature, however, reside outside the Southeast in all seasons. The largest deterministic hindcast skill over the Southeast is found in wintertime precipitation. At the same time, the boreal winter two-meter air temperature hindcasts have the smallest skill. The large wintertime precipitation skill, the lack of corresponding two-meter air temperature hindcast skill, and a lack of precipitation skill in any other season are features common to all three types of models (atmospheric models forced with observed SSTs, atmospheric models forced with predicted SSTs, and coupled ocean–atmosphere models). Atmospheric models with observed SST forcing demonstrate a moderate skill in hindcasting spring-and summertime two-meter air temperature anomalies, whereas coupled models and atmospheric models forced with predicted SSTs lack similar skill. Probabilistic and categorical hindcasts mirror the deterministic findings, i.e., there is very high skill for winter precipitation and none for summer precipitation. When skillful, the models are conservative, such that low-probability hindcasts tend to be overestimates, whereas high-probability hindcasts tend to be underestimates.

1 Introduction

In recent years, there has been a growing interest in the dynamical or statistical downscaling of Global Circulation Model (GCM) seasonal and climate forecasts to produce regional-scale predictions. For such an approach to produce useful regional forecasts, the GCMs driving the regional predictions must have a reasonable fidelity of the large scale. The aim of this study is to document the existing seasonal predictability and forecast skill—or lack thereof—of GCMs over the Southeast United States for near-surface variables, and to identify the sources of predictability and skill and their limitations.

The southeast sector of the United States is an area strongly influenced by ENSO (e.g., Ropelewski and Halpert 1987, 1986; Kiladis and Diaz 1989). This influence is most pronounced in precipitation and surface air temperature (e.g., Barnston 1994; Higgins et al. 2000; Markowski and North 2003; Saha et al. 2006). Here, El Niño winters tend to be wet and cold, and La Niña winters tend to be warm and dry. In summer, the impact is reversed, and much weaker.

L. Stefanova (✉) · V. Misra · J. J. O'Brien · E. P. Chassignet
Center for Ocean-Atmospheric Prediction Studies,
Florida State University, Tallahassee, FL, USA
e-mail: lstefanova@fsu.edu

S. Hameed
Asia–Pacific Economic Cooperation (APEC) Climate Center,
Busan, Korea

S. Hameed
University of Aizu, Fukushima, Japan

Published online: 07 January 2011

A number of studies have shown high predictive skill for winter precipitation in various dynamical models (e.g., Saha et al. 2006; Cocke et al. 2007), but a comprehensive study documenting inter-model comparisons for the Southeast United States' climate has been lacking thus far.

Despite the continuous advancements in numerical modeling, dynamical forecasting on time scales of a month or longer remains a challenging problem because of the chaotic nature of the atmospheric system and the large number of different processes that act at different space and time scales. The potential feasibility of dynamical seasonal forecasting is based on the premise that there is long-term predictability associated with the slowly varying (and therefore relatively predictable) surface boundary conditions. In other words, anomalies of the atmospheric states are largely governed by anomalies in the boundary condition forcing. One of the most important surface boundary conditions for the atmosphere is the sea surface temperatures (SSTs).

SST anomalies associated with ENSO have been widely shown to be responsible for the bulk of the variability of the tropical circulation on monthly and seasonal time scales (e.g., Shukla et al. 2000). The extent to which skillful seasonal predictions due to anomalous ENSO forcing are feasible in the extratropics remains subject to some debate (Kumar and Hoerling 1995; Shukla 1998). The prediction skill in general varies both geographically and seasonally and is dependent on the specific variable being forecast. Additionally, some boundary condition regimes–such as a strong ENSO–are associated with increased predictability and skill (Brankovic and Palmer 2000).

A standard tool for assessing the predictability inherent to a modeling system is the comparison of the external (i.e., boundary-forced) variability to the internal variability. In practice, the external and internal variability are estimated by using an ensemble forecast approach. A situation in which the externally forced variability exceeds the internal variability is an indication of the forecast system's ability to distinguish effectively between different forcing regimes (see "Appendix A1"). However, there is no certainty that, under such circumstances, the response to the different forcing regimes would be accurate. To assess the accuracy of the forecast system response, measures of forecast skills must be applied.

The present assessment is conducted in a multimodel framework using a collection of GCM hindcasts of precipitation and two-meter air temperature for a 20-year period. The model skill is evaluated in terms of anomaly correlations between the hindcast and observed seasonal means for each grid point in the domain. The collection of hindcasts used in this paper encompasses a range of models. Some are atmospheric models that use prescribed observed SSTs, some are atmospheric models that use prescribed predicted SSTs, and others are coupled ocean–atmosphere models that predict their own SSTs consistent with the evolution of the atmosphere. We find systematic differences in the forecast skills among the three types of models in terms of forecast skill, as will be discussed further.

The remainder of this paper is organized as follows. The models used for the assessment are described in Sect. 2. The methodology for the model evaluation is presented in Sect. 3. Section 4 presents and discusses the study findings. A summary and conclusions are given in Sect. 5.

2 Models

The model data used in this diagnostic study consist of seasonal hindcasts for 1982–2001 from several global models hosted by the Asia Pacific Economic Cooperation Climate Center (APCC). APCC maintains a diverse international collection of GCM seasonal forecasts that are freely available in real time, along with an archive of the GCM hindcasts. Since very few other sources systematically provide GCM seasonal forecast data in real time with open access, it is important to assess the potential usefulness of this collection for future regional predictions.

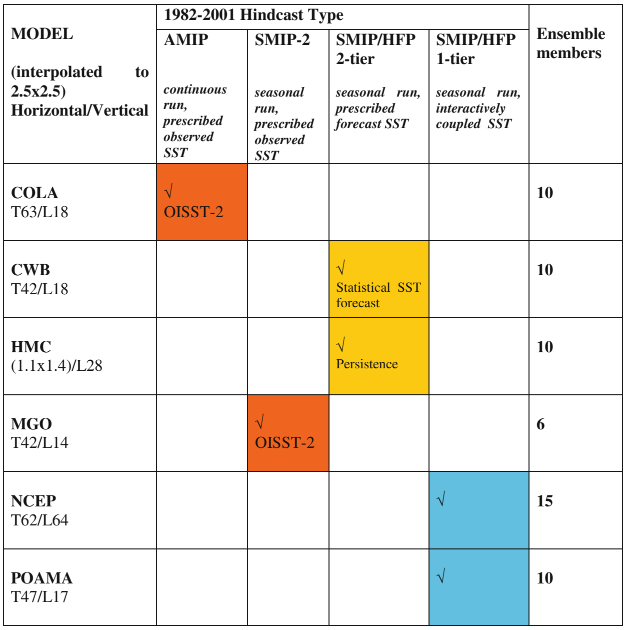
From all the models available in the APCC collection, we selected the six that had the longest hindcast period in common and contained the variables of interest. The salient features of these models, such as their original resolution, number of ensemble members, and SSTs used, are summarized in Table 1. Although this sample of models is limited, we feel that the results presented here represent the nature of seasonal forecast skill in current global circulation models. The study includes both coupled and uncoupled global atmospheric models (the latter using a variety of prescribed SSTs), encompassing the current practices for issuing dynamical seasonal forecasts.

The models used in this study fall into one of three types, depending on the source of SSTs used in the integration, as described below. With the exception of the Center for Ocean-Land–Atmosphere (COLA) model, which is run in the Atmospheric Model Intercomparison Project (AMIP) mode (Gates et al. 1999) (i.e., as a restart of a continuous multi-decadal integration with observed SST), all hindcasts are initialized roughly a month before the beginning of the target season. The number of ensemble members for each model is different: roughly ten, but for some models as few as six, and for others as many as fifteen (see Table 1 for details).

2.1 GCMs using prescribed observed SSTs

Two of the models use prescribed observed SSTs as a boundary condition over oceans. The first is the COLA

Table 1 Summary of the models used for the multimodel skill comparison and the type of SSTs that they use



MODEL **(interpolated to 2.5x2.5) Horizontal/Vertical**	**1982-2001 Hindcast Type**				**Ensemble members**
	AMIP *continuous run, prescribed observed SST*	**SMIP-2** *seasonal run, prescribed observed SST*	**SMIP/HFP 2-tier** *seasonal run, prescribed forecast SST*	**SMIP/HFP 1-tier** *seasonal run, interactively coupled SST*	
COLA T63/L18	√ OISST-2				**10**
CWB T42/L18			√ Statistical SST forecast		**10**
HMC (1.1x1.4)/L28			√ Persistence		**10**
MGO T42/L14		√ OISST-2			**6**
NCEP T62/L64				√	**15**
POAMA T47/L17				√	**10**

model (Schneider 2002). This is the only AMIP-type run in the set, with atmospheric conditions initialized in 1981. It uses prescribed weekly SSTs from the NOAA Optimum Interpolation Sea Surface Temperature Analysis (OISST-V2) data set (Reynolds et al. 2002).

The second model using prescribed observed SSTs is the Russian MGO model (Shneerov et al. 2002). The SSTs used by this model are also from OISST-V2. The model is initialized using the 00Z and 12Z analyses from the Hydrometeorological Centre (HMC) in Moscow for the 3 days immediately preceding the target season.

2.2 Uncoupled GCMs using prescribed predicted SST

Two of the models use prescribed predicted SSTs. One is the Central Weather Bureau (CWB) model from Taiwan (Liou et al. 1997). The SSTs used by this model are those predicted by OPGSST (Optimized Global SST system). OPGSST is an SVD-based statistical SST forecast system, developed by CWB, whose predictors include the Niño 3.4 SST index, tropical Pacific Ocean heat content, sea-level pressure around the Philippines, and the predictions of two simple coupled dynamical models (see Weng et al. 2005 for details). The CWB model is initialized with the National Centers for Environmental Prediction/Department of Energy (NCEP/DOE) Reanalysis Version 2 (Kanamitsu et al. 2002) at 12Z on the last 10 days of the month 2 months prior to the target season (i.e., initialized on the last 10 days of April for a summer, June-July–August, forecast). The other model using prescribed predicted SSTs is the Russian HMC model (Trosnikov et al. 2005). For this model, the monthly mean SST anomaly prior to the target forecast season is persisted for the duration of the integration. The ensemble members are generated by successively initializing the atmosphere from the last 5 days of the previous month at 12-h intervals.

2.3 GCMs using interactively coupled SSTs

The remaining two models in our subset are coupled ocean–atmosphere models in which the SSTs are freely evolving instead of being prescribed. The first is the NCEP Coupled Forecasting System (CFS) model (Saha et al. 2006), a fully coupled GCM. The ensemble members are initialized at 15 different times, ranging from the 9th day of the month 2 months prior to the target season to the 3rd day of the month immediately preceding the target season. The final model used in this assessment is the POAMA-1.5 Coupled Forecast System (Wang et al. 2008), developed by the Centre for Australian Weather and Climate Research/Bureau of Meteorology. This is a one-tier coupled model that has the temperature assimilated into the top 500 m of the ocean every 3 days, using Optimum Interpolation (Smith et al. 1991). It is initialized from an atmosphere/land pseudo-reanalysis, wherein the corresponding AMIP integration run is nudged to the 3-D atmosphere of the European Centre for Medium-Range Weather Forecast (ECMWF) 40-year Reanalysis (ERA-40). The model is initialized on the first day of the month preceding the target season using 6-hourly lagged atmospheric initial conditions.

3 Methodology

All model outputs are converted from the corresponding original grid (see the first column in Table 1) to a common 2.5° by 2.5° global grid. We have defined the Southeast United States domain to be all the land points included within 85°W–75°W and 22.5°N–40°N. The APCC model anomalies are calculated relative to the respective model climatology.

Following the recommendations for standardized verification set forward by the World Meteorological Organization's Lead Centre for the Long range Forecast Verification System (http://www.bom.gov.au/wmo/lrfvs/datasets.shtml), verification observations are retrieved from the Climate Prediction Center (CPC) Merged Analysis of Precipitation (CMAP) for the rainfall field (Xie and Arkin 1996) and the European Centre for Medium-Range Weather Forecasts (ECMWF) 40-Year Reanalysis (ERA-40) for the two-meter air temperature field.

A comparison between signal and noise is used to estimate the potential predictability in the multimodel system (see "Appendix 1"). The skill of forecasts in a deterministic sense is assessed using temporal anomaly correlations. The anomaly correlation between the hindcast (here, the hindcast is the multi-model average of multiple models' ensemble means) and observed time series is calculated as

$$AC_{ij} = \frac{\sum_{n=1}^{N}\left(f_{ij}(n) - \overline{f_{ij}}\right)\left(o_{ij}(n) - \overline{o_{ij}}\right)}{\sqrt{\sum_{n=1}^{N}\left(f_{ij}(n) - \overline{f_{ij}}\right)^2 \sum_{n=1}^{N}\left(o_{ij}(n) - \overline{o_{ij}}\right)^2}}$$

Here $f_{ij}(n)$ is the hindcast at the (i,j)th grid point at time n, $o_{ij}(n)$ is the corresponding observation, $\overline{a} = \frac{1}{N}\sum_{n=1}^{N} a(n)$, and N is 20 (the number of years in the data set).

Brier skill score (Murphy 1973) (see "Appendix 2" for details) is used as a measure of probabilistic skill, and a multicategory equitable threat score (Gerrity skill score, Gerrity 1992) is used for the assessment of categorical forecast skill. The commonly used equitable threat score (ETS) is a probabilistic forecast measure for dichotomous forecasts. It reflects the fraction of observed and/or forecast events that were correctly predicted, adjusted for hits associated with random chance. The value of ETS for a perfect deterministic forecast is 1, and for a climatological forecast, 0. Accounting for hits due to chance by the ETS allows scores to be compared more equitably across different regimes. The Gerrity skill score is an extension of this concept to a multicategory forecast. In this case, we have used three categories: "below normal," "normal," and "above normal," defined as the lower, middle, and upper, respectively, terciles of either precipitation or two-meter air temperature anomaly.

4 Results and discussion

Figure 1 shows the anomaly correlation of the APCC multimodel ensemble seasonal precipitation with the observations, along with contours of potential predictability, over land for the contiguous United States. This multimodel ensemble groups all three types of hindcasts (see Sect. 2): those forced with observed SSTs, those forced with prescribed forecast SSTs and those with coupled SSTs. While this is not an operationally useable combination, it is a useful theoretical construct that illustrates the potential skill and predictability from a diverse set of seasonal hindcasts. The anomaly correlation is shown by the color of the shading. Based on a one-tailed test for a series of twenty data points (the 20 years in our data set), correlations above 0.38 are significant at the 95% confidence level. The reason for displaying pixels with correlation values below the significance threshold is two-fold. First and foremost, it allows for visually estimating the degree of spatial homogeneity of the results. Second, the color scale is such that these results can be directly matched with those of Saha et al. (2006) for the NOAA/NCEP-CFS model, thus allowing for a visual comparison. The potential predictability, i.e., the ratio between the external and the total variance, is indicated with dark red contours; these contours are dashed for the 0.40 to 0.50 levels and solid for values above 0.50.

The Southeast is the only region in the United States that reaches a predictability ratio exceeding 0.5. This is found in both spring (MAM) and winter (DJF). For these two seasons, only the wintertime hindcasts are associated with remarkably strong anomaly correlations–exceeding 70%. The springtime forecasts, in contrast, show very low anomaly correlations–between 10 and 30%. These values are below the threshold for statistical significance (38%), but they are consistent across the Southeast domain.

In summer and autumn, the only skillful hindcasts are for summer over the Northwest. Elsewhere in the United States, there is neither predictability nor anomaly correlation skill for these seasons.

Figure 2 shows similar maps for the anomaly correlation and potential predictability of the two-meter air temperature anomalies for the four seasons. Interestingly, the regions of predictability are now shifted toward the west. The highest external-to-total variance ratios are found over Texas in the winter and over Arizona, New Mexico, Texas, and the Pacific Northwest in the spring. There is also a pocket of high predictability ratio for the summertime two-meter air temperatures over the state of Nevada. The areas of high predictability ratio for the temperature hindcasts tend to coincide with the regions of high hindcast skill. The Southeast region as a whole shows no noticeable predictability or skill for the two-meter air temperatures, regardless of season, although there is some suggestion of anomaly correlation skill for parts of the domain in some seasons.

The results in Figs. 1 and 2 are based on grouping the hindcasts from all member models. As discussed in Sect. 2, not all of these models have the same treatment of the oceanic temperatures. Since sea surface temperatures are a significant source of predictability on seasonal time scales, it is of interest to investigate the degree to which having perfect SSTs, versus SSTs predicted in either coupled or uncoupled mode, would affect the hindcast skill.

Figure 3 shows the anomaly correlation of precipitation over the Southeast for the four seasons and for three multimodel ensemble configurations. The first column corresponds to the multimodel ensemble including all member models. The second column corresponds to an ensemble of the two models that use observed ("perfect") SSTs (COLA and MGO), and the third column, to an ensemble of the two models using coupled SSTs (NCEP and POAMA). The predictability ratio is shown only in the first column, i.e., when all models are included.

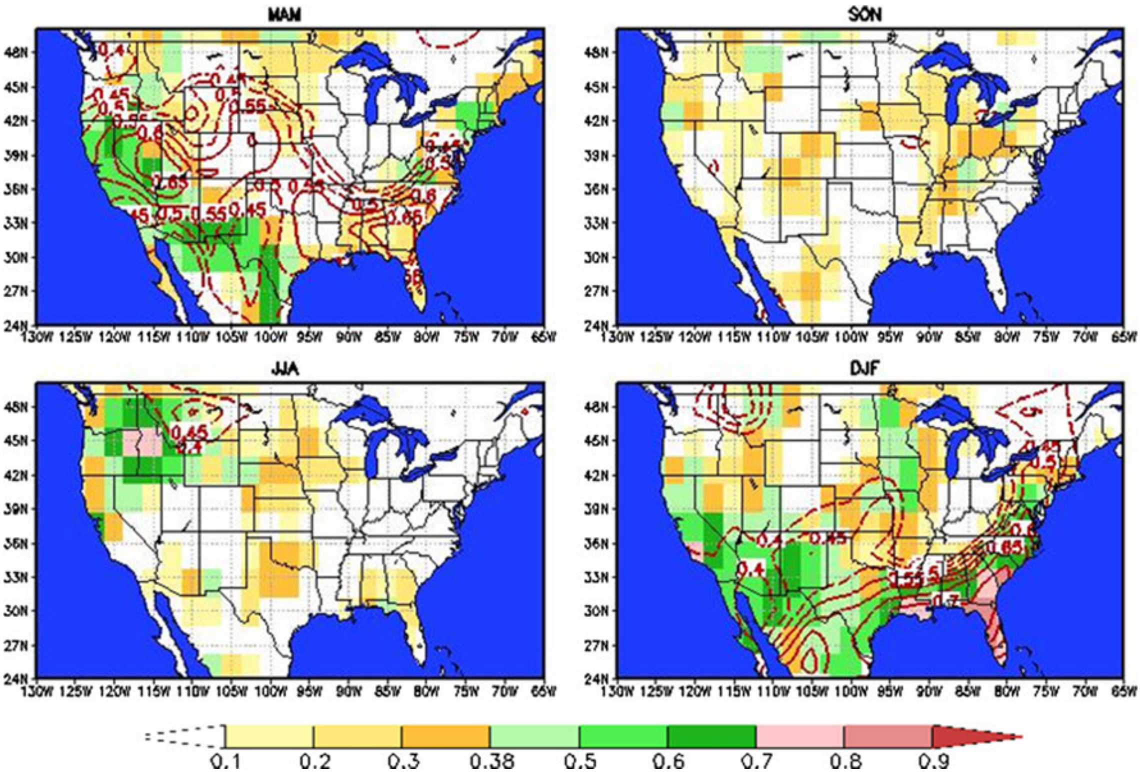


Fig. 1 20-year anomaly correlation (*shading*) and predictability ratio (contours) of the six member multimodel ensemble mean seasonal hindcasts of precipitation for the United States. *Top left* spring (MAM), *bottom left* summer (JJA), *top right* autumn (SON), and *bottom right* winter (DJF). Values below 0.10 are masked out. Values above 0.38 are significant at the 95% level

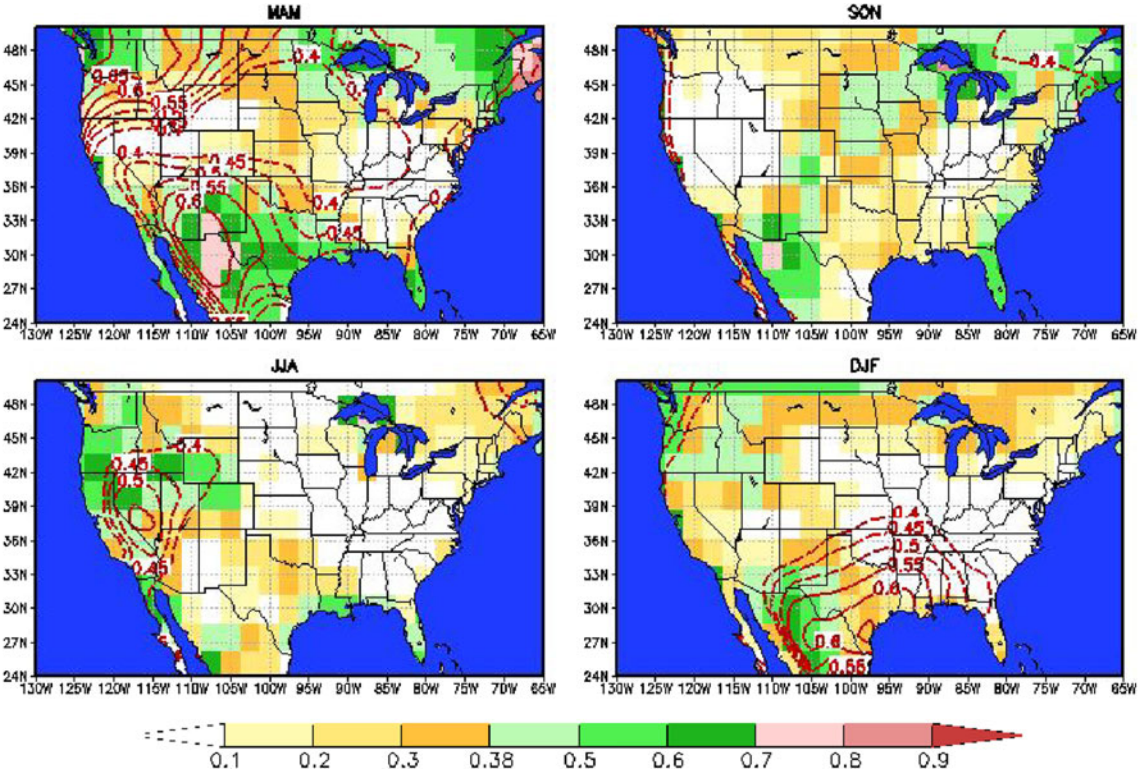


Fig. 2 As in Fig. 1 but for two-meter air temperatures

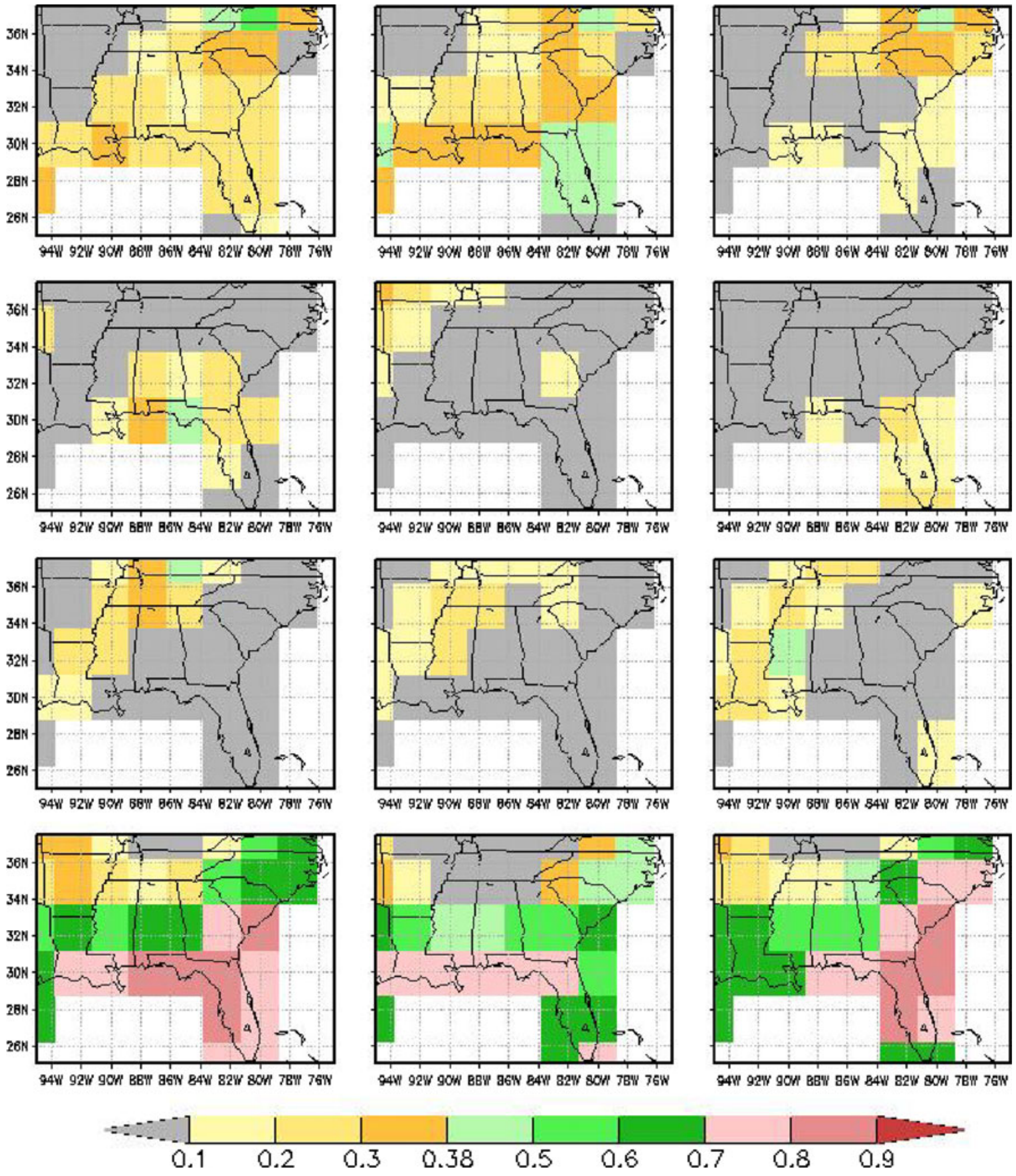


Fig. 3 20-year anomaly correlation (*shading*) of (**a**) the six-member multimodel ensemble mean (*left column*), (**b**) the two-member mean of models using prescribed observed SSTs (*central column*), and (**c**) the two-member mean of models using coupled SSTs (*right column*) for precipitation. The verification season is: MAM (*top row*), JJA (*second row*), SON (*third row*) and DJF (*bottom row*)

A number of features stand out from this intercomparison. Foremost, there is very large skill in wintertime hindcasts of precipitation, regardless of whether one is using prescribed or forecast SSTs. There is also some suggestion of skill (although below the significance threshold) for springtime precipitation using perfect SSTs, which is lacking in the case of coupled SSTs. Finally, there is no skill in the hindcasts of seasonal precipitation, regardless of the SST source, in summer and autumn.

A similar intercomparison for the two-meter air temperatures is shown in Fig. 4. It is striking that models with prescribed SSTs are significantly more skillful than the coupled models, with particularly high skill in spring and summer, especially for the coastal grid points. These results should be interpreted with caution, since the land-sea mask of the models has not been taken into account. It is quite possible that near the coast ocean temperature information may be directly incorporated into the calculation of the

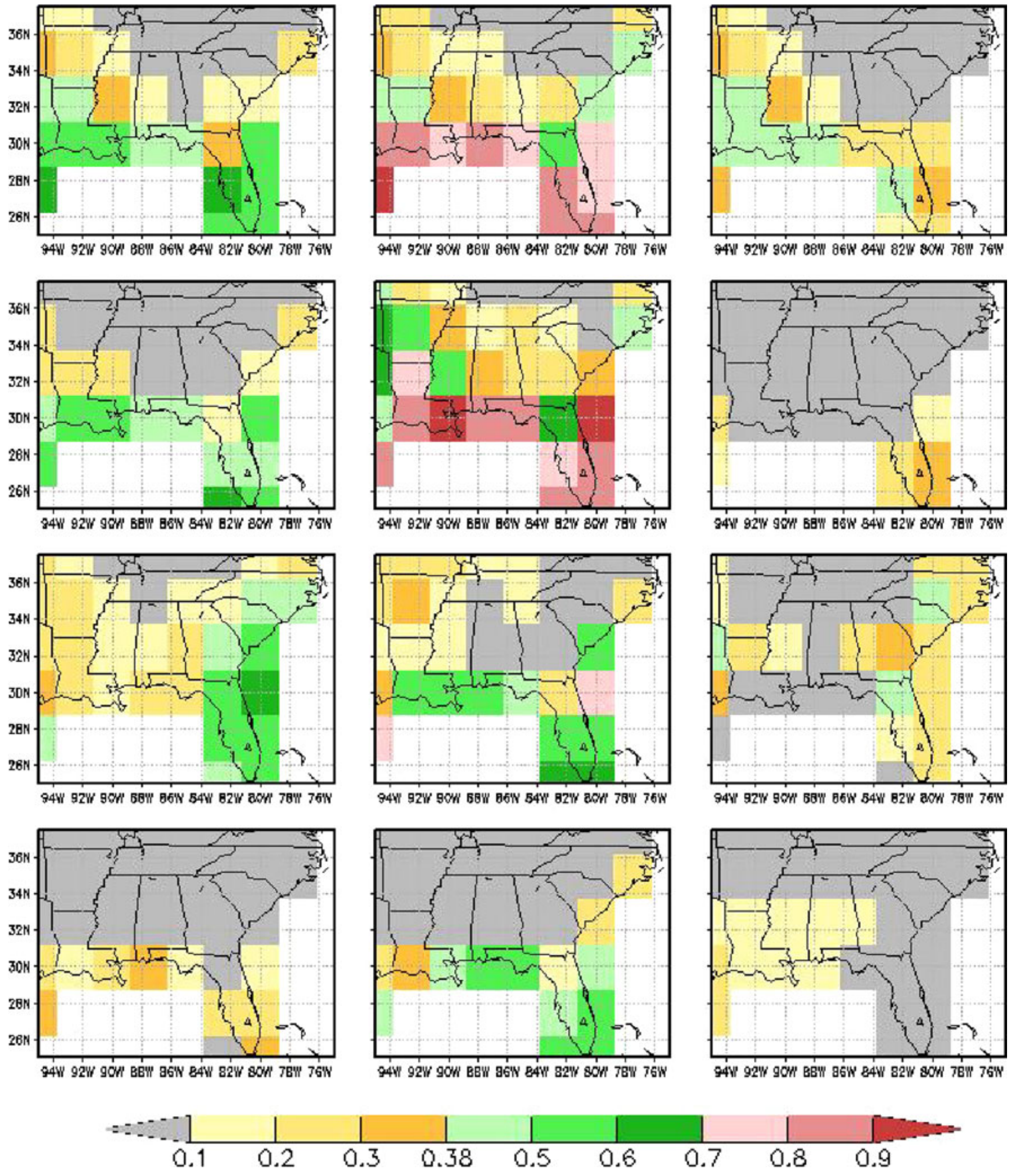


Fig. 4 As in Fig. 3 but for two-meter air temperatures

"land" grid-points' two-meter air temperature. The coupled models have hardly any skill in hindcasting the two-meter air temperature anomalies for the Southeast. Surprisingly, winter, which is the season with the largest skill for precipitation hindcasts, is the season with the smallest skill for two-meter air temperature hindcasts.

Figure 5 shows the area-averaged anomaly correlations for the individual ensemble mean of each of the six member models (see Table 1); the six-member multimodel mean; and the two-member multimodel means for (i) the models with prescribed observed SSTs, (ii) the models with prescribed predicted SSTs, and (iii) the models with coupled SSTs. The area averaging is done only over the land points of the Southeast domain (85°W–75°W and 22.5°N–40°N). The results are stratified by season and shown for (a) precipitation, (b) two-meter air temperatures, and (c) 850 mb zonal wind.

In addition to summarizing the results seen in the spatial maps of anomaly correlation, Figure 5 illustrates the degree of spatial skill homogeneity across the different

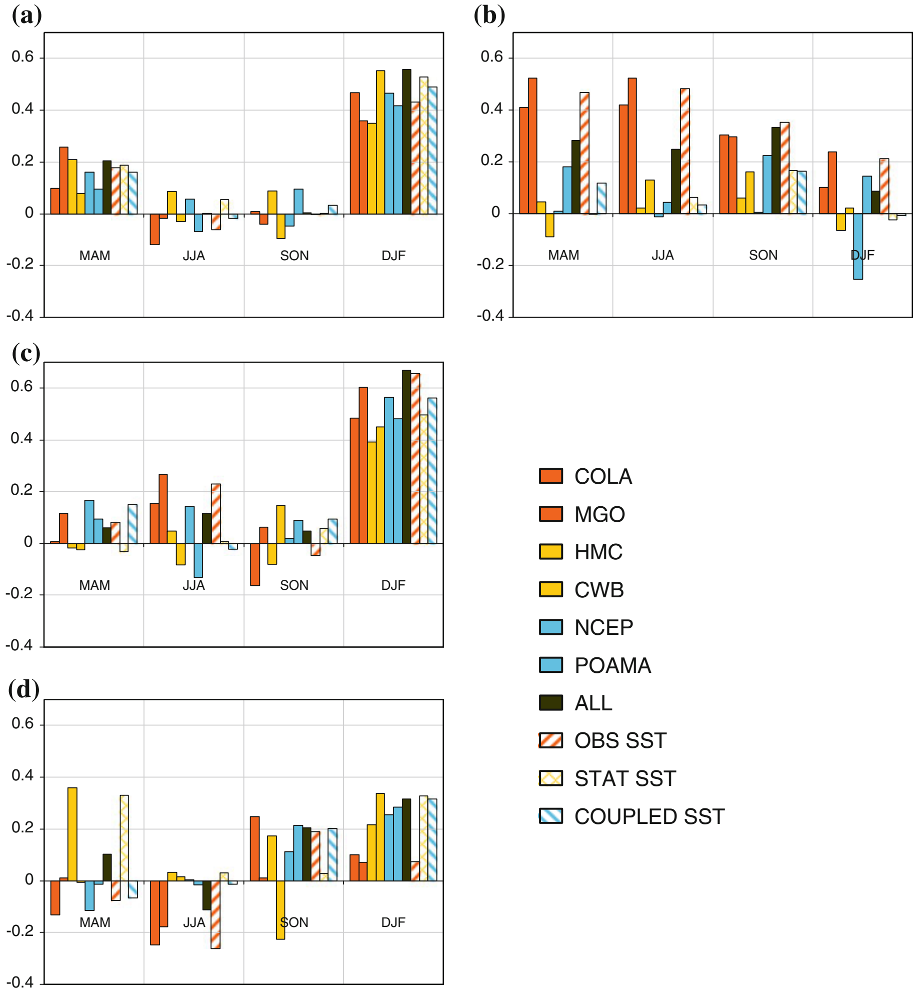


Fig. 5 Area-averaged anomaly correlations for each member model's ensemble mean (*bar colors* correspond to the colors in Table 1), the six-member multimodel ensemble mean ("ALL"), and the two-member multimodel ensemble mean of the models with (i) prescribed observed SST ("OBS SST"), (ii) prescribed predicted SSTs ("STAT SST"), and (iii) coupled model SST ("COUPLED SST"). The panels show the anomaly correlations of (**a**) precipitation, (**b**) two-meter air temperatures, (**c**) zonal wind at 850 mb, and (**d**) meridional wind at 850 mb, grouped by season, as indicated on the abscissa. The area averaging is done all over land points in the Southeast United States

models in the collection. For precipitation, the model hindcast skill appears consistent across models, regardless of their SST treatment. In terms of area-averaged skill, winter hindcasts of precipitation anomalies are by far superior. This skill homogeneity across models reflects the high degree of ENSO influence on the region's precipitation and the high skill of coupled and statistical models in forecasting Niño 3.4 SSTs (Kirtman et al. 2002; Wang et al. 2009). The complete lack of skill of summer and autumn precipitation anomaly hindcasts is likely due to the inability of the coarse resolution global models to resolve tropical storm activity and local convection and to the declining influence of ENSO. Interestingly, the large hindcast skill in winter precipitation anomalies coincides with a relatively large hindcast skill in the zonal wind anomaly at 850 mb. The hindcast skill of the 850 mb meridional wind anomaly (not shown) is smaller, although it also peaks during the winter season.

In the case of two-meter air temperature hindcasts, the skill homogeneity across models is absent. The two models with prescribed observed SSTs exhibit similar hindcast skills, with large area-averaged correlations for spring and summer and decreasing correlations toward the second half of the year. In contrast, neither the models with prescribed predicted SSTs nor the coupled models show a similar pattern. Since the Niño 3.4 temperatures of the coupled models are relatively well simulated (Wang et al. 2009), this leads us to conclude that ENSO has a negligible effect on the modeled seasonal anomalies of two-meter air temperatures over the Southeast United States. This suggests that knowledge of the SST anomalies in regions outside of the highly predictable Niño 3.4 region or other surface boundary conditions (e.g., soil moisture) may be important for simulating the interannual variability of the near-surface temperatures of the Southeast United States.

Figure 6 shows two reliability diagrams (plots of the observed conditional versus forecast event frequency) for wintertime precipitation anomalies exceeding 0 and 0.5 mm/day respectively. The values of the Brier skill score (BSS), which measures the probabilistic hindcast skill compared to a climatological hindcast, are labeled on the diagrams. Shown underneath each reliability diagram is the

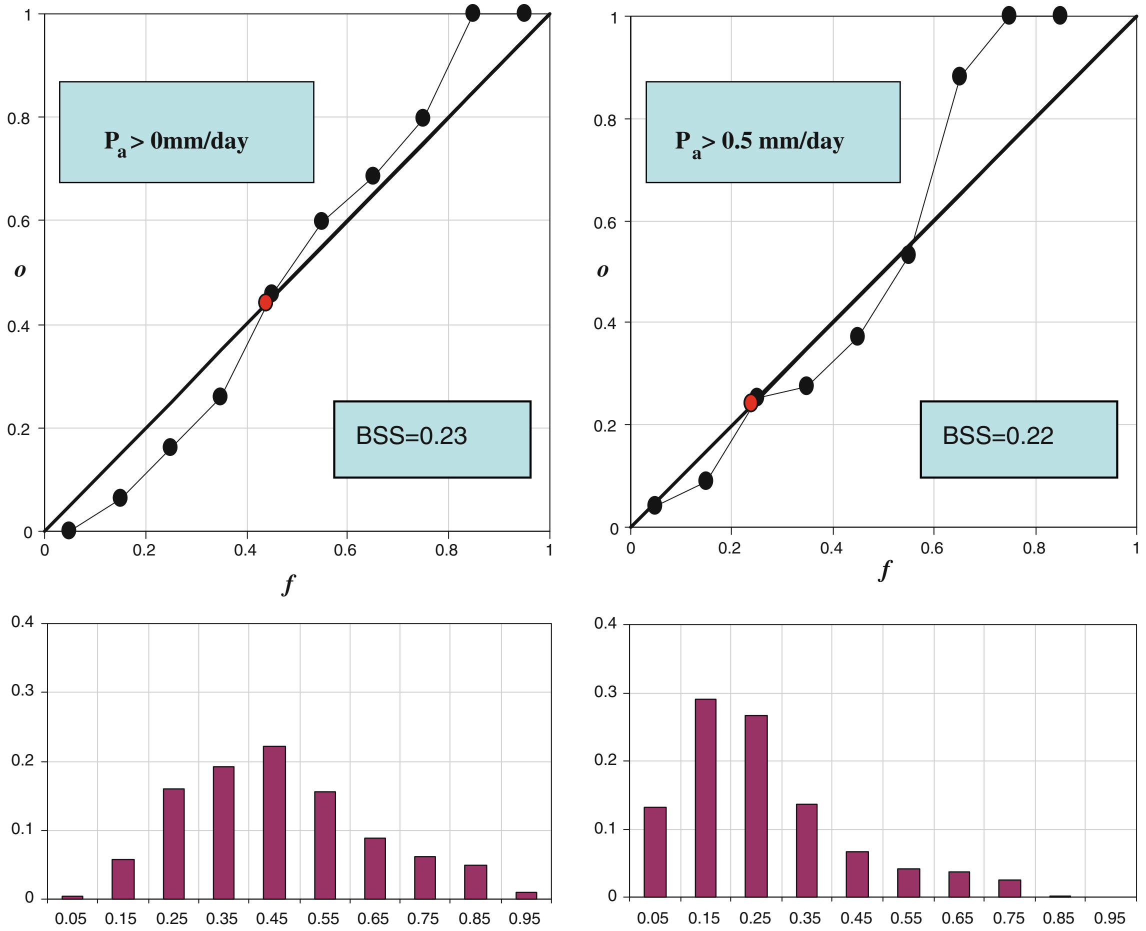


Fig. 6 Reliability diagram (*top*) and relative frequency distribution (*bottom*) for DJF precipitation anomaly for all land points in the Southeast United States for a precipitation anomaly P_a exceeding a threshold of 0 mm/day (*left*) and 0.5 mm/day (*right*). The Brier skill score BSS is indicated on the respective plot. The forecast probability is calculated as the fraction of all models' ensemble members forecasting threshold exceedance. The observed unconditional event frequency is indicated with a *red dot*

relative frequency of each forecast probability bin. These results are obtained using all land points within the Southeast domain and all model ensemble members. A similar set of diagrams is shown for summer precipitation in Fig. 7. The diagrams demonstrate that in probabilistic terms, anomaly precipitation skill is present in winter and absent in summer. When present, the skill is the result of both the reliability and the resolution terms of the BSS decomposition. Low probability forecasts tend to be overestimates, whereas high probability forecasts tend to be underestimates. Similar diagrams for the two-meter air temperatures (not shown) demonstrate lack of probabilistic skill.

Table 2 shows the three-category Gerrity skill score for both precipitation and two-meter air temperatures and for (i) the multimodel ensemble, (ii) models with observed SSTs, (iii) models with prescribed predicted SSTs, and (iv) models with coupled SSTs. The hindcast categories are "below normal," "normal," and "above normal," as determined from distribution terciles. To illustrate the meaning of a GSS score of a given value, sample contingency tables corresponding to the multimodel ensemble winter precipitation and autumn two-meter air temperature are shown in Tables 3 and 4, respectively. All models have high skill in categorical hindcasts for winter precipitation. Models using observed SSTs have high categorical skill for spring and summer two-meter air temperature, whereas models without observed SSTs show no skill in categorical hindcasts for two-meter air temperatures in any season.

5 Summary and conclusions

The season-and-variable combination associated with the largest deterministic hindcast skill for the Southeast United States is the boreal winter season precipitation. At the same time, wintertime two-meter air temperature hindcasts have the smallest skill. The large wintertime precipitation skill, the lack of corresponding two-meter air temperature

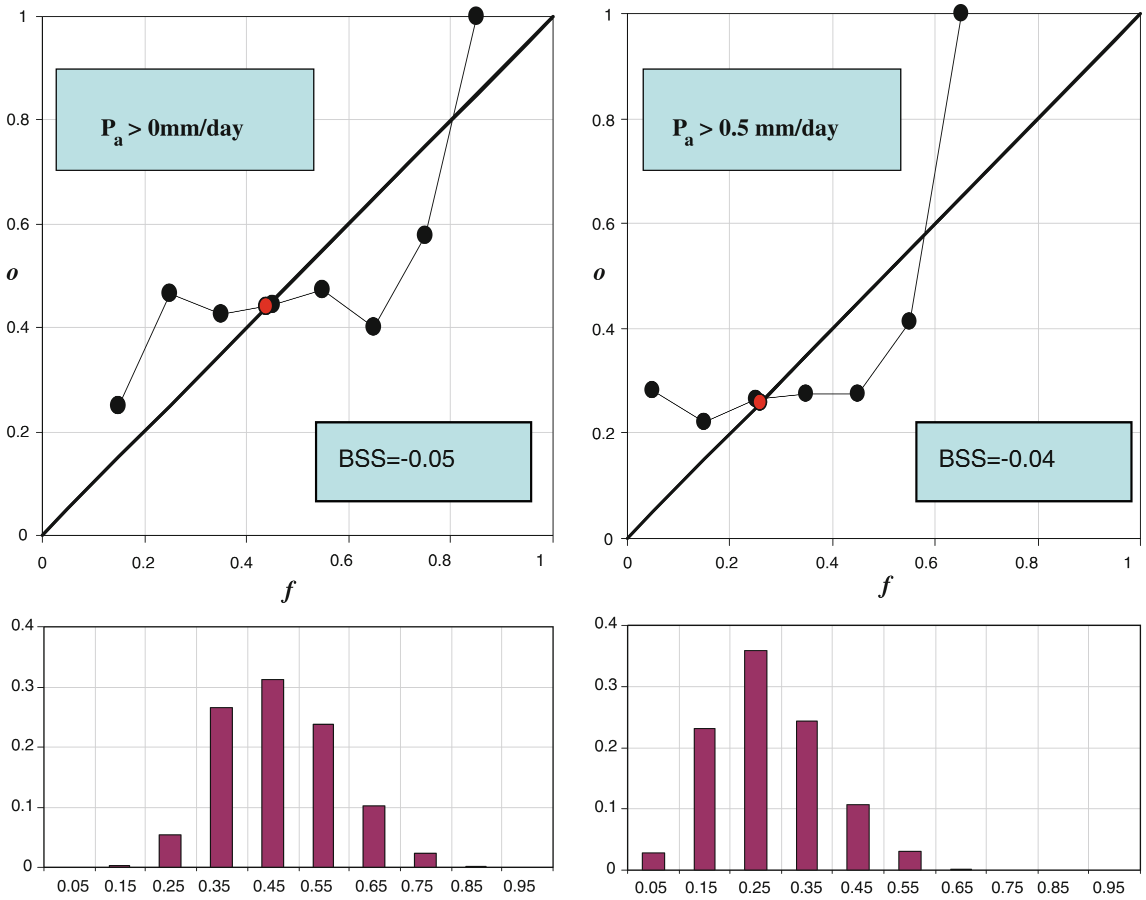


Fig. 7 As in Fig. 6 but for JJA precipitation

Table 2 Three-category Gerrity skill score for both precipitation and two-meter air temperatures, for (i) the multi-model ensemble, (ii) models with prescribed observed SSTs, (iii) models with prescribed predicted SSTs, and (iv) models with coupled SSTs

	Precipitation				Two-meter air temperature			
	Multi-model	Observed SSTs	Predicted SSTs	Coupled SSTs	Multi-model	Observed SSTs	Predicted SSTs	Coupled SSTs
MAM	0.06	0.07	0.01	0.06	0.17	0.35	−0.01	0.05
JJA	0.01	−0.04	0.06	−0.03	0.16	0.28	0.06	0.01
SON	0.01	0.01	0.06	0.04	0.17	0.15	0.01	0.12
DJF	0.37	0.30	0.35	0.30	0.00	0.15	−0.07	−0.02

The hindcast categories are "below normal," "normal," and "above normal," as determined from distribution terciles. A score of 1.0 corresponds to a perfect forecast, and a score of 0.0 corresponds to climatology

hindcast skill, and a lack of precipitation skill in any other season are features common to all three types of models (the atmospheric models forced with observed SSTs, the atmospheric models forced with predicted SSTs, and the coupled ocean–atmosphere models). Models with observed SST forcing demonstrate a moderate hindcast skill for two meter air-temperature anomalies in the spring and summer seasons, whereas the coupled models and the models forced with predicted SSTs do not have any skill.

The high predictability of winter precipitation can be attributed to the remote ENSO forcing. A large fraction of the rainfall in the winter over the Southeast United States is of large scale and associated with fronts. Above normal winter precipitation in the Southeast United States is the result of increased frequency and intensity of winter cold fronts (Hardy and Henderson 2003; Winsberg 2003). This frequency is controlled by the relative locations of upper air troughs and ridges above the North American continent

Table 3 Contingency table for the multimodel ensemble precipitation in winter (DJF), with a GSS score of 0.37

Obs Fcst	Below	Normal	Above
Below	**23**	10	5
Normal	10	**11**	9
Above	3	12	**18**

Table 4 As Table 3 but for two-meter air temperature in autumn (SON), with a GSS score of 0.17

Obs Fcst	Below	Normal	Above
Below	**15**	12	8
Normal	14	**11**	7
Above	9	10	**15**

and the strength and position of the subtropical jet stream. Stronger than usual subtropical jet westerlies over the region result in increased frequency and intensity of cold fronts and Gulf of Mexico winter cyclones (Kiladis and Diaz 1989; Hardy and Henderson 2003), and therefore in increased precipitation amounts. Such conditions are strongly associated with increased large-scale meridional temperature gradients resulting from the relatively well-predicted positive ENSO signal in the Tropical Pacific SSTs.

The models' lack of hindcast skill in precipitation in summer and fall seasons can be attributed to the largely convective nature of precipitation during the warm months. Neither the processes responsible for organized convection (such as sea breeze or tropical storms) nor those governing convection (parameterized processes) in general are properly resolved in the coarse resolution global circulation models.

A possible explanation for the presence of skill in hindcasts of warm season two-meter air temperature in models with observed SST and the lack of corresponding skill in coupled models is that local effects are the predominant factor. During the warm months, air temperatures are dominated by local forcing through surface fluxes (Misra and Dirmeyer 2009). In coastal areas, part of the underlying surface is ocean; models with prescribed perfect SSTs accurately reflect the local SST gradients for such grid points, whereas models with predicted SSTs do not.

The lack of hindcast skill in winter two-meter air temperatures, regardless of the SSTs used in the model, is more difficult to explain. As Misra and Dirmeyer (2009) indicate, during winter season, surface evaporation over the Southeast United States is energy limited. In other words, the near-surface temperature variations are dictated by radiative and sensible heat fluxes. As a result, precipitation and near-surface temperature are not significantly correlated, Therefore, forecast skill in precipitation will not necessarily translate to two-meter air temperature forecast skill. The low winter season hindcast skill of the two-meter air temperature could be reflective of poor fidelity of the radiative fluxes and cloud cover in the models. Since very few other sources systematically provide GCM seasonal forecast data in real time with open access, it is important to assess the potential usefulness of this collection for future regional predictions.

From all the models available in the APCC collection, we selected the six that had the longest hindcast period in common and contained the variables of interest. The salient features of these models, such as their original resolution, number of ensemble members, and SSTs used, are summarized in Table 1. Although this sample of models is limited, we feel that the results presented here represent the nature of seasonal forecast skill in current global circulation models. The study includes both coupled and uncoupled global atmospheric models (the latter using a variety of prescribed SSTs), encompassing the current practices for issuing dynamical seasonal forecasts.

Given the inability of models with predicted SST (either separately forecast or computed in a coupled mode) to produce skillful seasonal forecasts of near-surface variables (with the notable exception of winter precipitation), their practical utility for the Southeast is limited. It remains to be seen whether dynamical or statistical downscaling of seasonal and climate forecasts can overcome the global model forecast limitations described above.

Acknowledgments We thank the various model providers for making available, through APEC Climate Center (APCC), the hindcast datasets used in this study. We thank Ms. Kyong Hee An for facilitating the data access and for providing associated documentation and Ms. Kathy Fearon for her careful reading of the manuscript and helpful editorial comments. All model data used in this study are available online from APCC (http://www.apcc21.net). This research was supported by NOAA grant NA07OAR4310221 and USDA grant 2088-38890-19013.

Appendix 1: Potential predictability

Following Kumar and Hoerling (1995), consider an ensemble of forecasts for any forecast variable A. Let there be $I = 1,\ N$ ensemble members with $\alpha = 1{,}M$ years of external forcing. The ensemble mean forecast, or the most likely outcome, for a given year α is then

$$\overline{A_\alpha} = \frac{1}{N}\sum_{i=1}^{N} A_{i\alpha}.$$

The internal variance, or spread, of the ensemble members around this mean is

$$\sigma_\alpha^2 = \frac{1}{N}\sum_{i=1}^{N}\left(A_{i\alpha} - \overline{A_\alpha}\right)^2.$$

Since the spread can be dependent on the particular choice of year, the internal variance is then σ_α^2averaged over all possible α, or

$$\sigma_I^2 = \frac{1}{M}\sum_{\alpha=1}^{M}\sigma_\alpha^2 = \frac{1}{M}\frac{1}{N}\sum_{\alpha=1}^{M}\sum_{i=1}^{N}\left(A_{i\alpha} - \overline{A_\alpha}\right)^2.$$

The external variance is an estimate of the degree to which the difference between the ensemble mean forecast for different years is due to the boundary conditions rather than to "chance"; thus it is a measure of the forecast's ability to distinguish between different regimes associated with different boundary conditions. The overall mean forecast, averaged over all realizations and boundary conditions, is given by

$$\bar{A} = \frac{1}{M}\frac{1}{N}\sum_{\alpha=1}^{M}\sum_{i=1}^{N} A_{i\alpha},$$

then the external variance is given by

$$\sigma_E^2 = \frac{1}{M}\sum_{\alpha=1}^{M}\left(\overline{A_\alpha} - \overline{A}\right)^2.$$

The total variance of the system is then

$$\sigma_T^2 = \sigma_E^2 + \sigma_I^2.$$

By estimating the ratio of σ_E^2 to σ_T^2 or σ_I^2 we can then judge what part of the observed signal is due to boundary conditions and what part is due to the uncertainty of initial conditions, i.e., is effectively noise. The larger the ratio, the higher the predictability inherent to the system. The values of the ratio range between zero and one. In the case of zero, the ensemble does not see the boundary conditions, i.e., the outcome is entirely due to noise. In the case of one, the boundary conditions overwhelmingly mask out the effect of uncertainty of initial conditions.

The distance between the centroids of distributions for two different boundary condition regimes is a representation of the system's external variability. The sharpness of the distribution associated with a particular regime is representative of the internal variance and represents the range of possible incomes associated with the particular boundary conditions. If the distributions are well separated, either because they are narrow or because their center points are far apart, the two states can be easily distinguished. If the reason for the distinguishability is that the curves are narrow, it can be said that the predictability is due to the internal variability being low. If what makes the distinction possible is that the two states are far apart, then the predictability is attributable to the large external variance.

Appendix 2: Brier skill score and reliability diagram

A probabilistic forecast is one that estimates the probability of occurrence of a chosen event $\mathscr{E}$, such as a precipitation rate anomaly relative to the mean state exceeding a preselected threshold level. For an ensemble of equally reliable models the probability P of the event $\mathscr{E}$ is $(m/M) \times 100$, where m is the number of ensemble members forecasting $\mathscr{E}$, and M is the total number of ensemble forecasts. Since for a single realization a probability forecast is neither correct nor wrong, probability forecasts are verified by analyzing the joint (statistical) distribution of forecasts and observations.

The Brier score measures the magnitude of the probability forecast errors. It is defined as

$$b = \frac{1}{n}\sum_{k=1}^{n}(f(k) - o(k))^2,$$

where the index k refers to the forecast/observation pairs, and n is the total number of such pairs within the data set, and both the forecast (f) and the observations (o) are in terms of probabilities. The lowest possible value of the Brier score is zero, and it can only be achieved with a perfect deterministic forecast.

Let the probabilistic forecast for $\mathscr{E}$ be done within I discrete categories y_i. The frequency with which forecasts of y_i are issued is $p(y_i)$. The frequency within a category y_i forecast with which the event $\mathscr{E}$ actually occurs is the conditional frequency $\overline{o_i} = p(o(k) = 1|y_i)$. A reliability diagram is a plot of $\overline{o_i}$ versus y_i, accompanied by the forecast frequency distribution $p(y_i)$ versus y_i. For a perfect forecast, the reliability diagram would be a line at 45°.

As suggested by Murphy (1973), it is useful to decompose the Brier score into three terms: reliability, resolution, and uncertainty:

$$b = \underbrace{\sum_{i=1}^{I} p(y_i)(y_i - \overline{o_i})^2}_{\text{reliability}} - \underbrace{\sum_{i=1}^{I} p(y_i)(\overline{o_i} - \overline{o})^2}_{\text{resolution}} + \underbrace{\overline{o}(1 - \overline{o})}_{\text{uncertainty}}$$
$$= b_{\text{rel}} - b_{\text{res}} + b_{\text{unc}},$$

where $\overline{o} = \frac{1}{n}\sum_{k=1}^{n} o(k)$ is the unconditional mean frequency of occurrence of the event $\mathscr{E}$.

The reliability term evaluates the statistical accuracy of the forecast–a perfectly reliable forecast is one for which the observed conditional frequency $\overline{o_i}$ is equal to the forecast probability (i.e., over all forecast for y percent chance of $\mathscr{E}$, $\mathscr{E}$ will occur in y percent of the times). The resolution term addresses the distance between the forecast frequency and the unconditional climatological frequency. Forecasts that are always close to the climatological frequency exhibit good reliability (since the forecast frequency matches the observed frequency) but poor resolution (since they are not able to distinguish between different regimes). The uncertainty term is a measure of the variability of the system and is not influenced by the forecast. The Brier skill score is calculated with respect to a reference forecast as

$$BS = \frac{b - b_{\text{ref}}}{b_{\text{perf}} - b_{\text{ref}}} = 1 - \frac{b}{b_{\text{ref}}}, \quad \text{since } b_{\text{perf}} \text{ is } 0.$$

For a *perfect* forecast system, $BS = BS_{\text{rel}} = BS_{\text{res}} = 1$, while for a *climatological* forecast $BS = BS_{\text{rel}} = BS_{\text{res}} = 0$.

References

Barnston AG (1994) Linear statistical short-term climate predictive skill in the Northern Hemisphere. J Climate 7:1513–1564

Brankovic C, Palmer TN (2000) Seasonal skill and predictability of ECMWF PROVOST ensembles. Q J Roy Meteor Soc 126(567): 2035–2067

Cocke S, LaRow TE, Shin DW (2007) Seasonal rainfall predictions over the southeast United States using the Florida State University nested regional spectral model. J Geophys Res 112. doi:10.1029/2006JD007535

Gates WL, Boyle J, Cove C, Dease C, Doutriaux C, Drach R, Fiorino M, Gleckler P, Hnilo J, Marlais S, Phillips T, Potter G, Santer BD, Sperber KR, Taylor K, Williams D (1999) An overview of the results of the Atmospheric Model Intercomparison Project (AMIP I). Bull Amer Meteorol Soc 80:29–55

Gerrity JP (1992) A note on Gandin and Murphy's equitable skill score. Mon Wea Rev 120:2709–2712

Hardy JW, Henderson KG (2003) Cold front variability in the southern United States and the influence of atmospheric teleconnection patterns. Phys Geogr 24:120–137

Higgins RW, Leetma A, Xue Y, Barnston A (2000) Dominant factors influencing the seasonal predictability of US precipitation and surface air temperature. J Climate 13:3994–4017

Kanamitsu M, Ebisuzaki W, Woollen J, Yang S-K, Hnilo JJ, Fiorino M, Potter GL (2002) NCEP-DOE AMIP-II reanalysis (R-2). Bull Amer Meteorol Soc 83:1631–1643

Kiladis GN, Diaz HF (1989) Global climate extremes associated with extremes of the Southern oscillation. J Climate 2:1069–1090

Kirtman BP, Shukla J, Balmaseda M, Graham N, Penland C, Xue Y, Zebiak S (2002) Current status of ENSO forecast skill: a report to the Climate Variability and Predictability (CLIVAR) Numerical Experimentation Group (NEG). CLIVAR Working group on seasonal to interannual prediction. p 31

Kumar A, Hoerling MP (1995) Prospects and limitations of seasonal atmospheric GCM predictions. Bull Amer Meteorol Soc 76:335–345

Liou C-S, Chen J-H, Terng C-T, Wang F-J, Fong C-T, Rosmond TE, Kuo H-C, Shiao C-H, Cheng M-D (1997) The second-generation global forecast system at the Central Weather Bureau in Taiwan. Weather Forecast 12(3):653–663

Markowski GR, North GR (2003) Climatic influence of sea surface temperature: evidence of substantial precipitation correlation and predictability. J Hydromet 4:856–877

Misra V, Dirmeyer PA (2009) Air, sea, and land interactions of the continental US hydroclimate. J Hydromet 10:353–373

Murphy AH (1973) A new vector partition of the probability score. J Appl Meteor 12:595–600

Reynolds RW, Rayner NA, Smith TM, Stokes DC, Wang W (2002) An improved in situ and satellite SST analysis for climate. J Climate 15:1609–1625

Ropelewski CF, Halpert MS (1986) North American precipitation and temperature patterns associated with the El Nino/Southern Oscillation (ENSO). Mon Wea Rev 114:2352–2362

Ropelewski CF, Halpert MS (1987) Global and regional scale precipitation patterns associated with the El Niño/Southern Oscillation. Mon Wea Rev 115:1606–1626

Saha S, Nadiga S, Thiaw C, Wang J, Wang W, Zhang Q, van den Dool HM, Pan H-L, Moorthi S, Behringer D, Stokes D, Peña M, Lord S, White G, Ebisuzaki W, Peng P, Xie P (2006) The NCEP climate forecast system. J Climate 19:3483–3517

Schneider EK (2002) Understanding differences between the equatorial Pacific as simulated by two coupled GCMs. J Climate 15:449–469

Shneerov BE, Meleshko VP, Matjugin VA, Spryshev PV, Pavlova TV, Vavulin SV, Shkolnik IM, Subov VA, Gavrilina VM, Govorkova VA (2002) The current status of the MGO global atmospheric circulation model (version-MGO-03). MGO Proceeding 550:3–43

Shukla J (1998) Predictability in the midst of chaos: a scientific basis for climate forecasting. Science 282(5389):728–731

Shukla J, Anderson J, Baumhefner D, Brankovic C, Chang Y, Kalnay E, Marx L, Palmer T, Paolino D, Ploshay L, Schubert S, Straus D, Suarez M, Tribbia J (2000) Dynamical seasonal prediction. Bull Amer Meteor Soc 81(11):2593–2606

Smith RL, Blomley JE, Meyers G (1991) A univariate statistical interpolation scheme for subsurface thermal analyses in the tropical oceans. Prog Oceanogr 28:219–256

Trosnikov IV, Kaznacheeva VD, Kiktev DB, Tolstikh MA (2005) Assessment of potential predictability of meteorological variables in dynamical seasonal modeling of atmospheric circulation on the basis of semi-Lagrangian model SL-AV. Russian Meteorol Hydrol 12

Wang G, Alves O, Hudson D, Hendon H, Liu G, Tseitkin F (2008) SST skill assessment from the new POAMA-1.5 system. BMRC Res Lett 8:2–6

Wang B, Lee J-Y, Kang I-S, Shukla J, Park C-K, Kumar A, Schemm J, Cocke S, Kug J-S, Luo J-J, Zhou T, Wang B, Fu X, Yun W-T, Alves O, Jun EK, Kinter J, Kirtman B, Krishnamurti T, Lau NC, Lau W, Liu P, Peigon P, Rosati T, Schubert S, Stern W, Suarez M, Yamagata T (2009) Advance and prospectus of seasonal prediction: assessment of the APCC/CliPAS 14-model ensemble retrospective seasonal prediction (1980–2004). Clim Dyn 33:93–117

Weng S-P, Tung Y-C, Huang W-H (2005) Predictions of global sea surface temperature anomalies: introduction of CWB/OP-GSST1.1 Forecast System. Proceedings, Conference on Weather Analysis and Forecasting, Taipei, Taiwan, pp 341–345

Winsberg MD (2003) Florida weather. University Press of Florida. p 218

Xie P, Arkin PA (1996) Global precipitation: a 17-year monthly analysis based on gauge observations, satellite estimates, and numerical model outputs. Bull Amer Meteor Soc 78:2539–2558