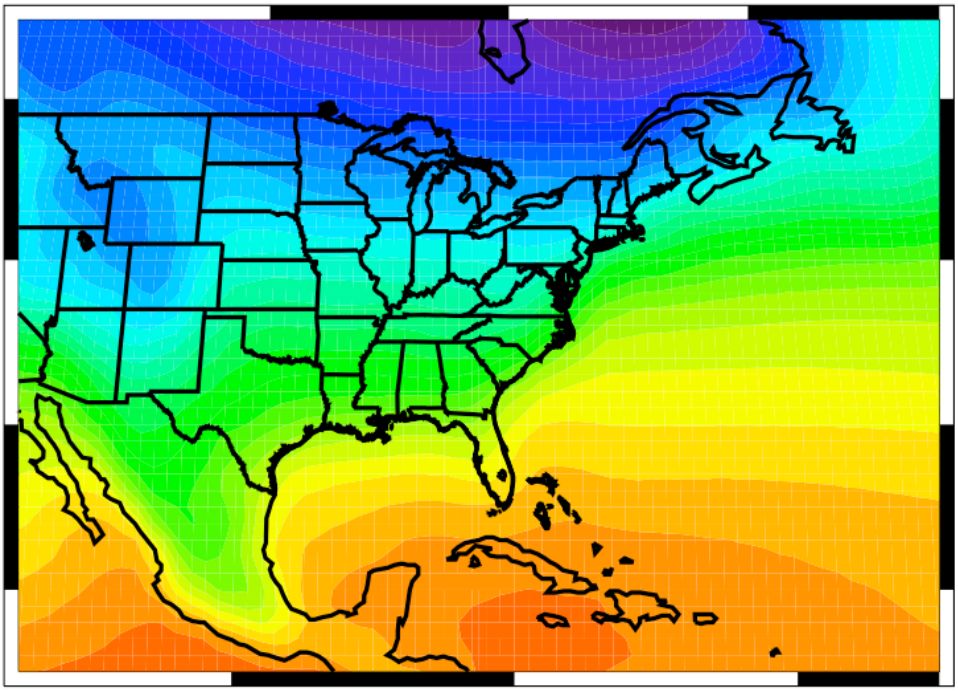
An Evaluation of CMIP5 20th Century Climate Simulations for the Southeast USA



OCCRI

An Evaluation of CMIP5 20th Century Climate Simulations for the Southeast USA

Prepared for the USGS Southeast Climate Science Center

by

David E. Rupp

Oregon Climate Change Research Institute

College of Earth, Ocean, and Atmospheric Sciences

Oregon State University

Corvallis, Oregon, USA

19 November 2014

Front cover image: CMIP5 multi-model average December-March temperature, 1960-1999.

Abstract

The 20th century climate for the Southeast United States (US) and surroundings as simulated by global climate models participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5) was evaluated. A suite of statistics that characterize various aspects of the regional climate was calculated from both model simulations and observation-based datasets. CMIP5 global climate models were ranked by their ability to reproduce the observed climate. Differences in models' performance between regions (i.e., the Southeast and Northwest US) argue for a regional-scale assessment and ranking of CMIP5 models.

1. Introduction

Simulations from the global climate models (GCMs) participating in the Coupled Model Intercomparison Project Phase 5 (CMIP5) provide the basis for many of the conclusions in the recent Intergovernmental Panel on Climate Change (IPCC) Assessment Report #5 (AR5) (IPCC 2013). The data from these simulations are also being relied upon for local and regional climate change assessments across the US, either in direct form, or following a either a statistical or dynamical downscaling transformation to bring the data to a spatial resolution compatible with local or regional matters. In response, the question of GCM reliability (sometimes also termed credibility or veracity), at the regional scale is often raised by users of these data. To help address this question, we evaluated CMIP5 GCMs with respect to how well they reproduce the observed climate of the Southeast United States (US).

Monthly temperature and precipitation data from 41 global climate models (GCMs) of the Coupled Model Intercomparison Project Phase 5 (CMIP5) were compared to observations for the 20th century, with a focus on the Southeast US and surrounding region. The methodology largely followed that used by *Rupp et al.* [2013] for the Pacific Northwest US, who calculated a suite of statistics, or metrics, that characterize various aspects of the regional climate. Performance, or credibility, was assessed based on the GCMs' abilities to reproduce the observed metrics. GCMs were ranked in their credibility using two methods. The first simply treated all metrics equally. The second method considered two properties of the metrics: 1) redundancy of information (dependence) among metrics, and 2) confidence in the reliability of an individual metric for accurately ranking models. Confidence was related to how robust the

estimate of the metric was to ensemble size, given that for most of the models only a small number of ensemble members (i.e. realizations of the 20th century) were available.

2. Data and Methods

The methodology very closely follows *Rupp et al*. [2013] but is repeated below for the reader's convenience. References to the methodology *per se* should cite *Rupp et al*. [2013] and not this report. Any changes to the methodology are explicitly stated in the text below.

2.1. Data

Simulated near surface temperature (T), daily minimum (Tmin) and maximum (Tmax) temperature, and precipitation rate (P) were acquired from 41 GCMs (see Table A1 of the Appendix) of the CMIP5 "historical" experiment [*Taylor et al.*, 2012]. The historical experiment includes both natural and anthropogenic forcings for the years 1850-2005. For a given GCM, the number of members per ensemble varied from 1 to 10, differing only by initial conditions. The data were obtained at the monthly frequency, with the exception of Tmin and Tmax for 3 GCMs. IPSL-CM5A-LR, IPSL-CM5A-MR, and IPSL-CM5B-MR had known problems with monthly mean Tmin and Tmax, so monthly means were calculated from daily data.

For historical observations, we relied on five gridded datasets of monthly means of the following variables: near surface daily minimum, maximum, and average temperature, and surface precipitation rate. The observation-based datasets were:

1) University of East Anglia Climatic Research Unit (CRU) TS3.10.01, 0.5° x 0.5°, 1901-2009 [*Harris et al.*, 2013].
2) Parameter-elevation Regressions on Independent Slopes Model (PRISM), 2.5′ x 2.5′, 1895-2012 [*Daly et al*., 2008].

3) University of Delaware Air Temperature and Precipitation (UDelaware) v.3.01, 0.5° x 0.5°, 1901-2010 [*Willmott and Matsuura*, 2012a;b].

4) National Center for Environmental Prediction/National Center for Atmospheric Research Reanalysis (NCEP), ~1.9° x 1.9°, 1948-2012 [*Kalnay et al*., 1996].

5) European Centre for Medium-Range Weather Forecasts 40 Year Re-analysis (ERA40), ~2.5° x 2.5° mid-1957 to mid-2002 [*Uppala et al*., 2005].

While CRU, PRISM, and UDelaware are based on surface station observations, NCEP and ERA40 are reanalysis datasets based on a numerical model of the atmosphere that assimilates observations to update model states.

CRU, UDelaware, NCEP, ERA40, and CMIP datasets were regridded to a common resolution of 1° x 1° using an inverse-distance-weighting interpolation algorithm. PRISM datasets were regridded by averaging all native cells within the coarser 1° x 1° cell. Grid cell centers were located on the whole degree.

2.2. Performance metrics

The metrics used consider both properties of the regionally averaged time series and larger-scale patterns having regional influence. The following metrics of temperature and precipitation were selected on the basis of having theoretical merits as well as being relevant for impacts modeling:

1) Climatological mean of annual value (Mean).

2) Mean seasonal amplitude (SeasonAmp).

3) Spatial standard deviation (SpaceSD) of the climatological mean field, by season.

4) Spatial correlation (SpaceCor) of the observed to modeled climatological mean fields, by season.

5) Linear trend of annual values (Trend).

6) Time series variance (TimeVar) of temperature and coefficient of variation (TimeCV) of precipitation: calculated at frequencies ranging from 1 to 10 years.

7) Time series variance (TimeVar) of temperature and coefficient of variation (TimeCV) of precipitation of seasonal means.

8) Persistence (Hurst) measured using the Hurst exponent.

9) Strength of ENSO teleconnection (ENSO) in winter.

Also, but for temperature only, we calculated one additional metric:

10) Mean diurnal temperature range (DTR), by season.

The above metrics are identical to those in *Rupp et al*. [2013] but for the seventh metric listed, which was added after consultation with the USGS Southeast CSC. A full list of the metrics, along with the observational datasets used to evaluate each metric, is given in Table 1.

We evaluated most metrics as spatial averages over the entire Southeast US, defined here as the land area in Fig. 1. However, because the climate of the Southeast US is driven by larger scale oceanic and atmospheric patterns that we want to be faithfully simulated, the spatial variance and correlation metrics were examined over a larger domain (115°W – 50°W, 15°N – 55°N). This expanded domain covers a large portion of North America and the northwestern Atlantic.

Table 1. Definitions of performance metrics, the confidence in the metrics for model ranking, and observational datasets used by metric.

Metric[a]	Confidence category	Description	Observation datasets
Mean-T	Highest	Mean annual temperature (T) and precipitation (P), 1960-1999	CRU, PRISM, UDelaware, ERA40[d], NCEP[d]
Mean-P	Highest		
DTR-*MMM*[c]	Highest	Mean diurnal temperature range, 1950-1999	CRU[e], PRISM[e], NCEP
SeasonAmp-T	Highest	Mean amplitude of seasonal cycle as the difference between warmest and coldest month (T), or wettest and driest month (P). Monthly precipitation calculated as percentage of mean annual total, 1960-1999.	CRU, PRISM, UDelaware, ERA40[d], NCEP[d]
SeasonAmp-P	Higher		
SpaceCor-*MMM*-T[b,c]	Highest	Correlation of simulated with observed the mean spatial pattern, 1960-1999.	ERA40, NCEP[e]
SpaceCor-*MMM*-P[b,c]	Higher		
SpaceSD-*MMM*- T[b,c]	Highest	Standard deviation of the mean spatial pattern, 1960-1999. All standard deviations are normalized by the standard deviation of the observed pattern.	ERA40, NCEP[e]
SpaceSD-*MMM*-P[b,c]	Higher		
TimeVar.1-T	Lower	Variance of temperature calculated at frequencies (time periods of aggregation) ranging for $N = 1$ and 8 years, 1901-1999.	CRU, PRISM, UDelaware
TimeVar.8-T	Lowest		
TimeCV.1-P	Lower	Coefficient of variation (CV) of precipitation calculated at frequencies (time periods of aggregation) ranging for $N = 1$ and 8 water years, 1902-1999.	CRU, PRISM, UDelaware
TimeCV.8-P	Lowest		
TimeVar-*MMM*-T[c]	Lower	Variance of seasonal mean temperature, 1901-1999.	CRU, PRISM, UDelaware
TimeCV-*MMM*-P[c]	Lower	Coefficient of variation of seasonal mean precipitation, 1901-1999.	CRU, PRISM, UDelaware
Trend-T	Lower	Linear trend of annual temperature and precipitation, 1901-1999.	CRU, PRISM, UDelaware
Trend-P	Lowest		
ENSO-T	Lower	Correlation of winter temperature and precipitation with Niño3.4 index, 1901-1999.	CRU, PRISM, UDelaware
ENSO-P	Lowest		
Hurst-T	Lowest	Hurst exponent using monthly difference anomalies (T) or fractional anomalies (P), 1901-1999.	CRU, PRISM, UDelaware
Hurst-P	Lowest		

[a]Unless otherwise noted, metrics are average over Southeast US. [b]Expanded domain: 115°W – 50°W, 15°N – 55°N.

[c]*MMM* is the season designation: DJF, MAM, JJA, and SON.

[d]Temperature only used in ranking, not precipitation. [e]Not used in ranking.

We calculated several metrics (Mean, SeasonAmp, SpaceSD, SpaceCor) over the latter four decades of the 20th century (1960-1999), and DTR over 1950-1999, in order to include the shorter NCEP and ERA40 datasets in the analysis. However, those metrics that are more sensitive to record length (i.e., those that do not simply describe the mean state of the time series) were calculated over the 20th century (1901-1999) and consequently only for CRU, PRISM and UDelaware.

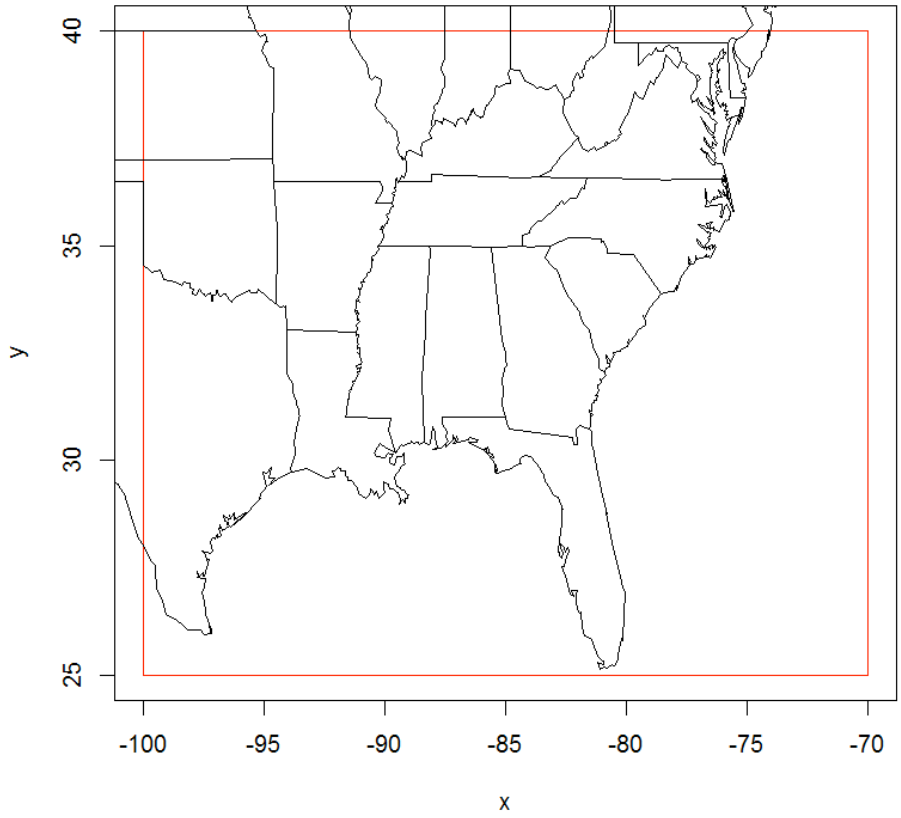


Figure 1. The Southeast US as defined in this study (ocean excluded).

In addition to calculating each metric for each ensemble member of each model, we also calculated a "multi-model mean" value. Given that models have different numbers of ensemble members, those models with larger ensembles will generally give better estimates of a particular statistic than those models with smaller ensembles. However, for simplicity, we gave each model equal weight when calculating a multi-model mean.

2.3. Model ranking by overall performance

A large number of metrics may help to elucidate the different strengths and weaknesses of models. At the same time, a suite of metrics also presents a challenge for selecting a subset of more credible models, for at least two reasons. For one, some metrics may be more relevant than others for a particular application, and the rankings may depend on which set of metrics are applied (e.g., *Santer et al*. [2009]). For another, there may be redundancy among metrics, given that not all are independent. In either case, treating all metrics equally might be inadvisable. We, therefore, applied two methods for ranking the models, as described below. The first simply treated all metrics equally, while the second did not.

The first method included all performance metrics and assigned equal weight to each metric. For a given model i and metric j, we defined an error $E_{i,j}$ as

$$E_{i,j} = \left| x_{obs,j} - x_{i,j} \right| \tag{1}$$

where x_{obs} and x_i are the observed and simulated ensemble mean metric, respectively. For x_{obs}, we used the mean of the ensemble of observations, where more than one observed dataset was examined. Application of Eq. (1) included correlations (where x_{obs} necessarily equaled 1). Furthermore, we defined a relative error $E^*_{i,j}$ as

$$E^*_{i,j} = \frac{E_{i,j} - \min(E_{i,j})}{\max(E_{i,j}) - \min(E_{i,j})} \tag{2}$$

and then summed the relative error across all *m* metrics:

$$E^*_{i,tot} = \sum_{j=1}^{m} E^*_{i,j} \tag{3}$$

to get the total relative error $E^*_{i,tot}$ per model. Ordering the models by their respective total relative error determined the ranking.

The second approach to ranking took into account both the redundancy in information among metrics and the confidence in the rankings of the individual metrics. To address the latter, we first excluded those metrics that were identified as not being robust. This exclusion of metrics is described in detail in Section 3.2 of *Rupp et al*. [2013]. Briefly, those metrics that show high *intra*-model (i.e., *intra*-ensemble) spread relative to *inter*-model spread were identified as not being robust metrics. We defined four categories of robustness, or confidence, in rankings: "highest", "higher", "lower", "lowest". Table 1 lists within which category each metric falls. Those metrics categorized as "lowest" were excluded from the following analysis. Also, so as not to so heavily weight those metrics calculated for each of four seasons, we used only the winter (DJF) and summer (JJA) values.

To address the matter of information redundancy, we conducted an empirical orthogonal function (EOF) analysis on the remaining metrics. This allowed us to reduce the large number of metrics, some of which co-vary and others of which add little information, down to a reduced number of orthogonal and more consequential metrics. We treated the observations just as if they were from another model, such that the EOF analysis was done on all values *x,* which include both observed (x_{obs}) and simulated (x_i) values for each metric. Note that for the EOF analysis we normalized the metric values by subtracting their mean and dividing by their standard deviation.

The leading EOFs provide a greatly reduced number of new orthogonal metrics, which can be examined separately. However, to arrive at a single metric from which to rank overall model performance, we simply calculated the Euclidean distance from the observations to each modeled value in EOF space across all dimensions of the leading EOFs. We used this distance as the overall error score per GCM, and normalized it to range from 0 (least error) to 1 (most error).

3. Results and Discussion

A discussion of the models' performance for each metric is provided in the Appendix. Here we focus on the results of the model ranking.

The ranking of models using the simple method on all 42 metrics is given in Fig. 2. Also shown are the relative errors for the individual metrics. Each model scored well in at least one metric, and there were several models that scored poorly in but a few metrics. Overall, the highest-ranked models include CNRM-CM5/CNRM-CM5-2 pair of models, the CESM1/CCSM4 family of models (with the exception of CESM1-WACCM), and the CMCC-CM/CMCC-CMS pair of models. Other high scoring models are MPI-ESM-LR, the "CC" versions of the GISS family of models, and HadGEM2-ES. There were some differences between the models that scored near the top for the Southeast US and those that scored best for the Pacific Northwest US in *Rupp et al*. [2013]. For example, while the CESM1/CCSM4 and CNRM-CM5 families fared well in both regions, the GISS family of models scored well in the Southeast but poorly overall in the Pacific Northwest.

From the EOF analysis on 22 of the full 42 metrics, we found that the leading 5 principal components (PC) cumulatively explained 22%, 40%, 52%, 62%, and 69% of the variance,

respectively. The models, ranked in order using the first 5 PCs, are shown in Fig. 3, upper panel. Fig. 3 also shows the effect of using just the first 2 or 4 PCs. In a few cases, using just the first 2 PCs makes a large change to a model's ranking (e.g., see CMCC-CESM in Fig. 3), which illustrates the sensitivity in ranking to how model metrics are weighted.

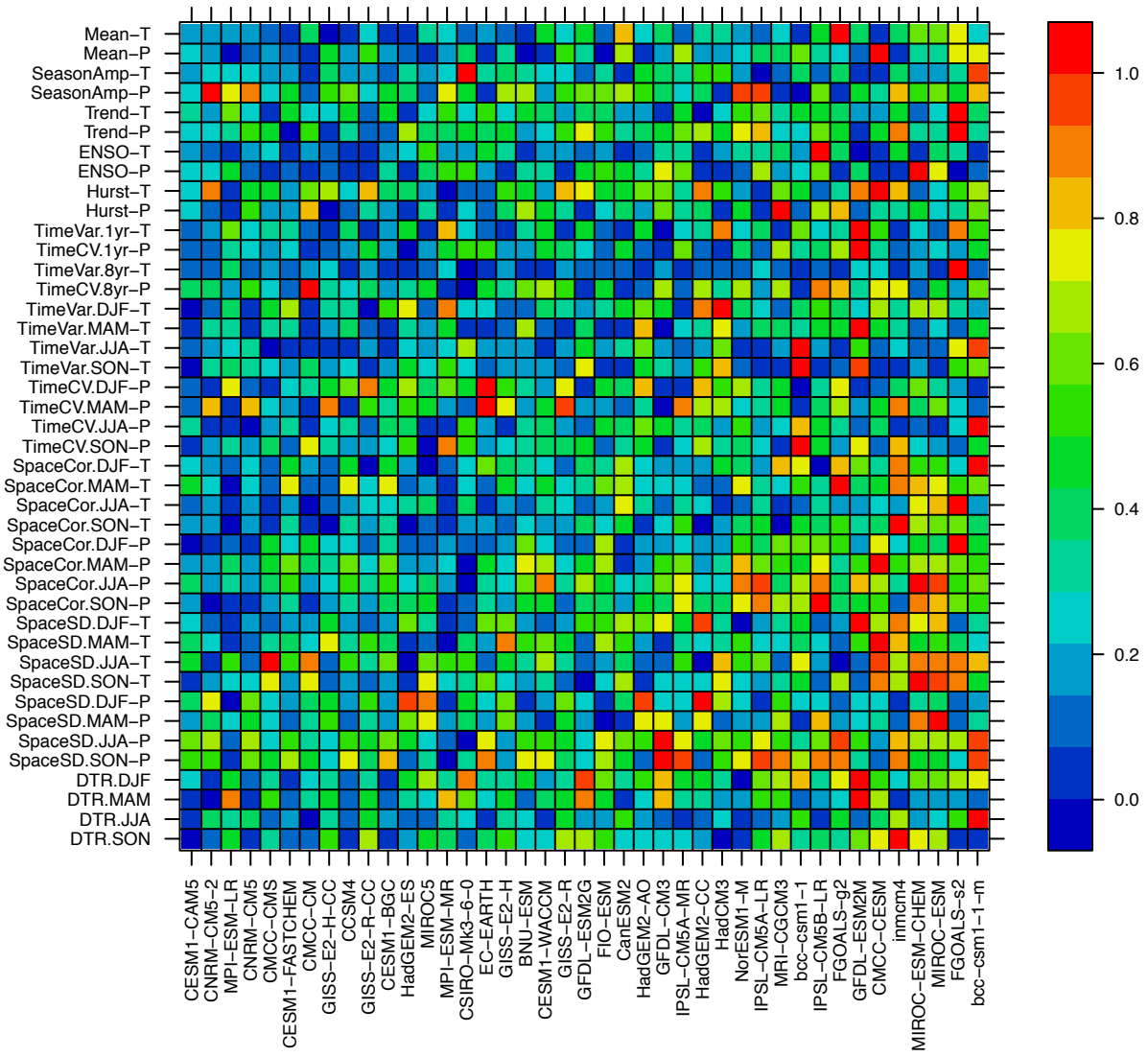


Figure 2. Relative error of the ensemble mean of each metric for each CMIP5 GCM. Models are ordered from least (left) to most (right) total relative error, where total relative error is the sum of relative errors from all metrics.

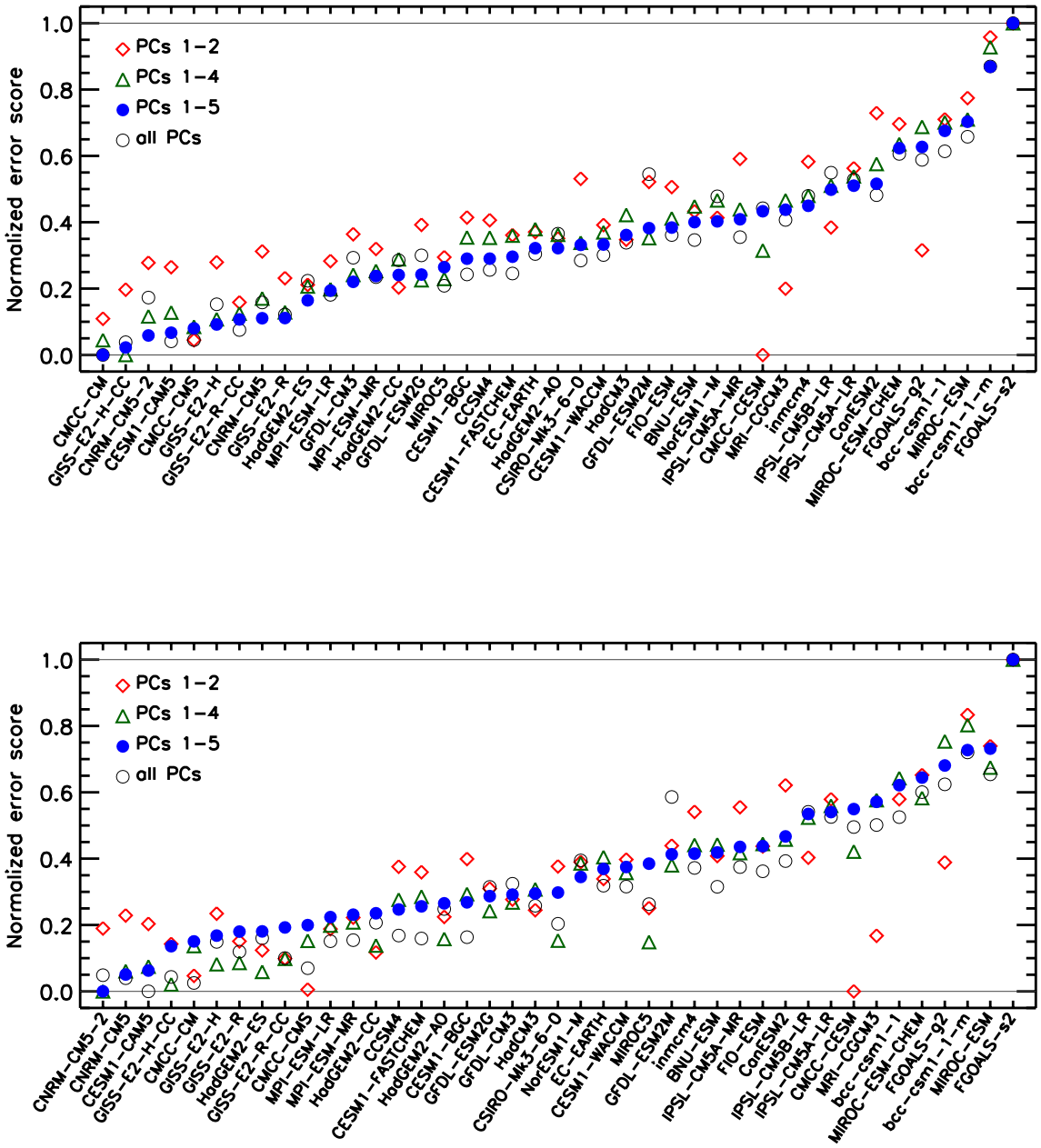


Figure 3. 41 CMIP5 GCMs ranked according to normalized error score from EOF analysis of performance metrics. Ranking is based on the first 5 principal components (filled blue circles). The open symbols show the models' error scores using the first 2, 4, and all 22 principal components (PCs). The best scoring model has a normalized error score of 0. Upper plot from using 22 metrics, lower plot from using the set of 18 metrics and identical observations datasets as *Rupp et al*. [2013].

To examine the influence of the additional metrics added for this evaluation of the Southeast US (TimeVar-DJF-T, TimeVar-JJA-T, TimeCV-DJF-P, TimeCV-JJA-P), we redid the EOF analysis with the same 18 metrics used in *Rupp et al*. [2013]. The inclusion of the additional metrics had only minor effect on the ranking of the models (compare upper and lower panels in Fig. 3). For example, CMCC-CM and GISS-E2-H-CC occupied the top 2 positions using 22 metrics, but were in 5th and 4th positions, respectively, when the set of 18 metrics were used.

Comparing the results from our initial simpler ranking to the more complex EOF-based analysis reveals minor differences. Nearly no models occupied precisely the same position in each method, but the general order was similar. For example, of the top 12 models resulting from the EOF method, 9 placed in the top 12 using the simple ranking method.

How the models scored in EOF "space" in shown in Fig. 4 for the 4 leading principal components (PC1 vs. PC2 and PC3 vs. PC4). The loadings, or weights, given to each metric within each of the leading 5 principal components is given in Table 2, where the metrics given the most weighting are highlighted. PC1 placed more weight on metrics related to temporal variability, though annual mean temperature and precipitation were also weighted heavily. PC2, in contrast, was more controlled by metrics quantifying the larger scale spatial patterns, similar to that seen by *Rupp et al*. [2013] for the Pacific Northwest.

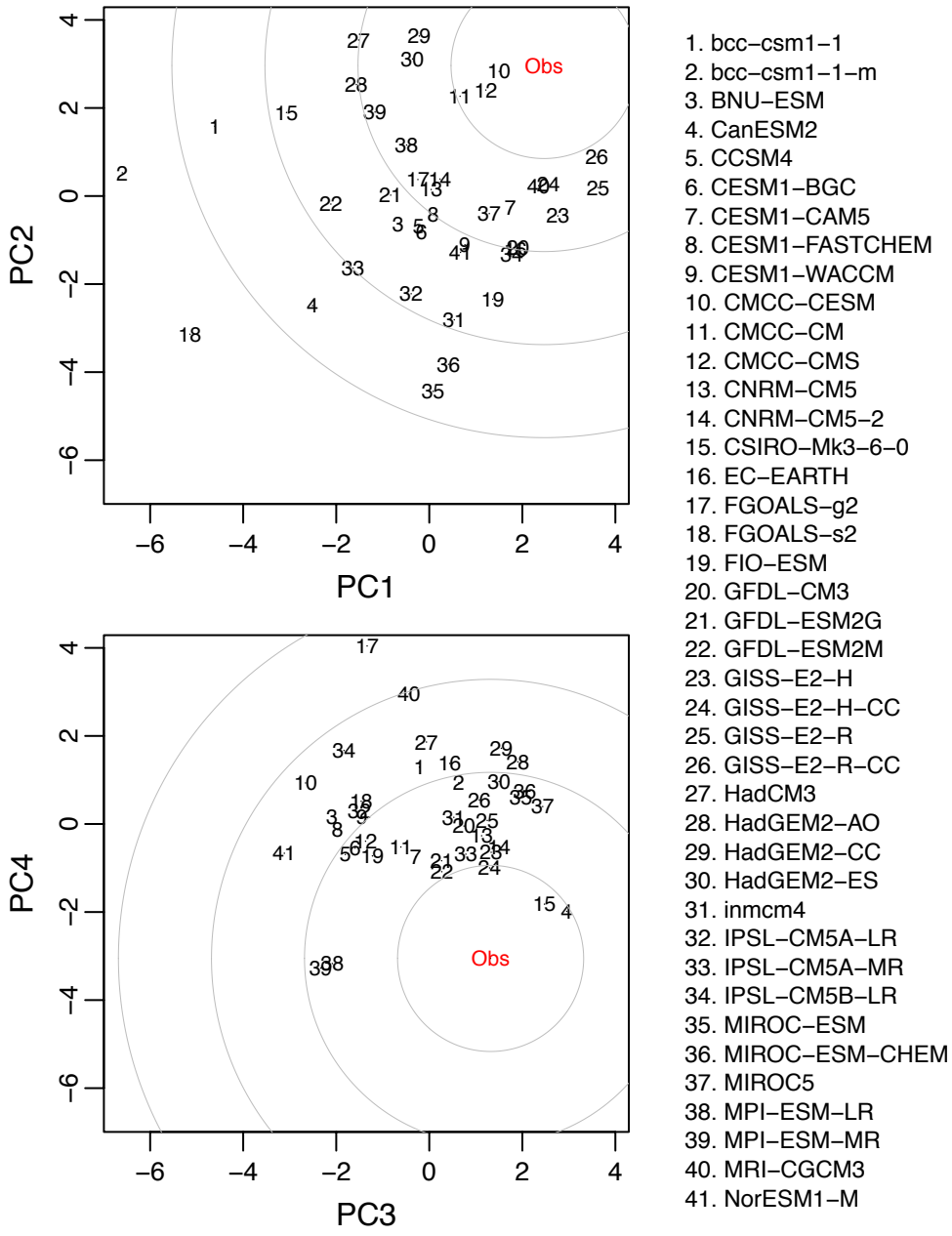


Figure 4. Loadings of the first four principal components (PC1, PC2, PC3, PC4) from EOF analysis of 22 evaluation metrics and 41 CMIP5 GCMs. "Obs" indicates the observation dataset.

Table 2. Loadings by metric of the leading 5 PCs. Absolute values of loadings greater than 0.20 and 0.30 have been shaded in yellow and red, respectively.

Metric	PC 1	PC 2	PC 3	PC 4	PC 5
Mean-T	0.34	0.20	-0.05	0.12	0.17
Mean-P	-0.33	0.19	0.12	0.16	-0.09
SeasonAmp-T	-0.08	-0.17	-0.04	-0.26	-0.12
SeasonAmp-P	-0.18	-0.24	-0.35	0.00	-0.01
Trend-T	0.06	-0.20	0.11	0.10	0.14
ENSO-T	-0.28	0.19	-0.27	0.00	0.14
TimeVar.1-T	-0.31	0.05	0.12	-0.36	0.13
TimeCV.1-P	-0.10	0.30	-0.34	0.14	-0.06
TimeVar-DJF-T	-0.16	-0.04	-0.29	-0.45	-0.07
TimeCV-DJF-P	-0.40	0.12	0.07	0.00	0.01
TimeVar-JJA-T	-0.31	0.15	0.33	-0.01	0.04
TimeCV-JJA-P	0.19	0.15	0.01	-0.29	0.22
DTR-DJF	0.15	0.28	-0.19	-0.10	-0.27
DTR-JJA	0.12	0.14	0.34	-0.31	-0.28
SpaceCor-DJF-T	0.08	0.28	0.22	-0.21	-0.26
SpaceCor-JJA-T	-0.07	0.35	-0.20	0.22	-0.07
SpaceCor-DJF-P	-0.10	0.39	-0.07	-0.02	0.10
SpaceCor-JJA-P	0.07	0.20	0.38	0.21	0.05
SpaceSD-DJF-T	0.08	0.28	-0.03	-0.30	0.34
SpaceSD-JJA-T	0.16	-0.01	-0.13	-0.17	-0.55
SpaceSD-DJF-P	-0.32	-0.09	0.13	0.18	-0.37
SpaceSD-JJA-P	0.34	0.20	-0.05	0.12	0.17

Sensitivity to observational dataset

It is important to note that we have not exhaustively quantified the sensitivity of the rankings to the choice of observational dataset, but simply averaged over observational datasets when more than one dataset was considered. However, we expect that the effect of observational dataset selection is not insignificant given that the spread among observational datasets is comparable to inter-model differences for a few of the metrics. As discussed in the Appendix, the noteworthy metrics in this regard are the mean annual precipitation (Mean-P; Fig. A1), the amplitude of the seasonal cycle of precipitation (SeasonAmp-P; Fig. A2), the diurnal temperature range, most notably in summer (DTR-JJA; Fig. A5), and the spatial pattern of precipitation, again most notably in summer (SpaceCorr-JJA-P, SpaceVar-JJA-P; Fig. A8). By far the largest differences occur between the station-based datasets (CRU, PRISM, and UDelaware) and the reanalysis datasets (ERA40 and NCEP), though sizable discrepancies also exist between ERA40 and NCEP for summer precipitation (Figs. A2 and A8). If we assume that the station-based datasets provide more reliable estimates of precipitation, this raises the question of whether reanalysis datasets should be used to evaluate precipitation from climate models in regions where there exist gridded station-based data of high quality (i.e., high station density and long records). In fact, in this study we departed from *Rupp et al*. [2013] in that we excluded ERA40 and NCEP when taking the average of observation datasets for the regionally-averaged metrics Mean-P and SeasonAmp-P, relying solely on CRU, PRISM, and UDelaware. However, we still used ERA40 and NCEP for evaluating large-scale spatial patterns that cover both land and ocean.

In regards to DTR, it is worth noting that near-coast grid cells of the Southeast US domain contained some influence of ocean cells in both the models and in NCEP, unlike CRU and PRISM. We might expect this ocean influence to have suppressed DTR somewhat in both

the models and NCEP, as compared to CRU and PRISM; in fact, DTR from the NCEP and the multi-model ensemble mean was 2-3°C lower than CRU and PRISM. *Rupp et al*. [2013] noted this, and as a test, removed the coastal grid cells from their domain and recalculated the regionally averaged DTR. While they found that this slightly increased DTR across all datasets, it only negligibly affected the relative values of DTR, thus the ranking of models based on DTR alone did not change. Also, the discrepancy between NCEP and CRU/PRISM remained unchanged, implying that the differences among observational datasets were not related to the influence of ocean cells. We conducted a similar test for the Southeast region, excluding the 1°x1° grid cells near the Atlantic coastline and all grid cells south of 32°, therefore excluding Florida and southern Texas entirely. Similar to *Rupp et al*. [2013], we found that while the reduced domain showed a slightly increased DTR, the pattern shown by the models and observational datasets in Fig. A5 showed no marked difference.

This implies other causes for the large differences between the station- and reanalysis-based values of DTR. One factor may be that observations of Tmin and Tmax are relatively instantaneous, whereas simulated Tmin and Tmax have been averaged over some time step that varies by GCM and reanalysis dataset, and therefore are effectively biased towards lower DTR. If this time-average smoothing is a major factor, then it is flawed to expect the modeled values to match the station-based instantaneous values; in effect, the reanalysis becomes a better reference against which to judge a model's ability to simulate DTR. In fact, when determining the above rankings, we used NCEP as the observational dataset to calculate the error in DTR, unlike *Rupp et al.* [2013], who used the average of NCEP, CRU, and PRISM.

Appendix. CMIP5 models' skill by performance metric

A.1. Climatologic mean

The simulated mean annual temperature of the Southeast US differed by 7°C from the coolest to the warmest model (Fig. A1, upper panel), though the coolest model was a full 2°C cooler than the next coolest model. The observation datasets differed only slightly amongst themselves, with a range of about 0.5°C between the warmest and coolest. Taken as an average, the 5 observation datasets were less than 0.5°C warmer than the median of the simulated mean annual temperatures. (Note: henceforth the observed values will be reported as the average of the observation datasets used, unless specifically stated otherwise).

For mean annual precipitation, the range across models was large: 65-cm $year^{-1}$ difference between the wettest and driest model (Fig. A1, lower panel). However, the observation datasets differed greatly, with a range of about 45 cm $year^{-1}$. The three station-based datasets (CRU, PRISM, and UDeleware) were comparably similar, with the reanalysis datasets giving the most (NCEP) and least (ERA40) precipitation of the 5 observations datasets. The average of the 3 station-based datasets were only about 5 cm $year^{-1}$ greater than the median of the simulated mean annual precipitation. Implications of the discrepancies among the observational datasets are discussed in Section 3.

A.2. Seasonal cycle

All the models reproduced the phase and general shape of the seasonal cycle of temperature (Fig. A2, upper panel), though the amplitude of the seasonal cycle varied widely among models (Fig. A3), ranging from 19.4°C to 28.9°C. The median of the modeled amplitude was 22.5°C, which was within 1°C of the observed amplitude.

Most models generated the general pattern of increased precipitation in the months of March through August that is evident in CRU, PRISM, and UDelaware (Fig. A2). Interestingly, large differences exist between the 3 station-based datasets and 2 reanalysis datasets; both ERA40 and NCEP show a strong seasonal cycle of dry winter and wet summer. This seasonal cycle in the reanalysis is both exaggerated in amplitude and shifted in phase by a few months with respect to the station-based datasets (and, interestingly, also to the overall pattern of the CMIP5 models).

Calculating mean monthly precipitation as a percentage of the mean annual total precipitation, simulated seasonal precipitation amplitude ranged from as small as 2.5% to as large as 7.2%. In comparison, the 3 station-based datasets gave precipitation amplitude of about 3%, implying the models on average are exaggerating the strength of the seasonal cycle. (Note that with mean monthly precipitation calculated as a percentage of annual, the percentages reported above are the differences of percent precipitation between the wettest and driest months).

Though the statistics above were averaged over the entire Southeast US, there are regional spatial gradients in the amplitudes of the seasonal cycle of both temperature and precipitation. Visual inspection showed good agreement between the observed (CRU and ERA40) and the multi-model mean spatial pattern of the temperature cycle amplitude (Fig. A4, left panels). The models as a whole accurately reproduced the strong coast-to-interior gradient.

Most of the Southeast US has a relatively weak seasonal-cycle amplitude in precipitation, with increased magnitude in Florida and the most western part of the SE region. The multi-model mean reproduces this spatial pattern, but to a lesser extent than the observational record (Fig. A4, right panels), and, not surprisingly, lacks some finer-scale features. In general, the

multi-model mean generates an amplitude that is too large over most of the region compared to CRU, and underestimates the amplitude over Florida and western part of the region.

A.3. Diurnal temperature range

The Southeast region as a whole exhibited small variations in the mean diurnal temperature range (DTR) throughout the year, with two maxima in April and October, and two minima in January and July. This feature was present in both the station-based and reanalysis datasets (Fig. A5). Some models reproduced this pattern, though others showed a unimodal cycle with a maximum in summer and minimum in winter.

Simulated mean DTR ranged from 6.2°C to 11.5°C in winter and from 7.3°C to 15.2°C in summer. For winter, DTR from all models was lower than observed DTR from CRU and PRISM; for summer, DTR from all but 3 models was lower than observed DTR from CRU and PRISM. In contrast, the multi-model mean simulated DTR for both seasons was more consistent with NCEP. Implications of the discrepancy between the station-based and reanalysis datasets are discussed in Section 3.

A.4. Large-scale spatial patterns

The multi-model mean temperature field over North America and northwest Atlantic accurately reproduced the ERA40 climatological fields (Fig. A7), with correlation coefficients (r) of 0.997 and 0.9741 in winter and summer, respectively (for NCEP, r = 0.998 and 0.983, respectively). Much of this high correlation resulted from simply matching the general latitudinal temperature gradient, but other continental features of the climatological fields were also reproduced. Individually, all models were very highly correlated to observations in winter ($0.982 \leq r \leq 0.996$) and highly correlated in summer ($0.84 \leq r \leq 0.98$). The variances of the modeled fields were also similar to the observed variance, though more so in winter, when all

standard deviations were within ±15% of the observed standard deviations. In summer, all models were within ±25% of observations (Fig. A9, upper panel).

The multi-model mean precipitation field over North America and northwest Atlantic generally reproduced the main large-scale climatological features of the ERA40 field in winter (r = 0.70), though was much less faithful in summer (r = 0.60) (Fig. A8) (for NCEP, r = 0.75 and 0.48, respectively). Individually, the spatial patterns of most, but not all, models correlated reasonable well with the spatial pattern of ERA40 precipitation in winter ($0.49 \leq r \leq 0.89$) while the correlations weakened in spring and fall, and were weakest in summer ($0.00 \leq r \leq 0.71$). Normalized standard deviations ranged from about 0.5 to 1.5 across all simulations and all seasons, with a large majority of models simulating too much spatial variability in winter and spring and too little variability in summer and fall (Fig. A9, lower panel).

A.5. 20th century trend

The average annual temperature in the Southeast region decreased during the 20th century by an estimated 0.27°C, calculated as the average from CRU, PRISM and UDelaware. Of 41 CMIP5 models, only 2 produced a decreasing trend; the multi-model mean trend was positive 0.60°C over the 20th century, with models ranging from 0.06°C to 1.96°C (Fig. A10, upper panel). This decreasing trend in observed temperatures over the region, known as the "warming hole", is discussed in the context of CMIP5 results by *Kumar et al.* [2013].

The linear trend in observed regional mean annual precipitation was +7% over the 20th century. Models produced ensemble-average trends ranging from -5% to +8% per century, while only 2 individual ensemble members from 2 GCMs exceeded the observed +7% per century (Fig. A10, lower panel). The multi-model mean trend in annual precipitation was +1.5% per century.

A.6. Temporal variability

Overall, the CMIP5 models tended to produce too much interannual variability in regionally averaged times series of temperature relative to the observations, though this bias is not large and the overall bias decreased as the temporal aggregation increased from annual to decadal (Fig. A11, upper panel). At the annual scale, simulated standard deviations ranged by a factor of about 2, from 0.43°C to 0.83°C (Fig. A12). At the octadal (i.e., 8-year) scale, simulated values ranged from 0.17°C to 0.68°C, or a factor of 4.

In case of precipitation, the observed annual variability was similar to the mean of the variability from all models (Fig. A11, lower panel), with the coefficient of variation (CV) ranging from 0.07 to 0.15 for simulated annual precipitation (Fig. A13, upper panel). Though both the simulations and observations showed apparent power-law scaling of the CV, the simulated CVs in general decreased too rapidly with increasing scale. A consequence is that by the octadal scale, most of the models were generating too little variability (Fig. A13, lower panel).

Separated by season, the above characteristics in year-to-year variability remain similar: small overall bias in standard deviation of temperature and coefficient of variability of precipitation (see Figs. A14 and A15). What is particularly notable, however, is that some models rank very differently across seasons. For example, of all models, NorESM1-M has the lowest CV of precipitation in summer but the second highest CV in winter.

A.7. Long-term persistence

The Hurst exponent [*Hurst*, 1951] of the observed temperature anomalies ranged from 0.70 to 0.73, depending on the dataset (CRU, PRISM, or UDelaware). Though the causes of observed Hurst exponent are not explored here, these values could, for example, indicate long-

term memory or non-stationarity in the mean [*Klemes*, 1974]. In either case, the Hurst exponent > 0.5 implies that the processes that determine temperature over the region occur over a wide range of scales [*Tessier et al*., 1996]. The mean Hurst exponent averaged over all models was 0.68. Individual simulations showed Hurst exponents all greater than 0.5 ($0.60 \leq H \leq 0.79$) with 90% of values falling between 0.63 and 0.75 (Fig. A16, upper panel).

The estimated Hurst exponent of the observed precipitation anomalies was 0.64 for all three datasets, and slightly less than that for temperature. The mean simulated Hurst exponent was, remarkably, also 0.64, with 90% of values falling between 0.55 and 0.68 (Fig. A17, upper panel) and all were greater 0.5.

A.8. ENSO teleconnections

Consistent with observations, a negative regional temperature response to ENSO was apparent in the models: all but 2 models had a negative response of winter (JFM) temperature to ENSO (Fig. A18, upper panel). The multi-model mean response was a 0.43°C decrease in winter temperature for every 1°C increase in the Niño3.4 index, which is slightly weaker than the observed response of -0.58 °C $°C^{-1}$. The agreement in the spatial pattern of the ENSO response is remarkable, though the location of the observed transition from negative to positive temperature response over the US occurs 2-3° latitude southward of the multi-model mean transition (Fig. A19, upper panels).

A precipitation response to ENSO was also apparent in the simulations, with all models showing increased JFM precipitation with warmer tropical Pacific temperatures (Fig. A19, lower panel). The multi-model mean response was 7.4% $°C^{-1}$, compared with the observed response of 3.7% $°C^{-1}$. The spatial patterns of the observed and mean simulated ENSO precipitation response were generally similar, though the CRU observations showed a negative precipitation

response over the Appalachians, which was not apparent in the multi-model mean (Fig. 19, lower panels). Because the precipitation response to ENSO varied in sign across the Southeast region (as given by CRU), the regionally averaged response has limitations as a performance metric.

Table A1. CMIP5 models used in this study and some of their attributes.

Model	Center	Number of ensemble members: T/ P/ Tmin/ Tmax/	Atmospheric resolution (lon. x lat.)	Vertical levels in atmosphere
BCC-CSM1-1	Beijing Climate Center, China Meteorological Administration	3/ 3/ 3/ 3	2.8x2.8	26
BCC-CSM1-1-M	Beijing Climate Center, China Meteorological Administration	3/ 3/ 3/ 3	1.12x1.12	26
BNU-ESM	College of Global Change and Earth System Science, Beijing Normal University, China	1/ 1/ 1/ 1	2.8x2.8	26
CanESM2	Canadian Centre for Climate Modeling and Analysis	5/ 5/ 5/ 5	2.8x2.8	35
CCSM4	National Center of Atmospheric Research, USA	6/ 6/ 6/ 6	1.25x0.94	26
CESM1-BGC	Community Earth System Model Contributors	1/ 1/ 1/ 1	1.25x0.94	26
CESM1-CAM5	Community Earth System Model Contributors	3/ 3/ 3/ 3	1.25x0.94	26
CESM1-FASTCHEM	Community Earth System Model Contributors	3/ 3/ 3/ 3	1.25x0.94	26
CESM1-WACCM	Community Earth System Model Contributors	1/ 1/ 1/ 1	2.5x1.89	66
CMCC-CESM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	3.75x3.71	39
CMCC-CM	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	0.75x0.75	31
CMCC-CMS	Centro Euro-Mediterraneo per I Cambiamenti Climatici	1/ 1/ 1/ 1	1.88x1.87	95
CNRM-CM5	National Centre of Meteorological Research, France	10/ 10/ 10/ 10	1.4x1.4	31
CNRM-CM5-2	National Centre of Meteorological Research, France	1/ 1/ 1/ 1	1.4x1.4	31
CSIRO-Mk3-6-0	Commonwealth Scientific and Industrial Research Organization/Queensland Climate Change Centre of Excellence, Australia	10/ 10/ 10/ 10	1.8x1.8	18
EC-EARTH	EC-EARTH consortium	5/ 7/ 4/ 4	1.13x1.12	62
FGOALS-g2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	5/ 5/ 5/ 5	2.8x2.8	26

FGOALS-s2	LASG, Institute of Atmospheric Physics, Chinese Academy of Sciences	3/ 3/ 3/ 3	2.8x1.7	26
FIO-ESM	The First Institute of Oceanography, SOA, China	3/ 3/ 3/ 3	2.81x2.79	26
GFDL-CM3	NOAA Geophysical Fluid Dynamics Laboratory, USA	5/ 5/ 5/ 5	2.5x2.0	48
GFDL-ESM2G	NOAA Geophysical Fluid Dynamics Laboratory, USA	3/ 3/ 1/ 1	2.5x2.0	48
GFDL-ESM2M	NOAA Geophysical Fluid Dynamics Laboratory, USA	1/ 1/ 1/ 1	2.5x2.0	48
GISS-E2-H	NASA Goddard Institute for Space Studies, USA	5/ 5/ 5/ 5	2.5x2.0	40
GISS-E2-H-CC	NASA Goddard Institute for Space Studies, USA	1/ 1/ 1/ 1	2.5x2.0	40
GISS-E2-R	NASA Goddard Institute for Space Studies, USA	3/ 3/ 3/ 3	2.5x2.0	40
GISS-E2-H-CC	NASA Goddard Institute for Space Studies, USA	1/ 1/ 1/ 1	2.5x2.0	40
HadCM3	Met Office Hadley Center, UK	10/ 10/ 10/ 10	3.75x2.5	19
HadGEM2-AO	Met Office Hadley Center, UK	1/ 1/ 1/ 1	1.88x1.25	38
HadGEM2-CC	Met Office Hadley Center, UK	1/ 1/ 1/ 1	1.88x1.25	60
HadGEM2-ES	Met Office Hadley Center, UK	5/ 5/ 5/ 5	1.88x1.25	38
INMCM4	Institute for Numerical Mathematics, Russia	1/ 1/ 1/ 1	2.0x1.5	21
IPSL-CM5A-LR	Institut Pierre Simon Laplace, France	6/ 6/ 1/ 1	3.75x1.8	39
IPSL-CM5A-MR	Institut Pierre Simon Laplace, France	3/ 3/ 1/ 1	2.5x1.25	39
IPSL-CM5B-LR	Institut Pierre Simon Laplace, France	1/ 1/ 1/ 1	3.75x1.8	39
MIROC5	Atmosphere and Ocean Research Institute (The University of Tokyo), National Institute for Environmental Studies, and Japan Agency for Marine-Earth Science and Technology	5/ 5/ 5/ 5	1.4x1.4	40
MIROC-ESM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	3/ 3/ 3/ 3	2.8x2.8	80
MIROC-ESM-CHEM	Japan Agency for Marine-Earth Science and Technology, Atmosphere and Ocean Research Institute (The University of Tokyo), and National Institute for Environmental Studies	1/ 1/ 1/ 1	2.8x2.8	80
MPI-ESM-LR	Max Planck Institute for Meteorology, Germany	3/ 3/ 3/ 3	1.88x1.87	47

MPI-ESM-MR	Max Planck Institute for Meteorology, Germany	3/ 3/ 3/ 3	1.88x1.87	95
MRI-CGCM3	Meteorological Research Institute, Japan	5/ 5/ 5/ 5	1.1x1.1	48
NorESM1-M	Norwegian Climate Center, Norway	3/ 3/ 3/ 3	2.5x1.9	26

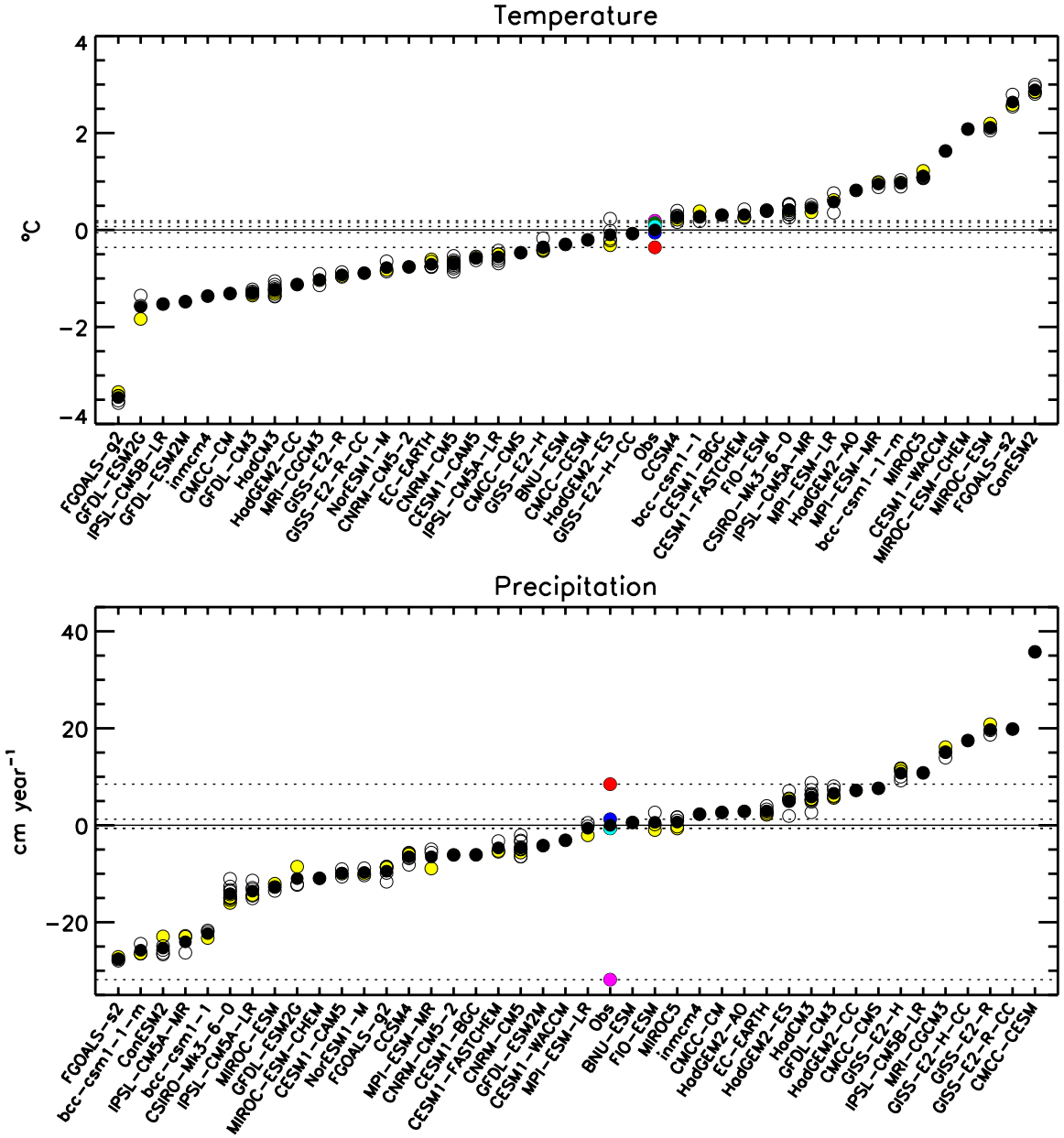


Figure A1. Mean annual temperature and precipitation bias for 41 CMIP5 GCMs averaged over the Southeast US. For each GCM, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from NCEP (red), ERA40 (magenta), CRU (dark green), PRISM (blue), UDelaware (Cyan), and average of observations (black). For precipitation, only CRU, PRISM and UDelaware are calculated in the average.

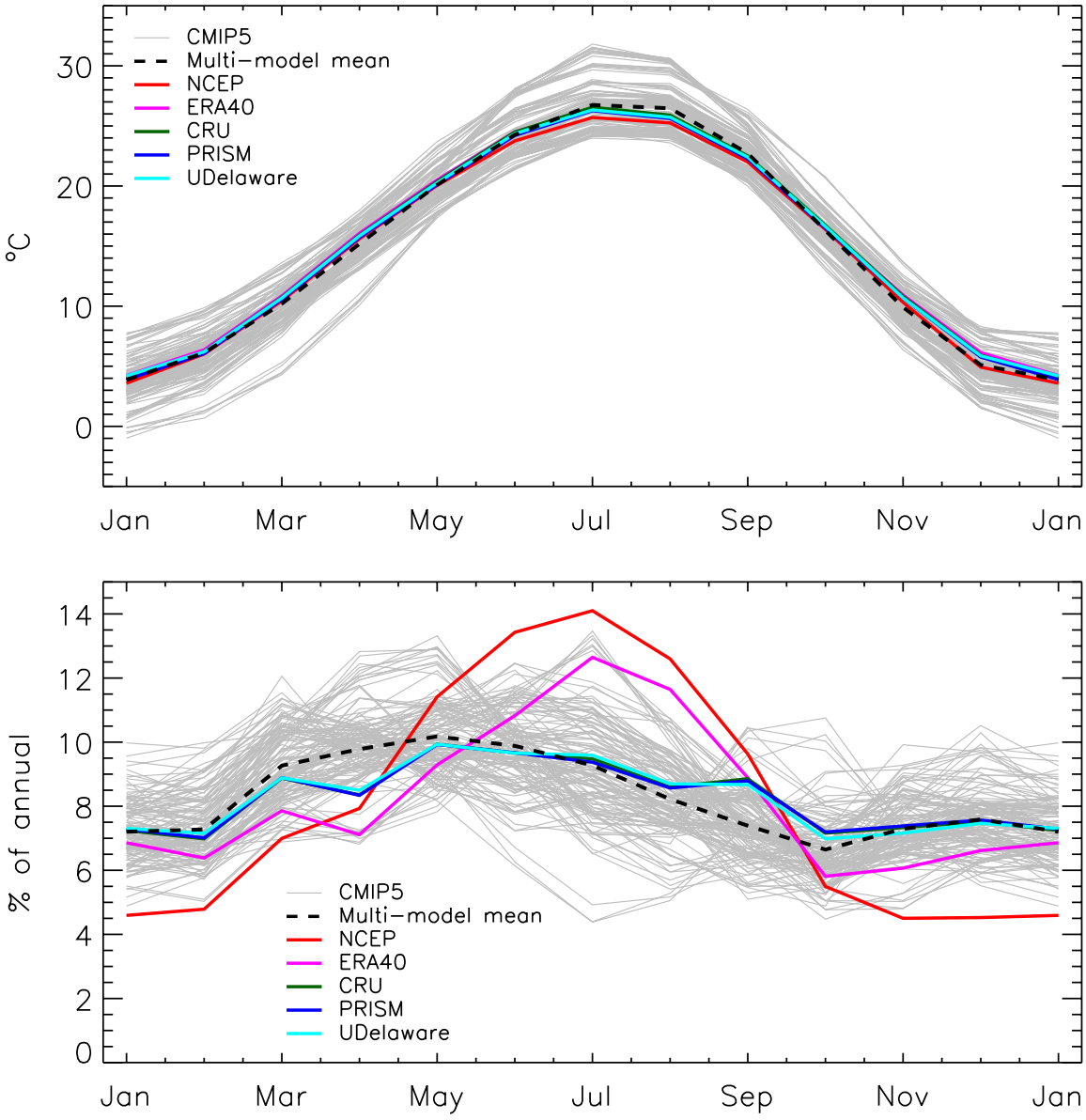


Figure A2. Mean seasonal cycle of temperature (upper panel) and relative precipitation (lower panel) averaged over the Southeast US. Monthly means are calculated from gridded observation datasets (NCEP, ERA40, CRU, PRISM, UDelaware) and from all ensemble members from 41 CMIP5 GCMs.

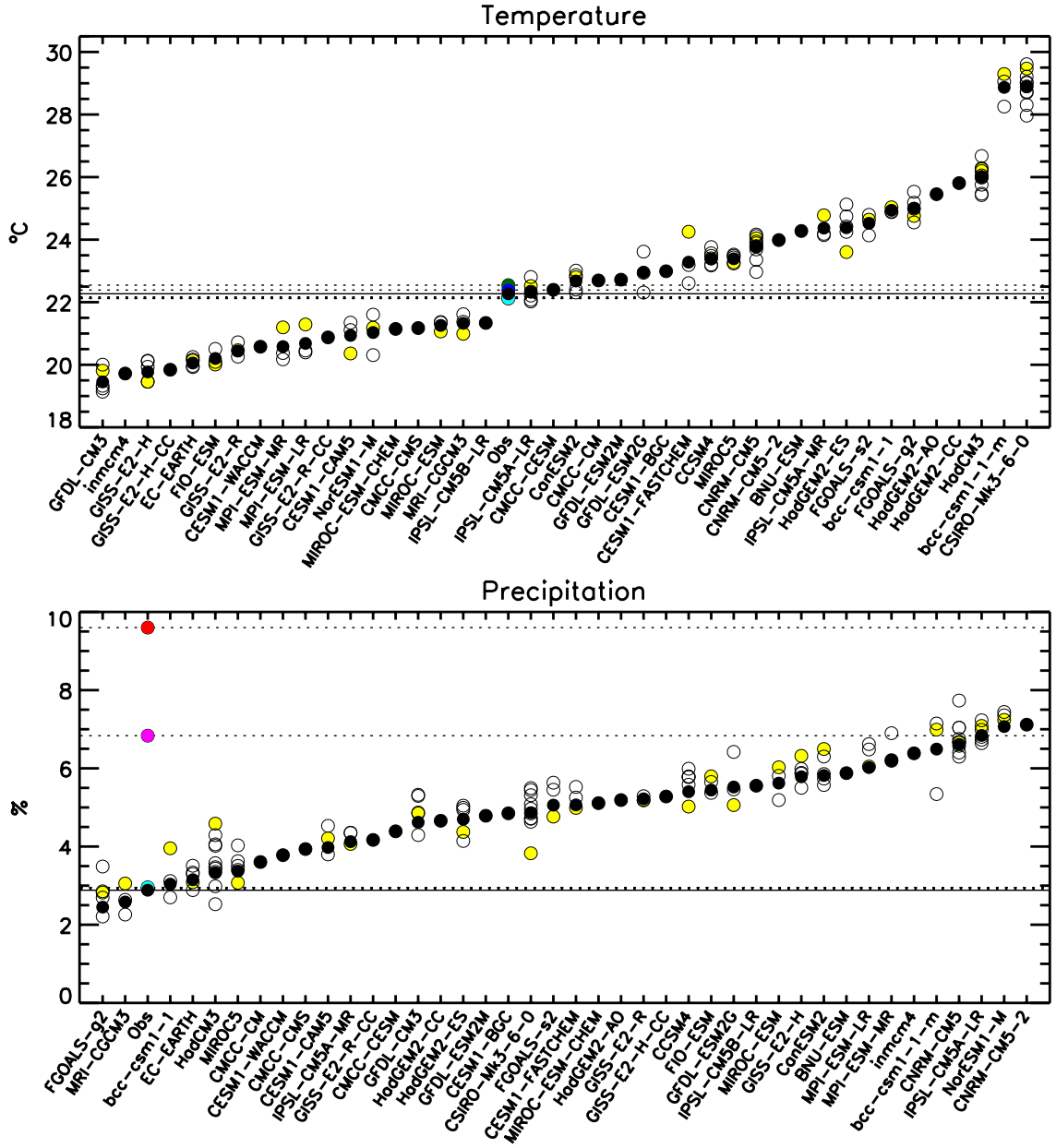


Figure A3. Mean seasonal cycle amplitude in temperature and relative precipitation for the Southeast US. For each CMIP5 GCM, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from NCEP (red), ERA40 (magenta), CRU (dark green), PRISM (blue), UDelaware (Cyan), and average of observations (black). Monthly precipitation is calculated as a percentage of the mean annual total, so the amplitude is the difference of percentages. For precipitation, only CRU, PRISM and UDelaware are calculated in the average.

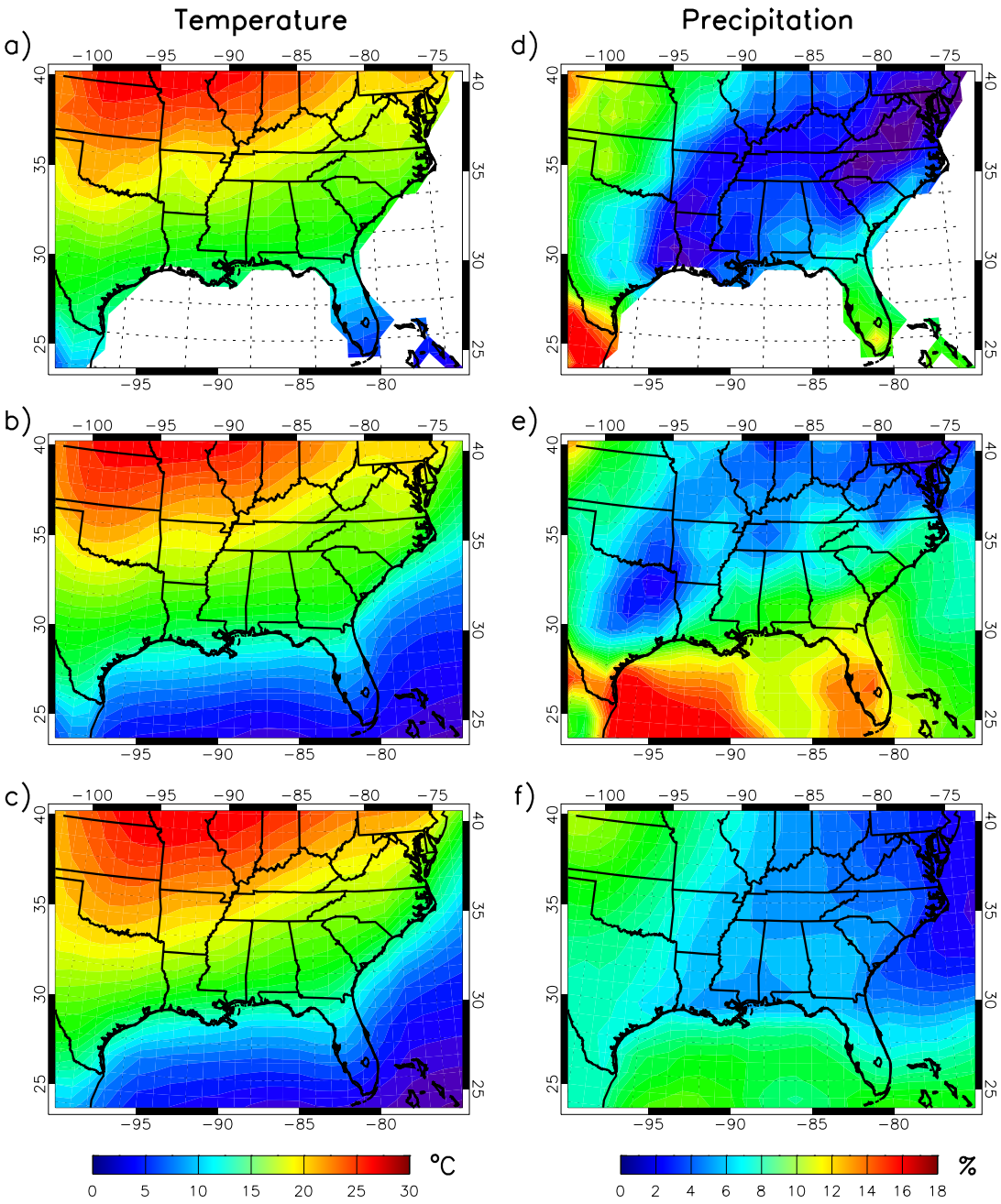


Figure A4. Mean seasonal cycle amplitude of temperature from CRU (a), ERA40 (b) and the CMIP5 multi-model mean (c), and mean season cycle amplitude of relative precipitation from CRU (d), ERA40 (e) and the CMIP5 multi-model mean (f).

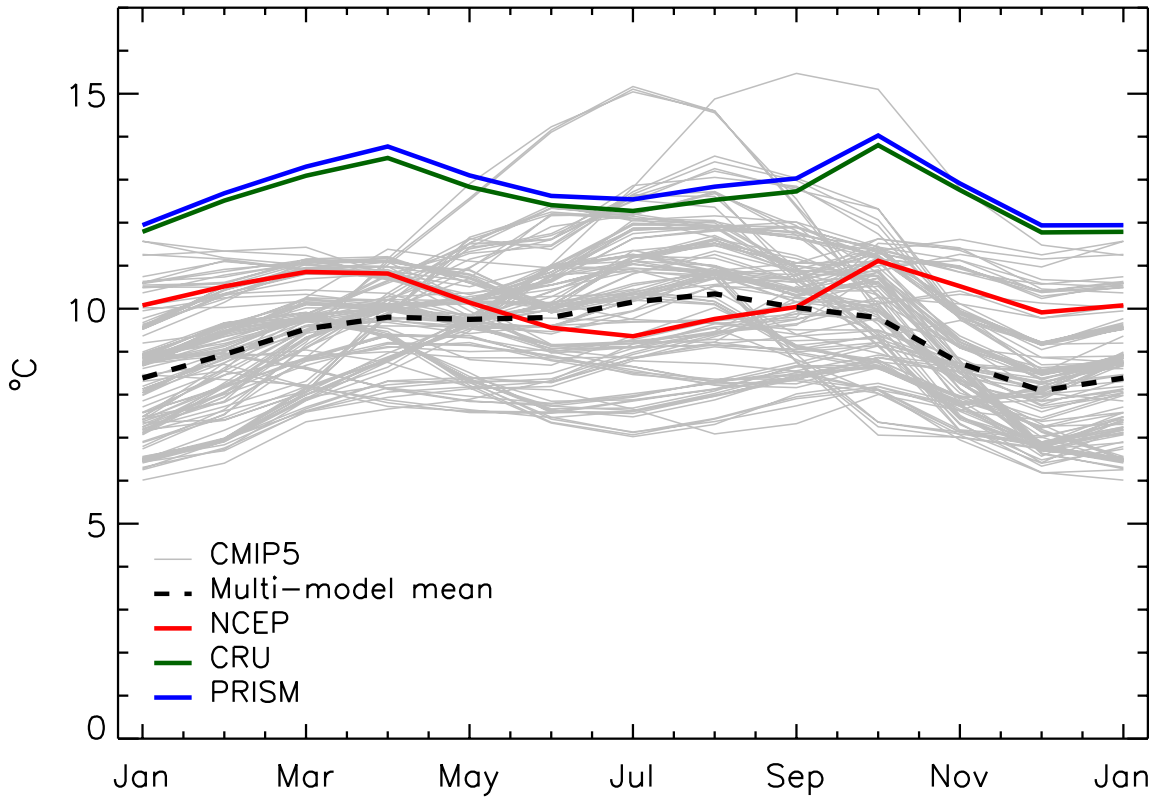


Figure A5. Mean seasonal cycle of diurnal temperature range averaged over the Southeast US. Monthly means are calculated from gridded observation datasets (NCEP, CRU, and PRISM) and from all ensemble members from 41 CMIP5 GCMs.

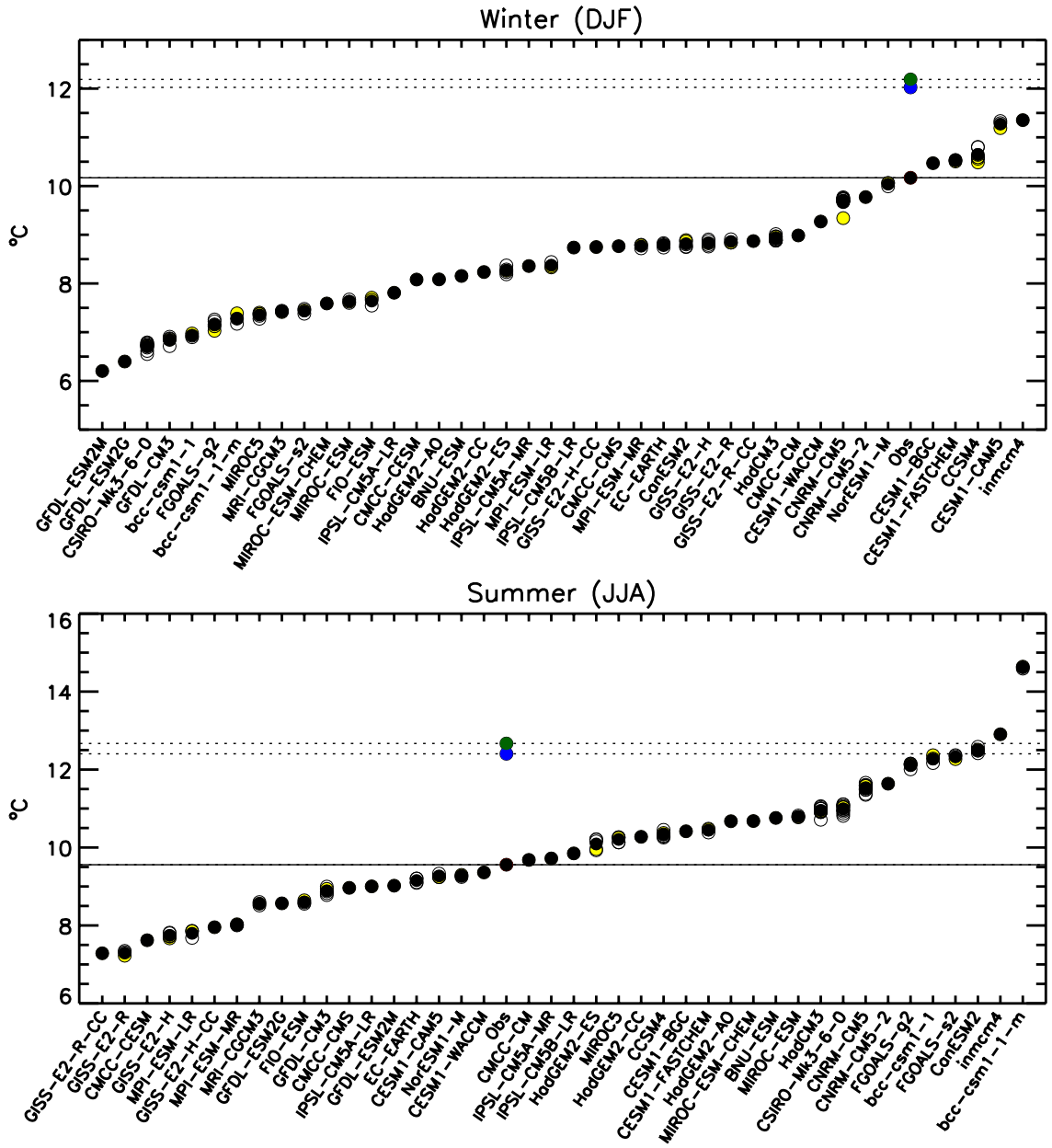


Figure A6. Mean diurnal temperature range (DTR) in winter (DJF) and summer (JJA) averaged over the Southeast US. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), and PRISM (blue), and NCEP (black).

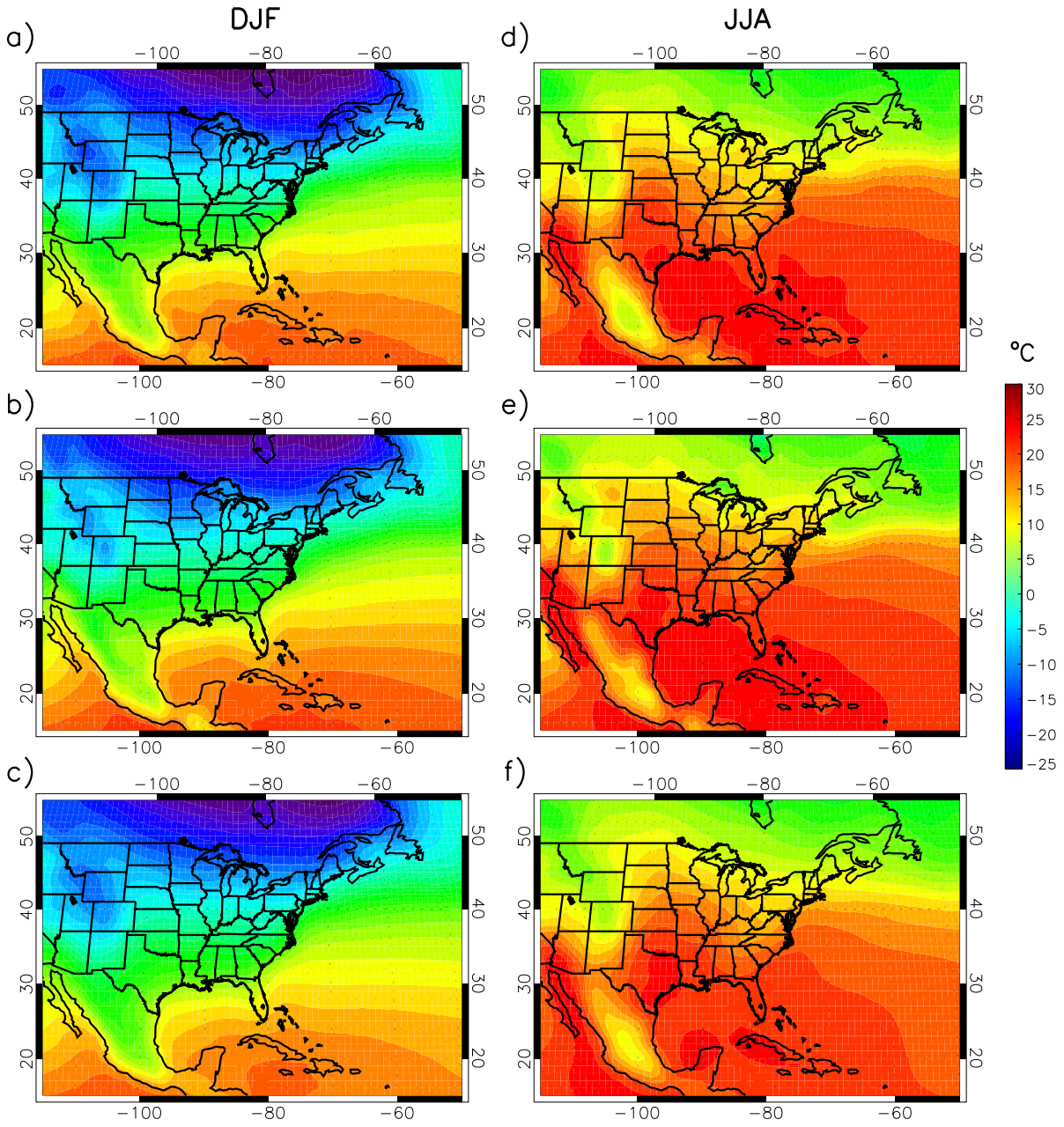


Figure A7. Mean winter (DJF) temperature from NCEP (a), ERA40 (b) and the CMIP5 multi-model mean (c), and mean summer (JJA) temperature from NCEP (d), ERA40 (e) and the CMIP5 multi-model mean (f).

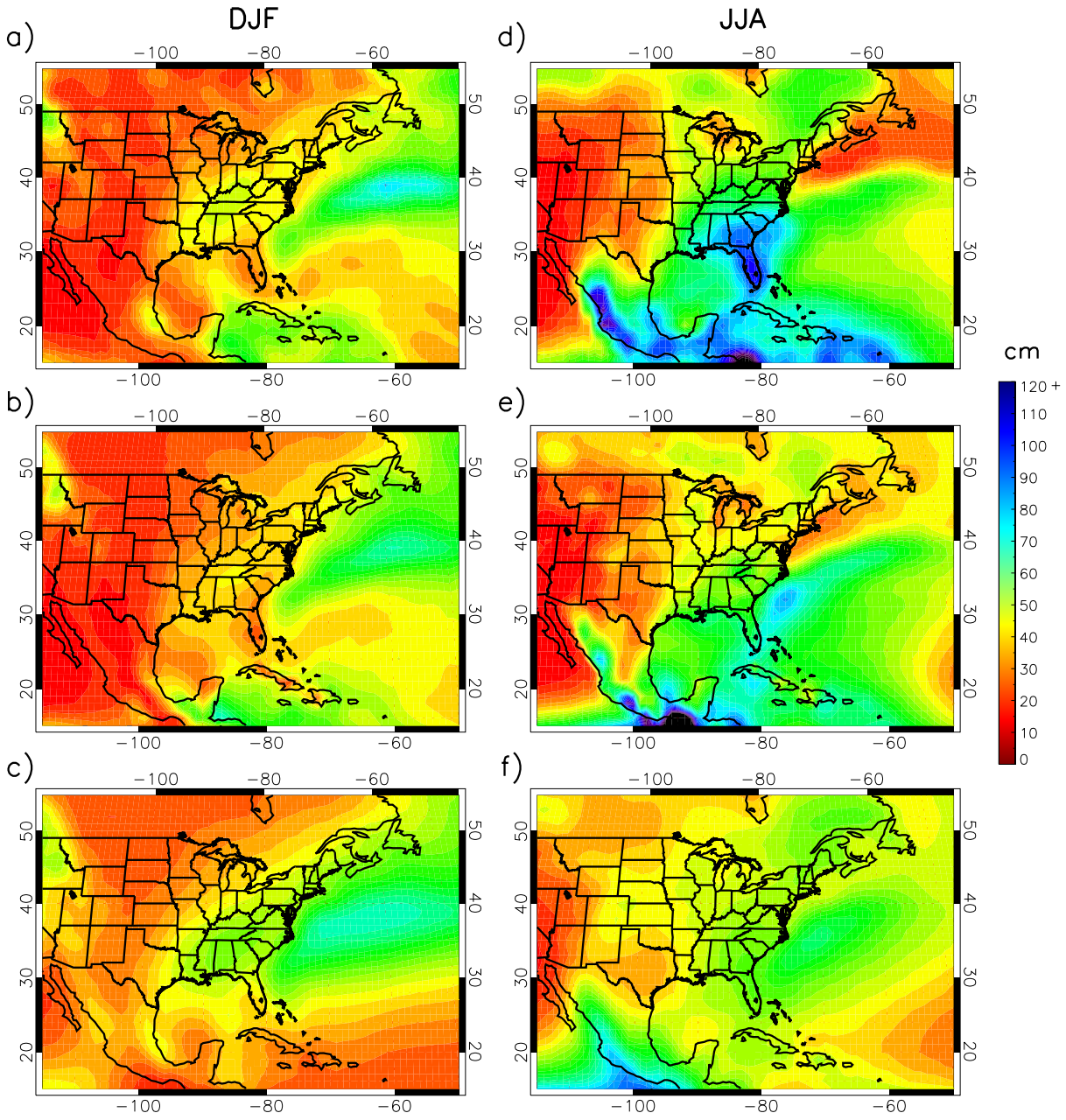


Figure A8. Mean winter (DJF) precipitation from NCEP (a), ERA40 (b) and the CMIP5 multi-model mean (c), and mean summer (JJA) precipitation from NCEP (d), ERA40 (e) and the CMIP5 multi-model mean (f).

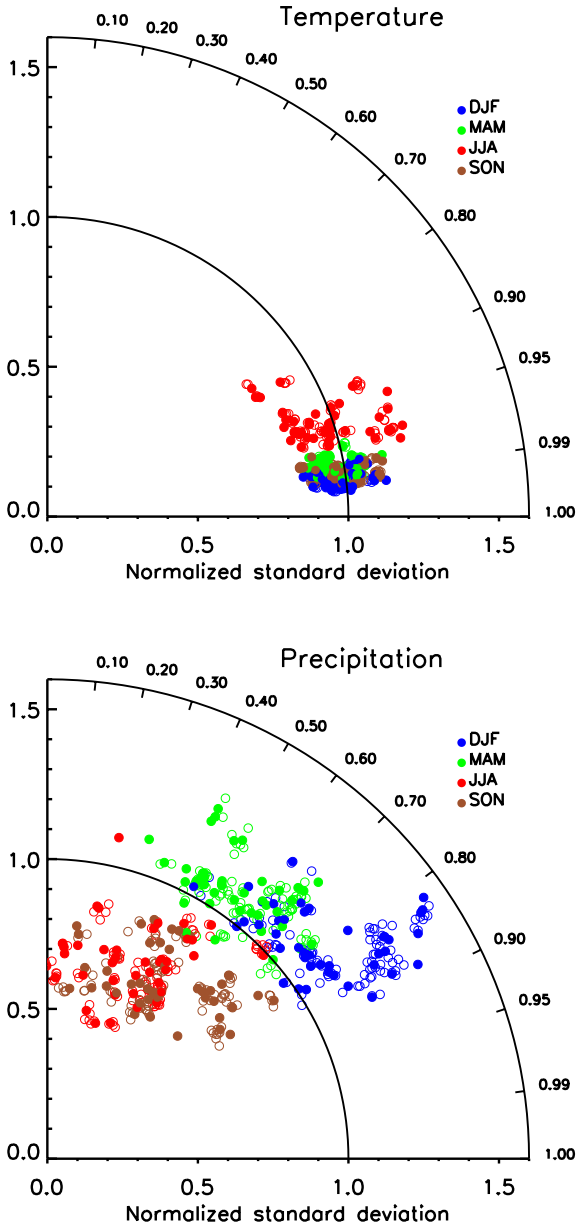


Figure A9. Normalized standard deviations (radius) and correlation coefficients (angle) by season for the climatological mean fields of temperature and precipitation from CMIP5. The spatial domain is approximately that shown in Figure A7. For each variable, the reference field for the normalization and the correlation is ERA40 reanalysis. Filled circles show the first ensemble members from each model and open circles show remaining ensemble members. Note that a perfect simulation would have both a normalized standard deviation and a correlation coefficient equal to unity.

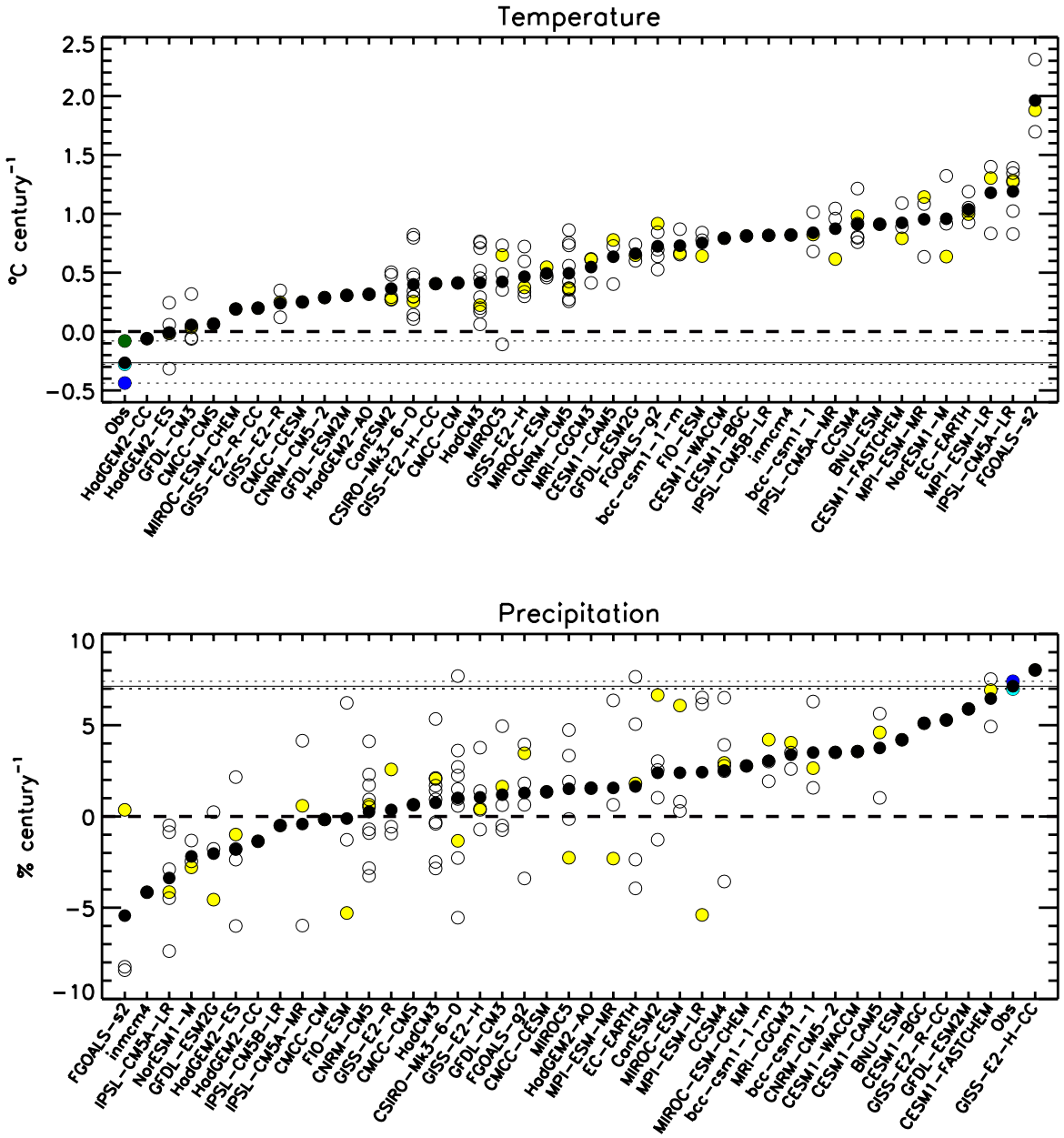


Figure A10. Southeast US-averaged trends in annual mean temperature and precipitation over the 20th century for all simulations and observations. For each of 41 CMIP5 GCMs, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

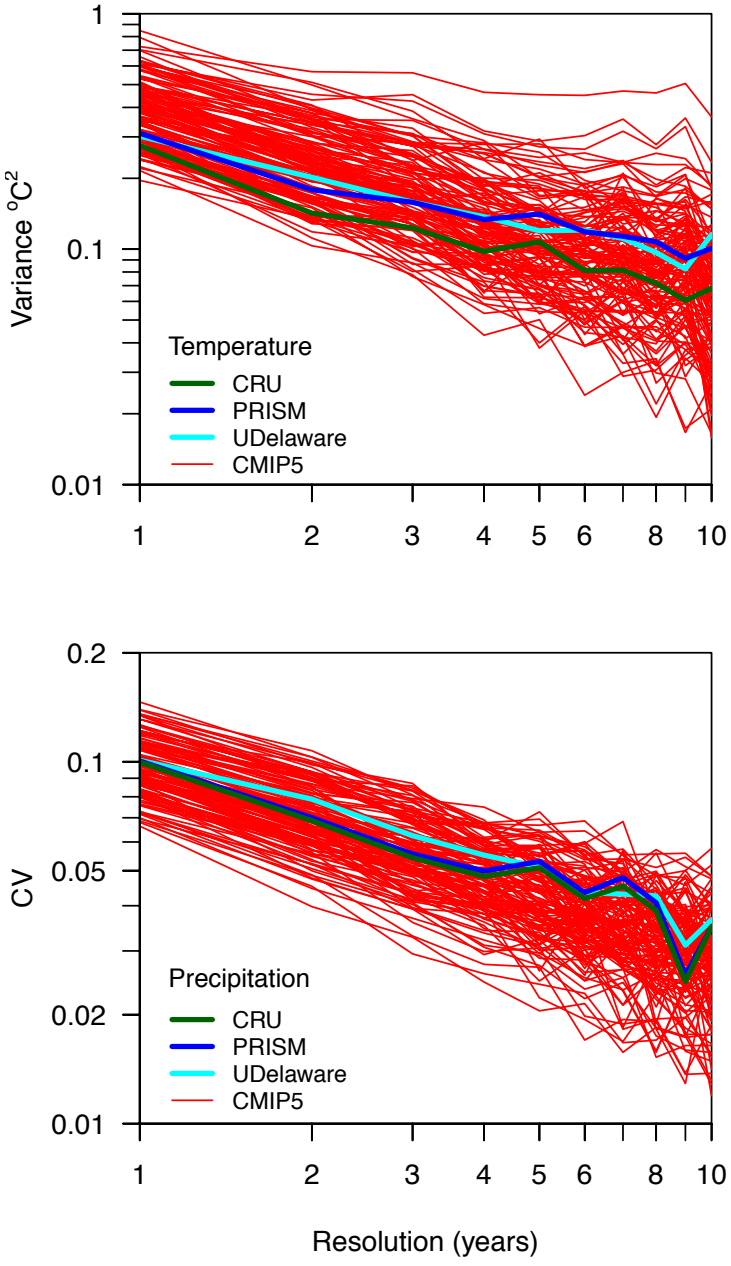


Figure A11. Variance of temperature anomalies (upper panel) and coefficient of variation of precipitation (lower panel) against temporal resolution for the Southeast US-averaged time series. Red lines show results from all ensemble members from 41 CMIP5 GCMs.

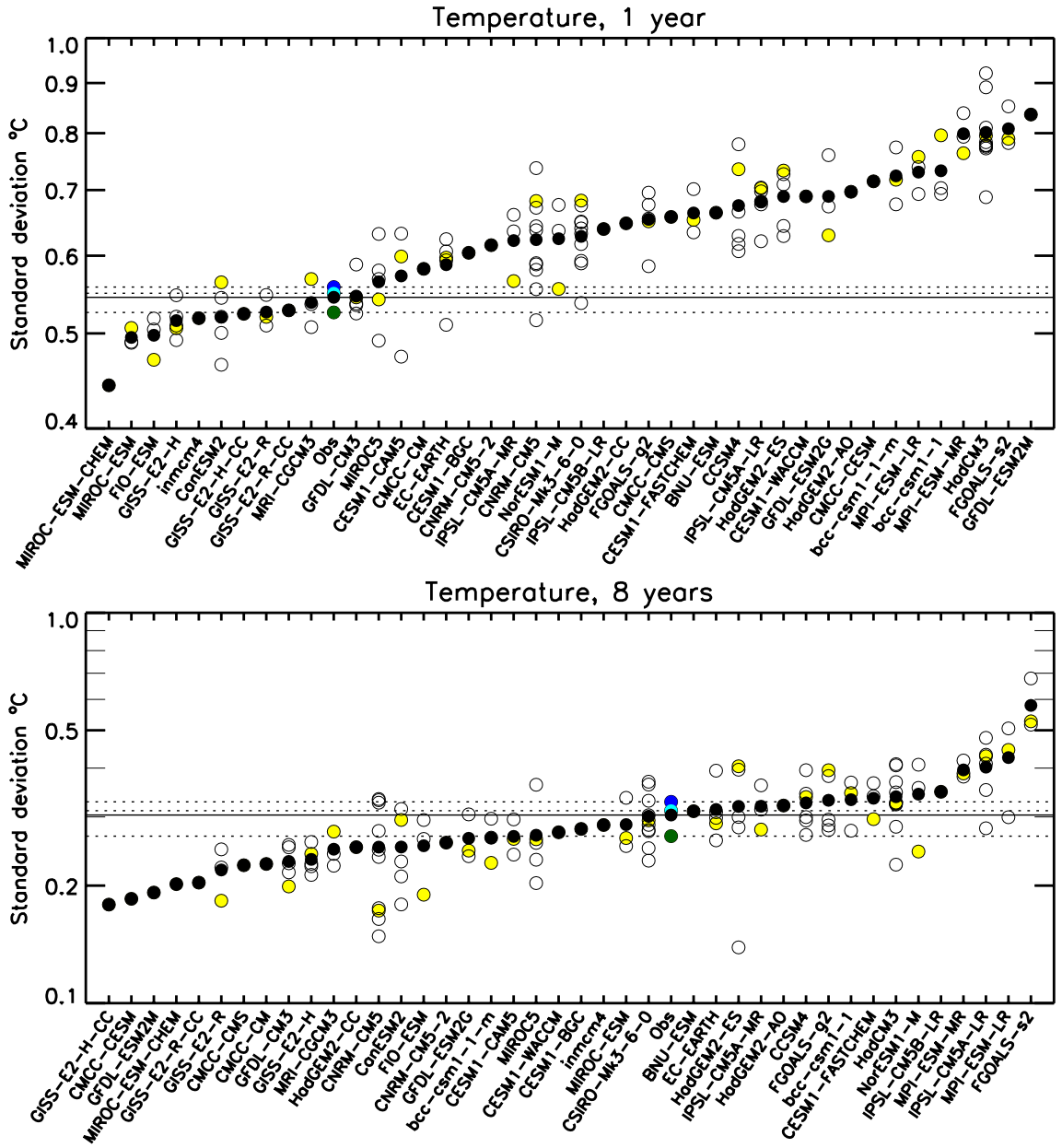


Figure A12. Standard deviation of temperature anomalies at resolutions of 1 year and 8 years. Values were averaged over the Southeast US domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

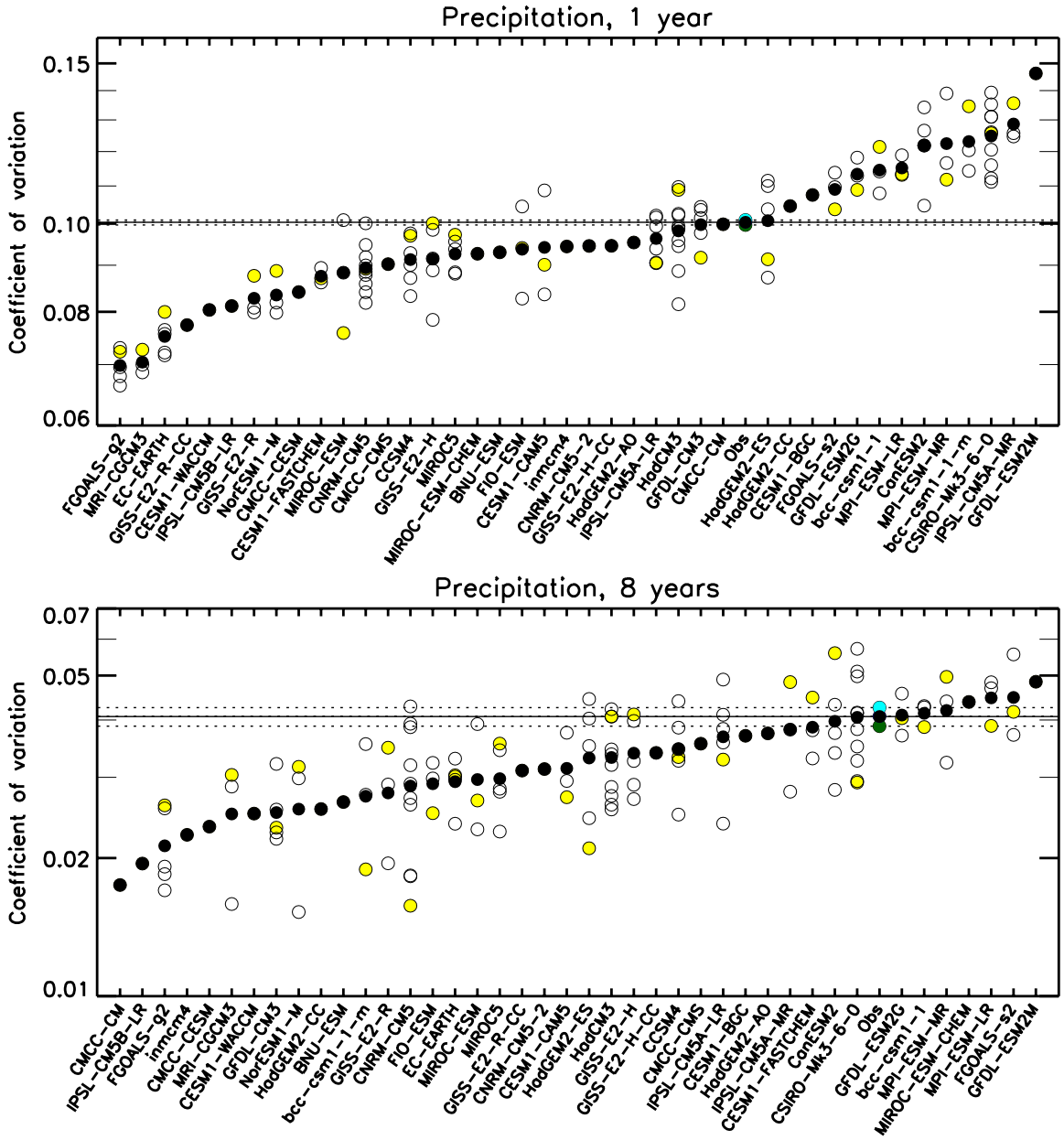


Figure A13. Coefficient of variation of precipitation at resolutions of 1 year and 8 years. Values were averaged over the Southeast US domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

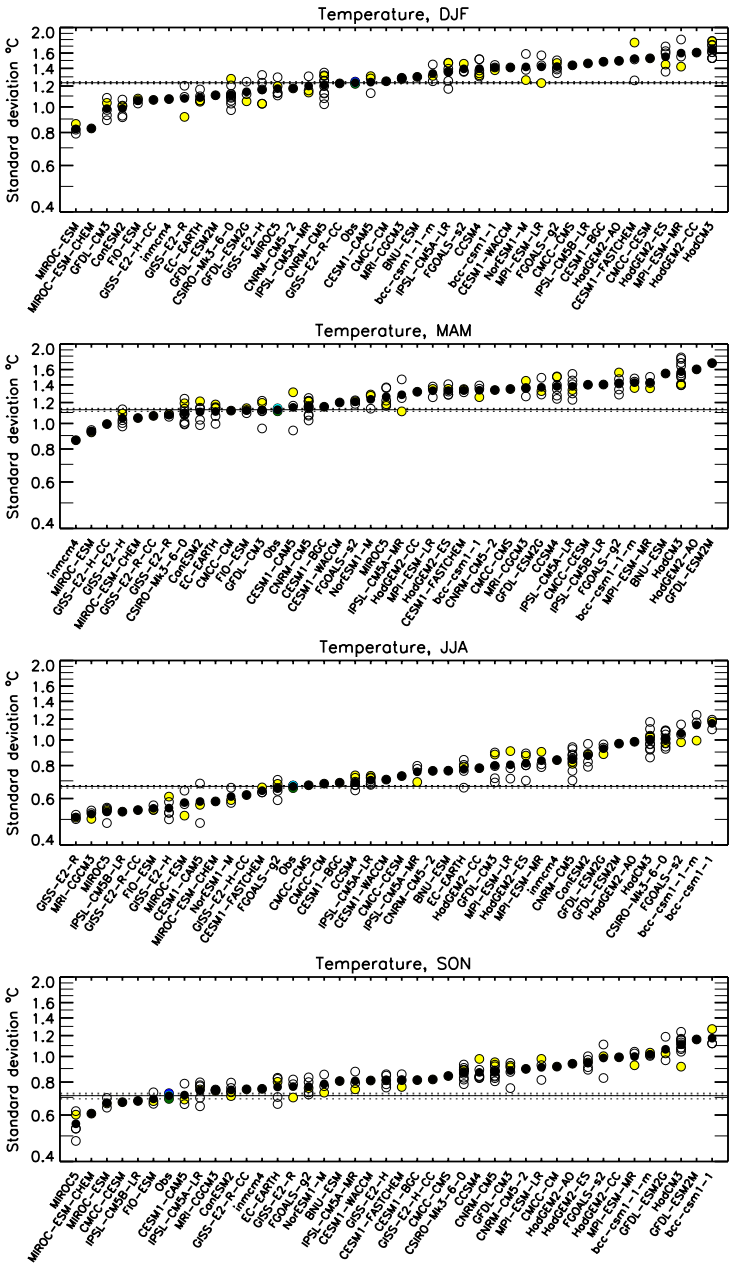


Figure A14. Standard deviation of seasonal mean temperature anomalies. Values were averaged over the Southeast US domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

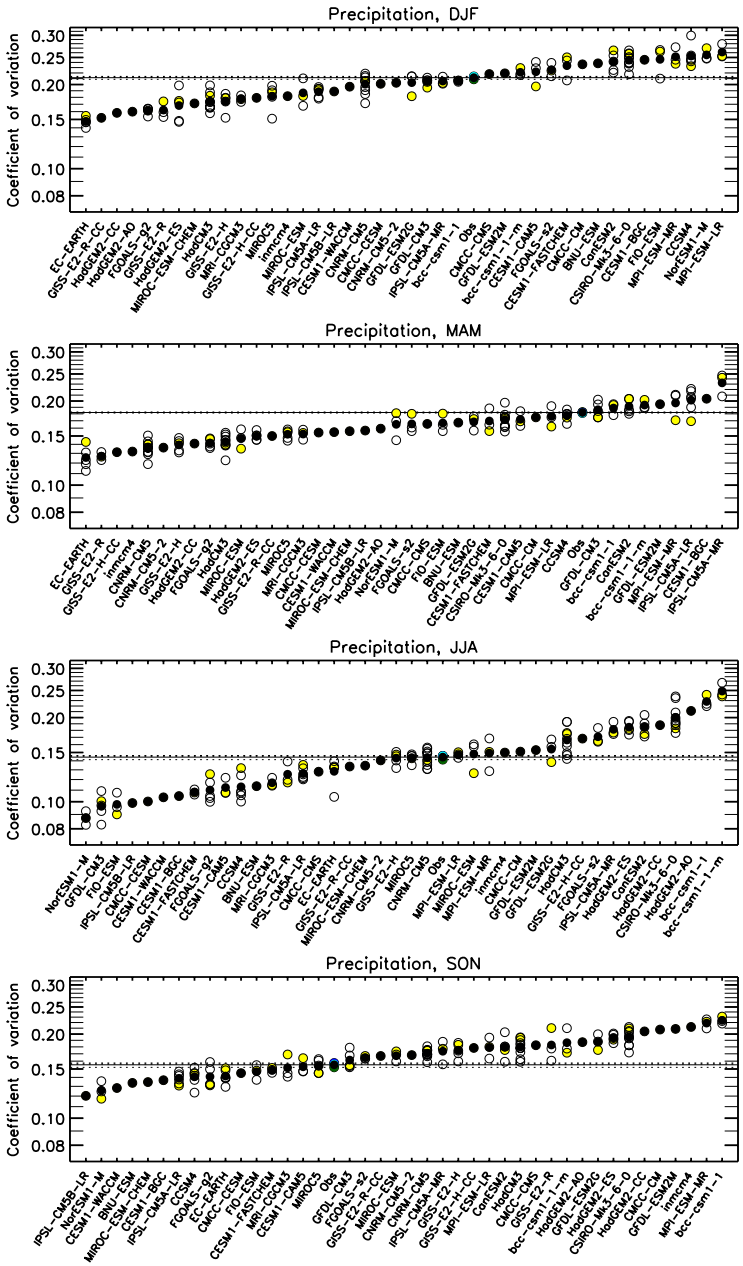


Figure A15. Coefficient of variation of seasonal mean precipitation. Values were averaged over the Southeast US domain. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

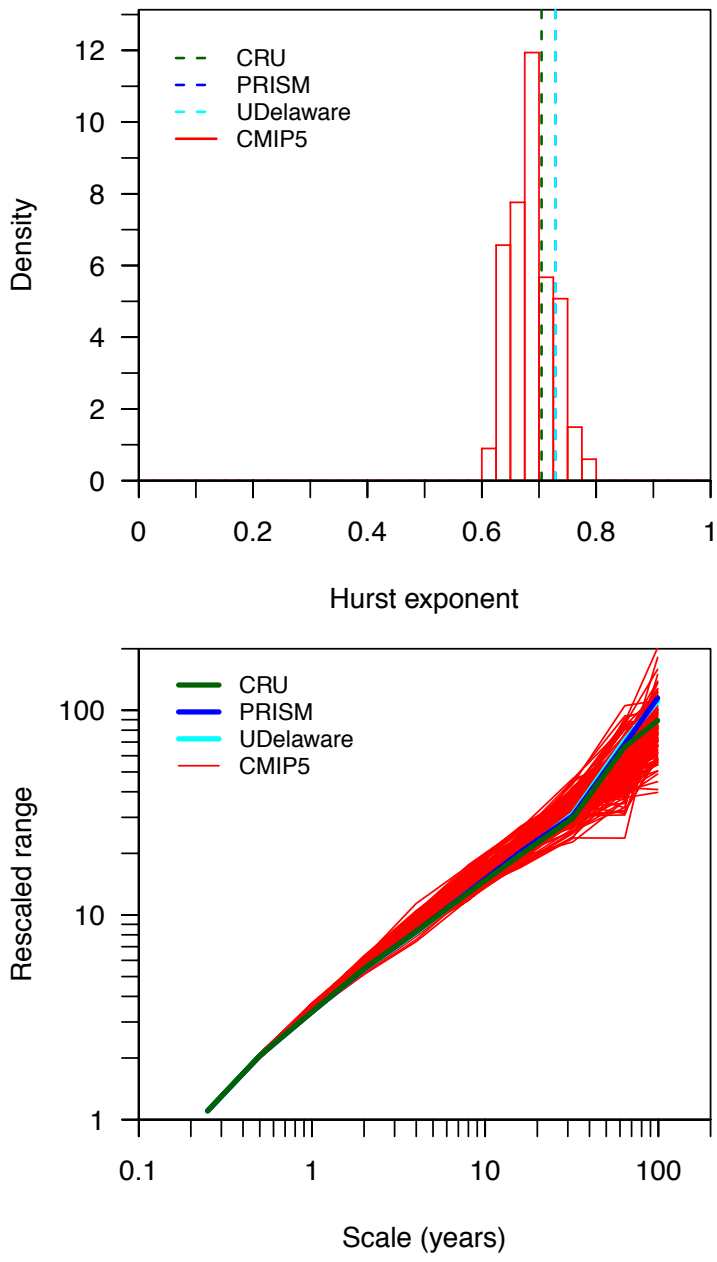


Figure A16. Upper panel: histogram of the Hurst exponent for Southeast US-averaged temperature in all CMIP5 simulations. The vertical dashed lines indicate the Hurst exponent estimated from observations. Lower panel: the rescaled range against the time scale calculated from the observations (heavy lines) and simulations (thin red lines).

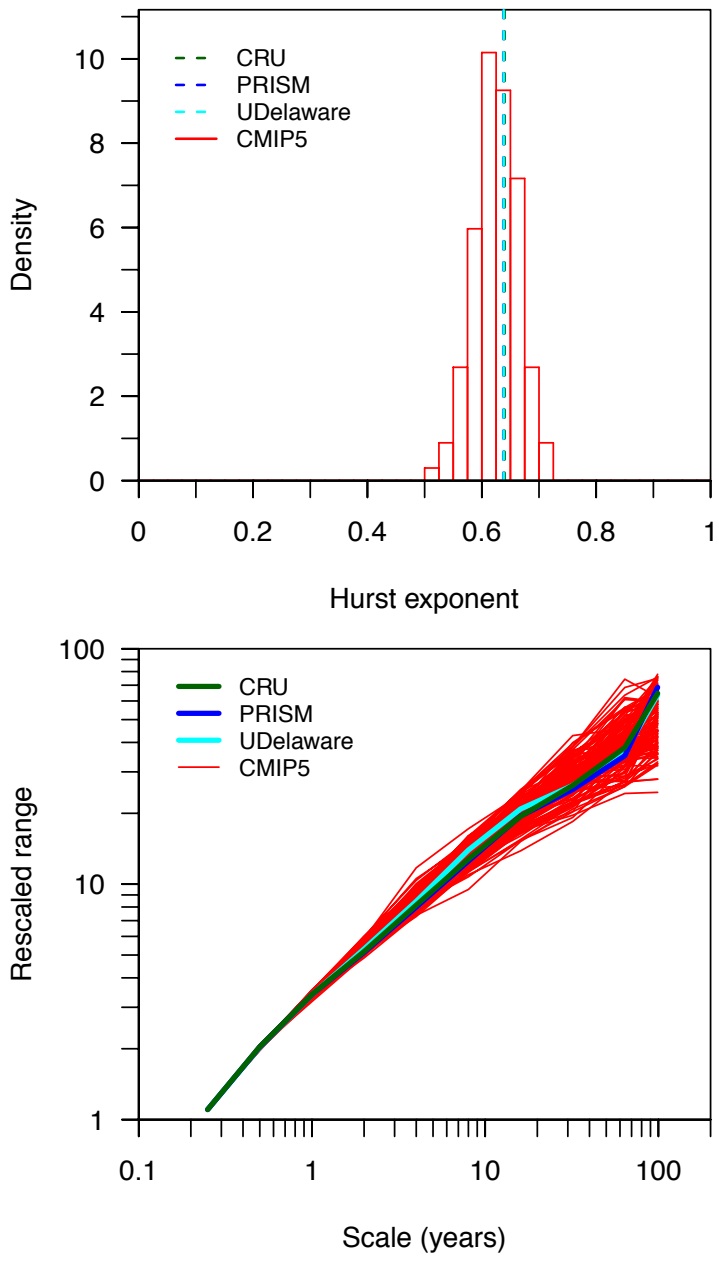


Figure A17. Upper panel: histogram of the Hurst exponent for Southeast US-averaged precipitation in all CMIP5 simulations. The vertical dashed lines indicate the Hurst exponent estimated from observations. Lower panel: the rescaled range against the time scale calculated from the observations (heavy lines) and simulations (thin red lines).

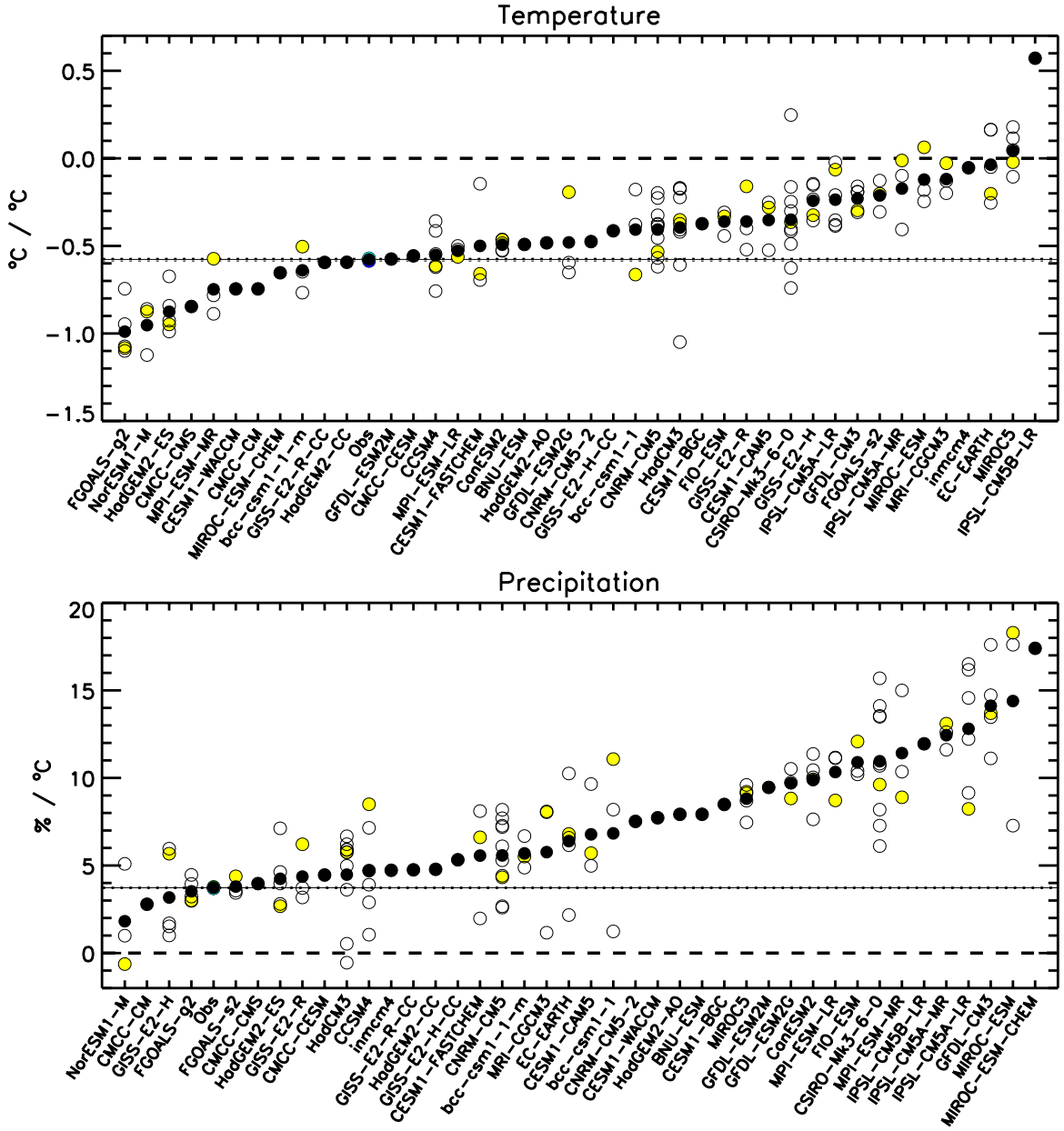


Figure A18. Response of Southeast US winter (JFM) temperature and precipitation to the Niño3.4 index averaged over NDJFM. For each of 41 CMIP5 models, black-filled circles show the ensemble average, yellow-filled circles show the first ensemble member, and the open circles show the remaining ensemble members. Observed (Obs) values are from CRU (dark green), PRISM (blue), UDelaware (cyan), and average of observations (black).

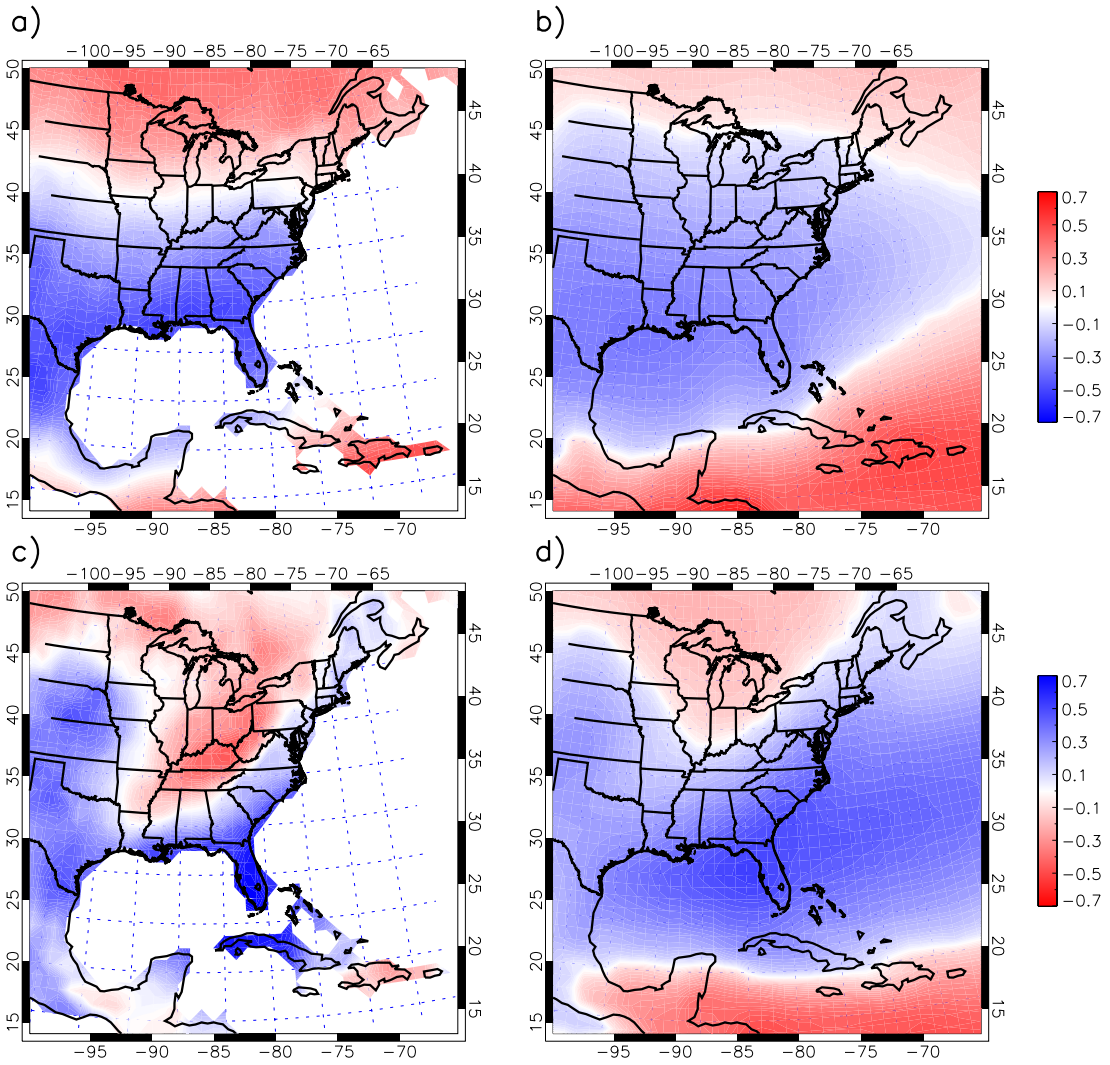


Figure A19. Correlation of CRU winter (JFM) (a) temperature and (c) precipitation with the Niño3.4 index averaged over NDJFM, and mean correlation of simulated winter (b) temperature and (d) precipitation to the same index. Simulations were from 41 CMIP5 models. Note that the legends have been reversed between the upper and lower plots so that blue implies both cooler and wetter conditions with a higher Niño3.4 index (i.e., El Niño conditions).

References

Daly, C., M. Halbleib, J. I. Smith, W. P. Gibson, M. K. Doggett, G. H. Taylor, J. Curtis, and P. A. Pasteris (2008), Physiographically-sensitive mapping of temperature and precipitation across the conterminous United States, *Int. J. Climatology*, *28*, 2031-2064, doi: 10.1002/joc.1688.

Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister (2013), Updated high-resolution grids of monthly climatic observations – the CRU TS3.10 Dataset. *Int. J. Climatol*., doi: 10.1002/joc.3711.

Hurst, H. E. (1951), Long-term storage capacity of reservoirs, *Trans. Am. Soc. Civ. Eng*., *116*, 770–799.

IPCC (2013), *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovern-mental Panel on Climate Change* [Stocker, T.F., D. Qin, G.-K. Plattner, M. Tignor, S.K. Allen, J. Boschung, A. Nauels, Y. Xia, V. Bex and P.M. Midgley (eds.)]. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 1535 pp.

Kalnay, E., et al. (1997), The NCEP/NCAR 40-year reanalysis project, *Bull. Amer. Meteor. Soc*., *77*, 437-470, 1996.

Klemes, V. (1974), The Hurst phenomenon: a puzzle? *Water Resour. Res*., *10*, 675-688.

Kumar, S., J. Kinter, P. A. Dirmeyer, Z. Pan, and J. Adams (2013) Multidecadal climate variability and the "Warming Hole" in North America: Results from CMIP5 Twentieth- and Twenty-First-Century climate simulations. *J. Climate*, 26, 3511–3527.

Rupp, D. E., J. T. Abatzoglou, K. C. Hegewisch, and P. W. Mote (2013), Evaluation of CMIP5 20th century climate simulations for the Pacific Northwest USA, *J. Geophys. Res. Atmos*., 118, 10,884–10,906, doi: 10.1002/jgrd.50843.

Taylor, K. E., R. J. Stouffer, and G. A. Meehl (2012), An overview of CMIP5 and the experiment design, *Bull. Am. Meteorol. Soc*., *93*, 485–498, doi: 10.1175/BAMS-D 11-973 00094.1.

Tessier, Y., S. Lovejoy, P. Hubert, D. Schertzer, and S. Pecknold (1996), Multifractal analysis and modeling of rainfall and river flows and scaling, causal transfer functions, *J. Geophys. Res.*, *101*(D21), 26,427–26,440, doi: 10.1029/96JD01799.

Uppala, S. M., and Coauthors (2005), The ERA-40 re-analysis, *Quart. J. R. Meteorol. Soc*., *131*, 2961-3012, doi: 10.1256/qj.04.176.

Willmott, C. J., and K. Matsuura (2012a), Terrestrial air temperature: 1900-2010 gridded monthly time series, version 3.01, http://climate.geog.udel.edu/~climate/html_pages/Global2011/README.GlobalTsT2011.html.

Willmott, C. J, and K. Matsuura (2012b), Terrestrial precipitation: 1900-2010 gridded monthly time series version, 3.02, http://climate.geog.udel.edu/~climate/html_pages/Global2011/Precip_revised_3.02/README.GlobalTsP2011.html.