https://doi.org/10.1007/s00382-016-3286-1

Evaluation of the skill of North-American Multi-Model Ensemble (NMME) Global Climate Models in predicting average and extreme precipitation and temperature over the continental USA

Louise J. Slater[1] · Gabriele Villarini[1] · Allen A. Bradley[1]

Received: 6 January 2016 / Accepted: 18 July 2016 / Published online: 4 August 2016

Abstract This paper examines the forecasting skill of eight Global Climate Models from the North-American Multi-Model Ensemble project (CCSM3, CCSM4, CanCM3, CanCM4, GFDL2.1, FLORb01, GEOS5, and CFSv2) over seven major regions of the continental United States. The skill of the monthly forecasts is quantified using the mean square error skill score. This score is decomposed to assess the accuracy of the forecast in the absence of biases (potential skill) and in the presence of conditional (slope reliability) and unconditional (standardized mean error) biases. We summarize the forecasting skill of each model according to the initialization month of the forecast and lead time, and test the models' ability to predict extended periods of extreme climate conducive to eight 'billion-dollar' historical flood and drought events. Results indicate that the most skillful predictions occur at the shortest lead times and decline rapidly thereafter. Spatially, potential skill varies little, while actual model skill scores exhibit strong spatial and seasonal patterns primarily due to the unconditional biases in the models. The conditional biases vary little by model, lead time, month, or region. Overall, we find that the skill of the ensemble mean is equal to or greater than that of any of the individual models. At the seasonal scale, the drought events are better forecast than the flood events, and are predicted equally well in terms of high temperature and low precipitation. Overall, our findings provide a systematic diagnosis of the strengths and weaknesses of the eight models over a wide range of temporal and spatial scales.

Keywords Seasonal forecasting · NMME · Flood · Drought · Multi-model ensemble · Model biases

1 Introduction

The North American Multimodel Ensemble (NMME) is an experimental project which was established in response to the U.S. National Academies' recommendation to support regional climate forecasting and decision-making over intraseasonal to interannual timescales (National Research Council 2010). Participating North-American agencies, which include the National Oceanic and Atmospheric Administration (NOAA)'s National Centers for Environmental Prediction (NCEP) and Geophysical Fluid Dynamics Laboratory (GFDL), the International Research Institute for Climate and Society (IRI), the National Center for Atmospheric Research (NCAR), the National Aeronautics and Space Administration (NASA)'s Global Modeling and Assimilation Office (GMAO), the Rosenstiel School of Marine & Atmospheric Science from the University of Miami (RSMAS), the Center for Ocean-Land–Atmosphere Studies (COLA), and Environment Canada's Meteorological Service of Canada—Canadian Meteorological Center (CMC), have been contributing model predictions from

This paper is a contribution to the special collection on the North American Multi-Model Ensemble (NMME) seasonal prediction experiment. The special collection focuses on documenting the use of the NMME system database for research ranging from predictability studies, to multi-model prediction evaluation and diagnostics, to emerging applications of climate predictability for subseasonal to seasonal predictions. This special issue is coordinated by Annarita Mariotti (NOAA), Heather Archambault (NOAA), Jin Huang (NOAA), Ben Kirtman (University of Miami) and Gabriele Villarini (University of Iowa).

✉ Louise J. Slater
louise-slater@uiowa.edu

[1] IIHR-Hydroscience and Engineering, The University of Iowa, Iowa City, IA, USA

Table 1 Summary of the characteristics of the eight NMME models

Model name	Modeling center	Available period	Ensemble size	Lead times (months)	Reference	Retrieved from	NMME Phase I	NMME Phase II
CCSM3 (community climate system model, version 3)	National Center for Atmospheric Research (NCAR); Center for Ocean–Land–Atmosphere studies (COLA); Rosenstiel School for Marine and Atmospheric Science, University of Miami (RSMAS)	1982-present	6	0.5–11.5	Kirtman and Min (2009)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.COLA-RSMAS-CCSM3/	✓	
CCSM4 (community climate system model, version 4-subset of CESM1)	NCAR/COLA/RSMAS (as above)	1982-present	10	0.5–11.5	Lawrence et al. (2012)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.COLA-RSMAS-CCSM4/		✓
CanCM3 (3rd generation canadian coupled global climate model)	Environment Canada's meteorological service of Canada–Canadian Meteorological Centre (CMC)	1981-present	10	0.5–11.5	Merryfield et al. (2013)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.CMC1-CanCM3/	✓	✓
CanCM4 (4th generation canadian coupled global climate model)	CMC (as above)	1981-present	10	0.5–11.5	Merryfield et al. (2013)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.CMC2-CanCM4/	✓	✓
CFSv2 (operational climate forecast system version 2)	NOAA's National Centers for Environmental Prediction (NCEP)	1982-present	28 (24 used/ 4 are incomplete)	0.5–9.5	Saha et al. (2014)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.NCEP-CFSv2/	✓	✓
GEOS5 (goddard earth observing system model, version 5)	National Aeronautics and Space Administration (NASA)'s Global Modeling and Assimilation Office (GMAO)	1981-present	12	0.5–8.5	Vernieres et al. (2012), Molod et al. (2012)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.NASA-GMAO-062012/	✓	✓
GFDL2.1 (climate model, version 2.1)	National Oceanic and Atmospheric Administration (NOAA)'s Geophysical Fluid Dynamics Laboratory (GFDL)	1982-present	10	0.5–11.5	Zhang et al. (2007); Delworth et al. 2006	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.GFDL-CM2p1-aer04/	✓	
FLORb01 (climate model version 2.5)	NOAA's GFDL (as above)	1982-present	12	0.5–11.5	Vecchi et al. (2014)	http://iridl.ldeo.columbia.edu/SOURCES/.Models/.NMME/.GFDL-CM2p5-FLOR-B01		✓

The available period does not reflect the presence of gaps in the forecasts. The number of ensemble members indicates the largest number of members per GCM and is not reflective of missing data for one or more members. The 0.5-lead time is the shortest available lead time and refers to the forecast for a month issued at the beginning of the month itself (e.g., the 0.5 lead time forecast for January 2000 is issued at the beginning of January 2000) their hindcasts (dating back to the early 1980s) and real-time forecasts since August 2011. Each model consists of between 6 and 28 "members," and the forecasts are provided at lead times that range between 0.5 and 11.5 months ahead of the forecast (Table 1). The two key advantages of the NMME, in comparison with other projects, are that the data are made freely available and that the focus is not just on retrospective forecasts, but also on real-time information.

A central component of the NMME project consists in quantifying model ensemble skill (Kirtman et al. 2014) to generate the most reliable climate forecasts. Model accuracy can be measured on several levels, by comparing each model's individual members, each model's ensemble mean (of model members), or the multi-model ensemble mean, against the observed climate data. Typically, multi-model means are found to have greater skill than single models (Hagedorn et al. 2005). Such averaging schemes are usually computed either by giving the same weight to each model's ensemble mean, or by giving equal weight to all members (thus assigning more weight to the models with more members) (e.g., Tian et al. 2014). The first assessments of NMME skill consistently suggest that the multi-model ensemble mean performs as well as, or better than, the best model (Becker et al. 2014; DelSole and Tippett 2014; Wood et al. 2015; Ma et al. 2015a; Thober et al. 2015). This increased skill of the NMME multi-model ensemble in contrast with the individual models appears to be related to the addition of new signals (from new models), rather than to the reduction of noise due to model averaging (DelSole et al. 2014).

However, because of the broad spatial and temporal scope of the NMME, most analyses of model skill are limited by necessity to specific lead times, regions, or seasons. Global, 1°-by-1° resolution studies tend to focus either on just one model, or on the shortest available lead time. For instance, Jia et al. (2015) characterize the skill of the high-resolution GFDL model FLOR, while Saha et al. (2014) investigate the skill of the NCEP Climate Forecast System (CFSv2) at the global scale. Conversely, Becker et al. (2014) provide a comprehensive analysis of temperature, precipitation, and sea surface temperature forecasts for multiple models at the global scale, but focus mainly on the shortest available lead time. Wang (2014) examines the global skill of NMME precipitation forecasts for the summer months and only at the shortest lead time. Mo and Lettenmaier (2014) interpolate the NMME forecasts bilinearly to a 0.5° grid over the continental United States to evaluate runoff and soil moisture forecasts, but only up to the 3-month lead time.

In contrast, analyses of the NMME conducted at the sub-continental scale often allow for a more comprehensive examination of model skill and of the relationship between ensemble forecasts and climate oscillations, and reveal regional agreement between models (Infanti and Kirtman 2016). In the southeastern United States, for example, it is shown that temperature and precipitation forecasts become increasingly skillful in the winter months at short lead times (Infanti and Kirtman 2014). Studies found that the predictability of precipitation (Mo and Lyon 2015), and/or temperature (Roundy et al. 2015) and drought (Ma et al. 2015b) generally improves in regions that are significantly affected by the El Niño-Southern Oscillation (ENSO). In North America, the majority of high correlations between temperature/precipitation forecasts and observations are found in the south-east (SE), south-west (SW), and north-west (NW) during Eastern Pacific El Niño events (Infanti and Kirtman 2016). Such analyses also help determine which models are the most useful at the regional/seasonal scale; for instance, over continental China, the CFS models performed the best, followed by GFDL, NASA, the Canadian models, IRI and CCSM3 (Ma et al. 2015b) (see Table 1 for an overview of models and acronyms—note that we did not include IRI's fourth-generation atmospheric GCM (ECHAM4p5) in our model selection because it no longer issues real-time forecasts). In an analysis of four NMME models over the continental United States and the Atlantic Warm Pool (AWP), the CFSv2 and GFDL models showed the most skill for predicting seasonal rainfall anomalies in the July–October season (Misra and Li 2014).

Thus, despite an increasing number of analyses focused on the quantification of NMME skill, a systematic investigation across different models, regions, seasons, and lead times is still lacking. Additionally, very little is known regarding the skill of these models for forecasting extended periods of high temperature and/or low precipitation leading to drought conditions, as well as extreme precipitation leading to flooding. For instance, we know that most NMME models were unable to forecast the 2012 North American drought correctly, while those that correctly predicted its occurrence did so fortuitously, and "for the wrong reason" (Kam et al. 2014). Therefore, a thorough evaluation of the NMME models' ability to forecast the occurrence of different extremes over extended periods of time is also missing.

To fill these gaps, the research questions that we address in this study are the following:

- At the intraseasonal scale, what is the skill of the eight individual NMME model ensembles in predicting precipitation and temperature patterns, for every available lead time, every month of the year, and for every subregion of the continental United States? How do their biases compare? Do certain models perform better than others for certain regions, lead times, and months, and does the eight-model ensemble mean outperform the individual models?

- At the seasonal scale, what is the ability of these eight models to forecast extended periods of high temperature and low precipitation leading to drought conditions, as well as prolonged periods of extreme precipitation leading to flooding?

To answer these questions, we conduct a systematic decomposition of the forecasting skill of the eight individual model ensembles (computed as the mean of all members in each model) as well as of the eight-model ensemble mean (computed by assigning the same weight to each model's mean), using the NMME forecast data and observed monthly data for verification. Section 2 presents the forecast and observed data, and Sect. 3 provides an overview of the statistical methods used to perform forecast verification and the diagnosis of each model's ability to predict seasonal extremes. The results are presented in Sect. 4, while Sect. 5 summarizes the main findings and conclusions of the study.

2 Data

2.1 NMME temperature and precipitation data

Here we focus on eight GCMs from the NMME project, for which temperature and precipitation forecasts are available from the early 1980s to the present. The GCMs we consider are: CCSM3 and CCSM4 from NCAR, COLA and RSMAS; CanCM3 and CanCM4 from Environment Canada's CMC; CM2.1 and FLORb01 from NOAA's GFDL; GEOS5 from NASA's GMAO; CFSv2 from NOAA's NCEP. The characteristics of the different models are summarized in Table 1.

The data were downloaded from the IRI/Lamont Doherty Earth Observatory (LDEO) Climate Data Library (http://iridl.ldeo.columbia.edu/) in netCDF format, on a 1.0° latitude by 1.0° longitude grid. Monthly total precipitation (variable name "prec", in mm/day) and monthly reference mean temperature at 2 meters (variable name "tref", in Kelvin units) were obtained for all available lead times and ensemble members over the continental United States. Temperature data were converted from Kelvin units to degrees Celsius. For CanCM3, CanCM4, and CFSv2, the hindcast and forecast data were downloaded separately and combined for the analysis. In the case of CFSv2 we used the pentad realtime forecasts which match the pattern of the CFSv2 hindcasts.

Data were extracted for each model from netCDF files in R using the ncdf4 package (Pierce 2014). The files typically contain five dimensions, which are the longitude, latitude, member, lead, and forecast reference time. The number of ensemble members ranges from 6 for COLA to 12 for GEOS5 and FLORb01, and 28 for CFSv2 (Table 1). To limit the scope of the analysis, we consider the mean of each model's ensemble members, rather than analyze each model member individually. The focus of our analysis is monthly to seasonal predictions, ranging from 0.5 to 11.5 month leads. The term "lead" indicates the period between the forecast initialization time and the month that is predicted (so a "0.5-month lead forecast" refers to a monthly forecast that was made about 15 days ahead of the forecast period). Model forecast lead times vary from 0.5 to 8.5 months for GEOS5, up to 9.5 months for CFSv2, and up to 11.5 months for all of the other models (Table 1). Here, the expression "forecast reference time" refers to the date when the forecasts were issued (e.g., July 2015).

To analyze forecast skill at the regional scale, we define seven major regions of the United States based on the boundaries described in Kunkel et al. (2013), which are a modification of the regions that were originally used in the 2009 National Climate Assessment Report (Karl et al. 2009) by dividing the Great Plains Region into North and South (Fig. 1). The NMME data are projected as stacked rasters and cropped to the dimensions of these seven regions using the 'raster' package in R (Hijmans 2015), to extract the mean weighted forecast value of all of the grid cells falling within each region (as defined by the polygons) for every month and lead time.

2.2 Reference temperature and precipitation data

To verify model skill, we use temperature and precipitation data from the Parameter-elevation Regression on Independent Slopes Model (PRISM) climate mapping system (Daly et al. 2002), which represents the reference dataset for the continental United States. PRISM's temporal and spatial resolutions are monthly and approximately 4 km. The data are freely available from the web (http://www.prism.oregonstate.edu/index.phtml) and cover the period from 1890 to the present. We divide precipitation monthly totals by the number of days in each historical month to obtain daily values, and to match the units of the NMME models. Extracted precipitation and temperature data time series are plotted against reference PRISM data for every model, region, month, and lead time for verification purposes (see Supplementary materials, pp 2–25).

Other studies (e.g., Becker et al. 2014; Infanti and Kirtman 2014) have used as verification field the station observation–based Global Historical Climatology Network and Climate Anomaly Monitoring System (GHCN + CAMS) for temperature, and the Climate Prediction Center (CPC) global daily Unified Raingauge Database (URD) gauge analysis for precipitation rate. Here we chose to use PRISM data instead because they account for elevation in the interpolation scheme, have a fine spatial resolution, and are the official product for the U.S. Department of Agriculture.

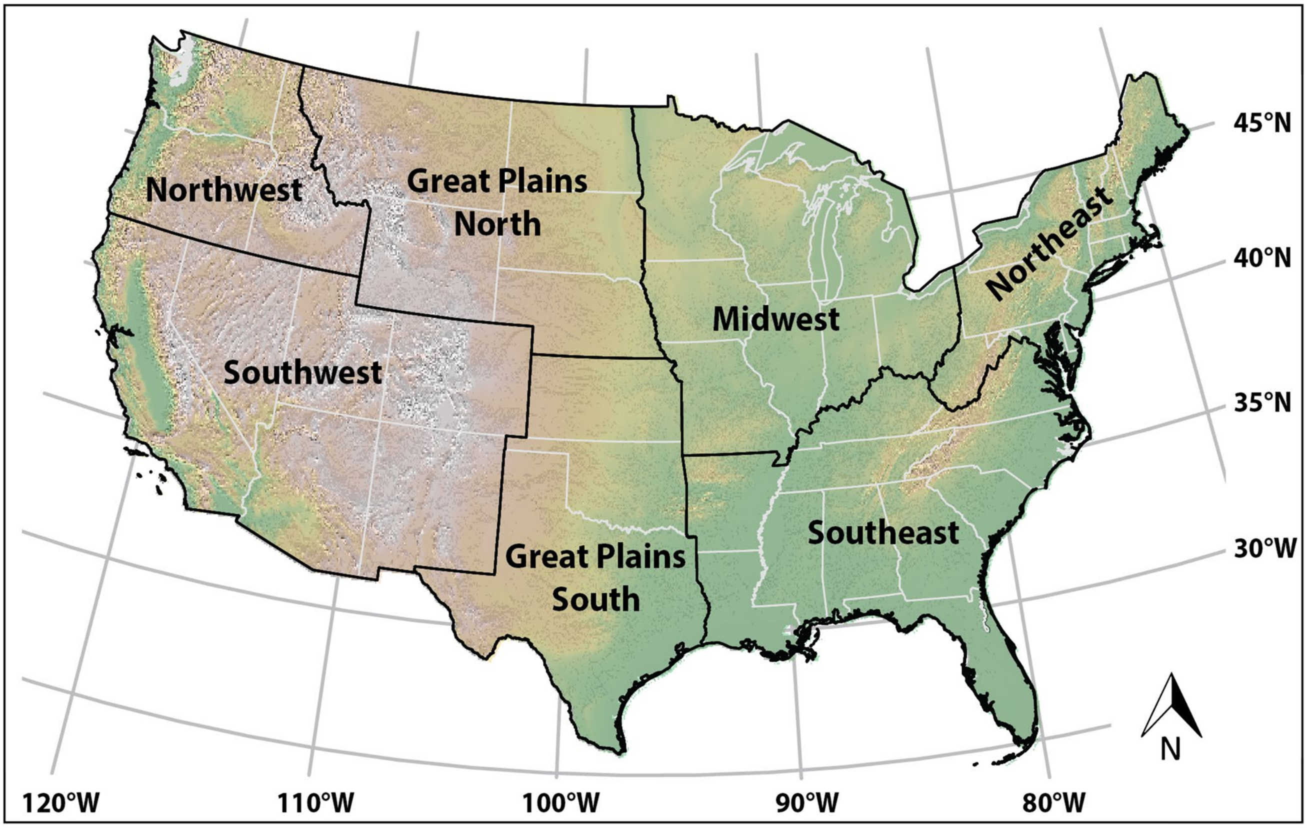

Fig. 1 Location of the seven regions across the continental United States. *Black* outline indicates the extent of the regions. *Pale gray* outline indicates the states within each region. *Colored* topographic *shaded* relief is shown in the background

3 Methodology

3.1 Forecast verification

Different approaches and methods have been developed to quantify the skill of a forecast system. Here we quantify the accuracy of the forecast relative to the climatology (used as reference) using the mean square error (MSE) skill score SS_{MSE} (e.g., Hashino et al. 2007):

$$SS_{MSE} = 1 - \frac{MSE}{\sigma_x^2} \tag{1}$$

where σ_x represents the standard deviation of the observations. A perfect forecast receives a skill score of 1. As the value tends to zero, the forecast skill decreases. A value of 0 indicates that the forecast accuracy is the same as would be achieved using climatology as the forecast. Negative values indicate that the accuracy is worse than the climatology forecast. The value of SS_{MSE} can be decomposed into three components (Murphy and Winkler 1992):

$$SS_{MSE} = \rho_{fx}^2 - \left[\rho_{fx} - \frac{\sigma_f}{\sigma_x}\right]^2 - \left[\frac{\mu_f - \mu_x}{\sigma_x}\right]^2 \tag{2}$$

where ρ_{fx} is the correlation coefficient between observations and forecasts and quantifies the degree of linear dependence between the two; μ_f and μ_x are the forecast and observation means, respectively; σ_f represents the standard deviation of the forecasts. Based on this decomposition, the value of the correlation coefficient (or its squared counterpart, the coefficient of determination) reflects the forecast accuracy in the absence of biases. For this reason, it represents the potential skill (PS), which is the skill that could be achieved without the quantification of the biases. Thus, it is commonly assumed (e.g., Boer et al. 2013; Younas and Tang 2013) that the difference between the potential and actual skill represents "room for model improvement"; however, as explained by Kumar et al. (2014), there is not necessarily a relationship between the potential and the actual skill of climate models, and assuming that there should be one amounts to expecting that the real-world data should behave identically to the model predictions.

The second term in the right hand side of Eq. (2) quantifies the conditional biases and is referred to as the slope reliability (SREL). The last term quantifies the unconditional biases and it is referred to as the standardized mean error (SME).

Forecast verification using the skill score and its decompositions in Eq. (2) is a diagnostic tool that produces a more realistic quantification of the forecast skill compared to taking the correlation coefficient at face value. Moreover, the decomposition of the skill in different bias sources can provide model developers with feedback about strengths and weaknesses of their models. In general, unconditional biases (large SME) can easily be removed with bias-correction methods (Hashino et al. 2007), while conditional biases (large SREL) may require more sophisticated calibration. Any forecasts with low potential skill (PS) will have limited predictability, even if biases are eliminated.

To perform the skill verification of the NMME, we tailor the PRISM and NMME data to cover the same months between January 1982 and December 2014. The verification is carried out for each model ensemble mean, region, and lead time following the above procedure, as also described in Bradley and Schwartz (2011). A separate skill verification is conducted on the eight-model ensemble mean, which is the mean forecast of all models (where one model already represents the arithmetic mean of its own ensemble members), for each region and lead time.

3.2 Extreme event diagnosis

The second part of the diagnosis is the assessment of each model's ability to predict extreme floods and droughts at the seasonal scale. To do this, we investigate the models' capacity to capture prolonged periods of extreme precipitation and temperature lasting several months. Eight extreme flood and drought events affecting different parts of the continental United States were selected based on their severity and duration. The event had to last at least one full month, and less than a year, so that we might evaluate its predictability for multiple lead times. The severity of the events was evaluated using the NOAA's Billion Dollar Weather and Climate Disasters Table of Events (https://www.ncdc.noaa.gov/billions/events). The chosen events include four floods (July–August 1993, January–March 1995, June–August 2008, and March 2010) and four droughts (June–August 1988, March–November 2002, March–August 2011, and May–August 2012). For the flood events, we focus on positive precipitation anomalies (high rainfall), and for the droughts, we observe positive temperature anomalies and negative precipitation anomalies (high temperature and lack of rainfall).

We first define the extent of each event based on the description given in the Billion Dollar Weather Table. The PRISM data are aggregated over the entire continental United States at the 1° × 1° resolution to match the spatial resolution of the NMME data. At each 1° pixel and for the period of interest for a given event, we compute the standardized anomalies with respect to the mean and standard deviation computed over the 1983–2014 period (the years 1982 and 2015 are excluded systematically because not all models have a complete forecast for 1982, and 2015 forecast data were not yet available for all events at the time of the analysis). We then extract all the cells with standardized anomalies larger than 1 and smaller than -1 (depending on whether we are considering excess temperature/precipitation or lack of rainfall). The resulting raster contains only the grid cells for that event which were "anomalously" high or low with respect to the 1983–2014 climatology. The boundaries of the event are tailored to the locations indicated in the Billion Dollar Weather Table (Fig. 2). We then average all of the pixels within the region for the months characterizing each event (e.g., total rainfall for June–August 2008, for each year between 1983 and 2014) and compute the "domain averaged" standardized anomalies. Confidence intervals are computed around the anomaly using the approach described in Stedinger et al. (1993, Sect. 18.4.2).

Last, we use a similar procedure to calculate the corresponding NMME anomalies within the defined region. One mean (spatially-averaged) model forecast is extracted for the entire region for the selected months between 1983 and 2014, for each lead time. To obtain the seasonal forecast we compute the sum of forecasts initialized ahead of the entire season. Thus, for an event such as the June–August 2008 flood, the seasonal forecast initialized in June 2008 (just before the event) is calculated as the sum of the 0.5-, the 1.5-, and the 2.5-month lead forecasts initialized in June. If we initialize the forecast one month earlier, in May, the forecast can be calculated as the sum of the 1.5-, the 2.5- and the 3.5-month lead forecasts initialized that month. The forecast is calculated for increasingly long initialization times by going back in monthly time steps, as far the available lead times will allow. The resulting seasonal forecasts are then computed as anomalies, to allow a direct comparison with the average PRISM climatological anomaly for the event.

4 Results

4.1 Regional temperature and precipitation forecast skill

4.1.1 Temperature

The potential skill of the eight-model ensemble mean, as measured by the squared correlation coefficient between model forecasts and PRISM observations, ranges between 0 and 0.6 (Fig. 3a). We find that the highest skill is displayed at the shortest lead time (0.5-month lead) and

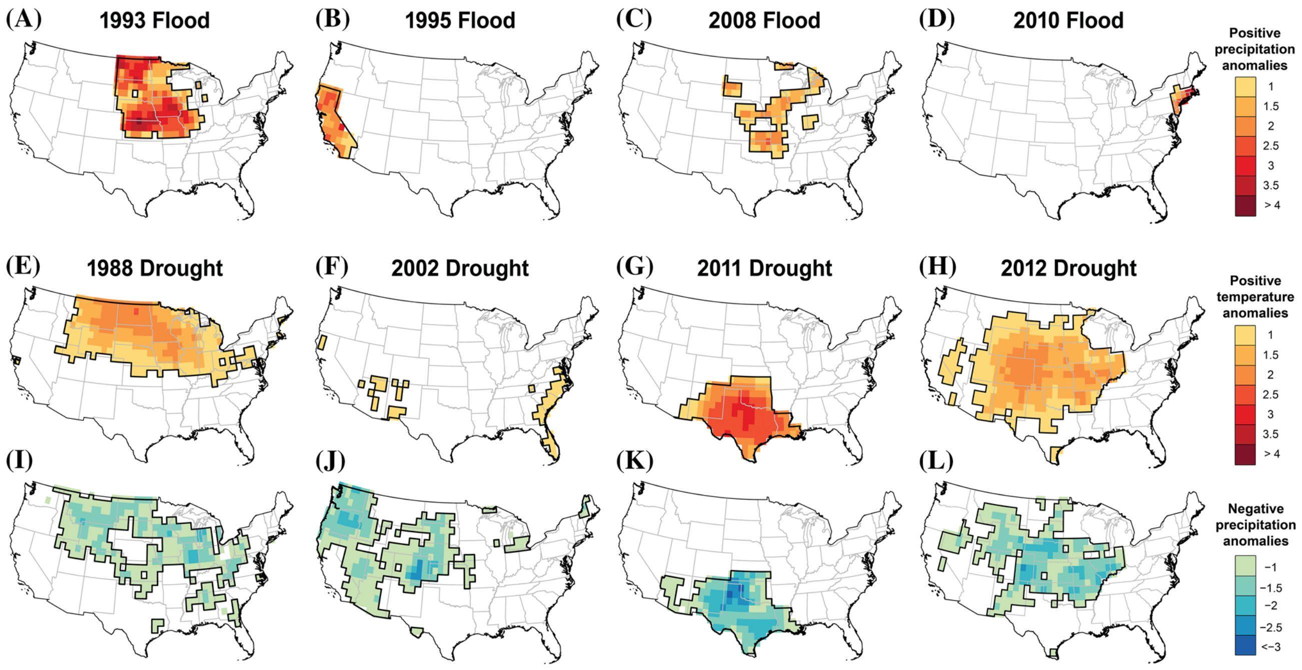

Fig. 2 Location of the studied flood and drought events across the continental United States. Computed climatological anomalies are indicated as *red shades* for temperature, and as *blue shades* for precipitation. *Thick black* outline indicates the spatial extent of the event. *Color intensity* indicates the anomaly of the observed climatology for the given season (>1 or <−1), as calculated on a pixel-by-pixel level across the entire United States. **a** 1993 July–August flood, precipitation anomalies. **b** 1995 January–March flood, precipitation anomalies. **c** 2008 June–August flood, precipitation anomalies. **d** 2010 March flood, precipitation anomalies. **e** 1988 June–August drought, temperature anomalies. **f** 2002 March–November drought, temperature anomalies. **g** 2011 March–August drought, temperature anomalies. **h** 2012 May–August drought, temperature anomalies. **i** 1988 June–August drought, precipitation anomalies. **j** 2002 March–November drought, precipitation anomalies. **k** 2011 March–August drought, precipitation anomalies. **l** 2012 May–August drought, precipitation anomalies

declines rapidly thereafter, so most regions and months display a skill of less than 0.1 by the 1.5-month lead time (Fig. 3a). The Northwest and Southwest tend to show better skill than the other regions at longer lead times, e.g., over the January–March and June–July periods respectively, possibly because of the good predictability of temperature anomalies arising from ENSO conditions during the same months (see e.g., Wolter and Timlin 2011, and mapping of the likelihood of seasonal extremes by the NOAA/ESRL Physical Science Division at http://www.esrl.noaa.gov/psd/enso/climaterisks/). Other regions such as the Midwest show almost no skill beyond the shortest lead time, possibly because of the weaker relationship with ENSO states.

Overall, the ensemble mean displays better ability than any of the individual models, with potential skill maxima that exceed that of any single model (see for example April temperatures in the Midwest at the 0.5-lead time, Figs. 3, 4), in agreement with other assessments of NMME model skill (Infanti and Kirtman 2014; Kirtman et al. 2014). There is not one model that clearly outperforms any of the others, although CCSM4, CanCM3, CanCM4, GEOS5 and CFSv2 do display better skill than CCSM3, GFDL2.1, and FLORb01 (Fig. 4). The same seasonal and regional patterns can be seen for the individual models as for the ensemble mean, with a clear peak in potential skill in the Southwest in the summer months (CCSM4, CanCM4).

The actual skill score is relatively low for all models and is mainly driven by the large unconditional biases (SME) in the models. The influence of the unconditional biases on the skill score is clearly detectable in the mirror-image pattern between the two (Figs. 3, 4). Dark blue colors indicating low skill scores are reflected by the dark red colors indicating a high unconditional bias. Overall, the skill score tends to be highest at the shortest lead times. The skill score of the ensemble mean can be quite high in specific regions such as the Midwest at the 0.5-month lead time during the cold season. Individual models, however, exhibit low skill scores over most regions and months, with values reaching below −10 most of the time (see Supplementary Materials pp. 26–29 for additional graphs indicating skill decomposition for the eight-model ensemble mean and for each individual model).

The unconditional biases display strong seasonal variability: they tend to be the lowest (white) in most regions in the winter/spring months, and tend to increase dramatically (red) in the summer. By contrast, the Northwest and

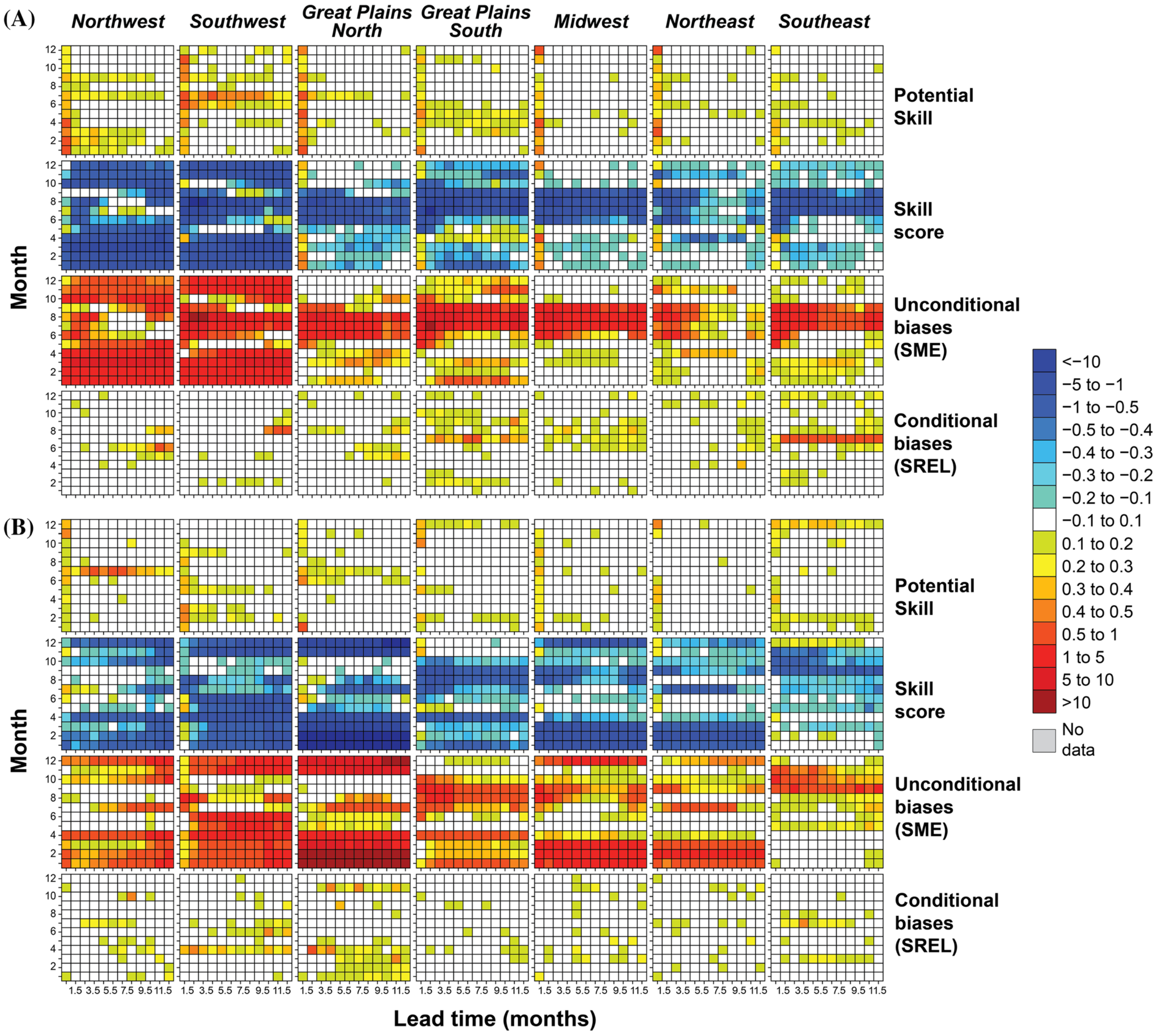

Fig. 3 *Color* maps indicating average skill of the eight-model ensemble mean for **a** Temperature and **b** Precipitation. For each individual *color* map (1 *box*), x-axis indicates the lead time of the climate forecast, ranging from 0.5 to 11.5 months; y-axis indicates the month that is forecast, ranging from 1 (January) to 12 (December). Labels at the *top* of the figure indicate each of the 7 regions shown in Fig. 1 (Northwest, Southwest, Great Plains North, Great Plains South, Midwest, Northeast, and Southeast). *Right side* of the figure indicates the computed components of the ensemble skill: potential skill, skill score, unconditional biases (SME), and conditional biases (SREL). The *color scale* on the *right side* of the figure is used for all components of the skill score, and ranges from less than −10 (*blue shades*) to more than 10 (*red shades*)

Southwest exhibit systematically higher biases in the winter and spring (particularly in the model ensemble). Therefore, as a result of this seasonality (e.g., better characterization of initial land surface conditions in the cold seasons), the unconditional biases also show some lead-dependence: during the summer months, they are the highest at the shortest leads (dark red), and decrease progressively with lead time (as is visible in the case of CanCM4/CanCM3, and to a lesser extent CFSv2). These seasonal fluctuations have a notable influence on the overall skill score, and suggest that forecasts made in the summer months could generally be improved by eliminating the unconditional biases.

The conditional biases (SREL) tend to range between 0 and 1, and are thus about an order of magnitude lower than the unconditional biases, which are mostly between about 0 and 10. Conditional biases are typically very low during most of the year (Fig. 3), and do not vary notably by lead time for most of the models (Fig. 4). One visible

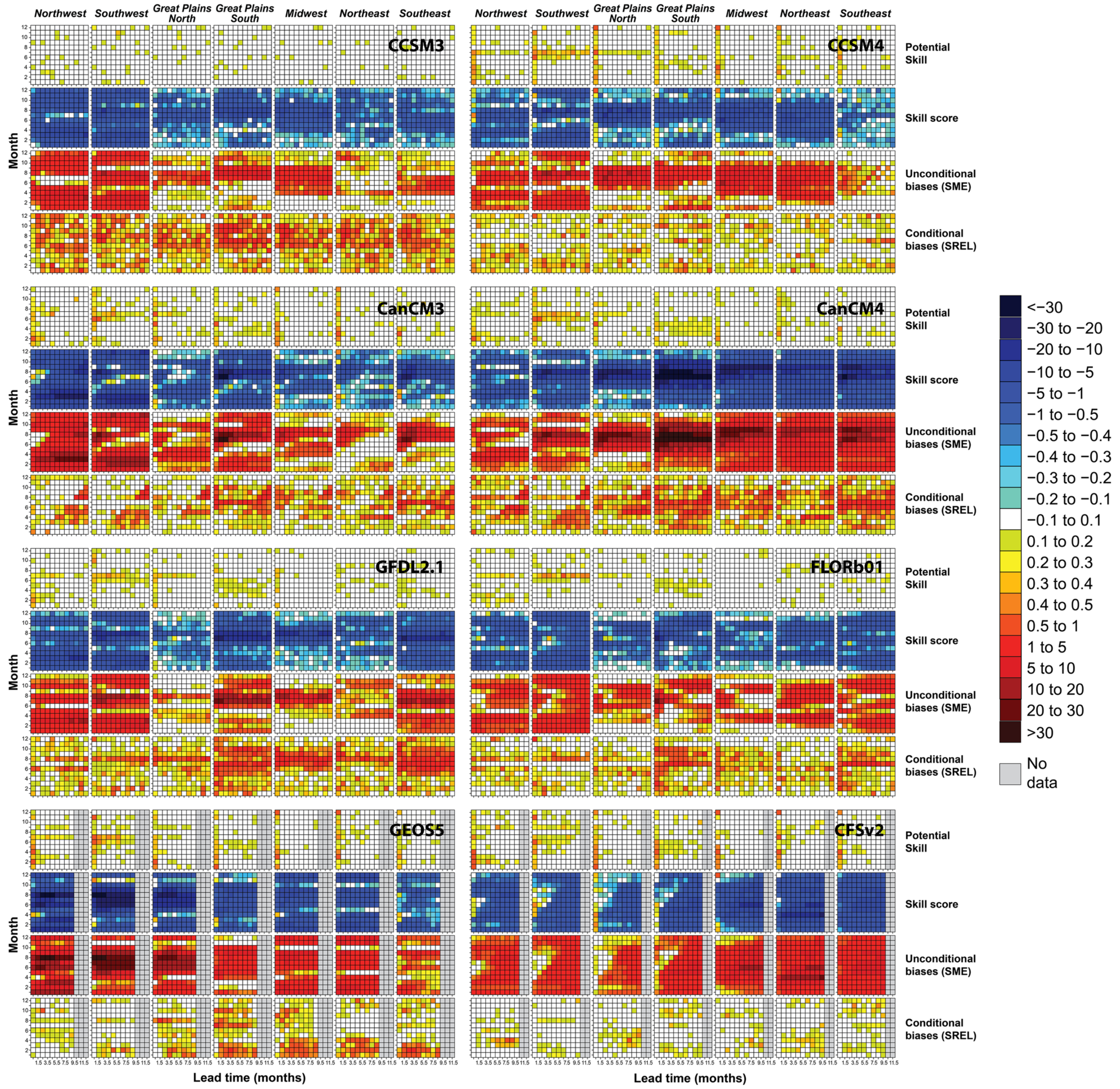

Fig. 4 Skill of the eight individual GCMs in forecasting temperature (CCSM3, CCSM4, CanCM3, CanCM4, GFDL2.1, FLORb01, GEOS5, and CFSv2). *The layout of the panels* is the same as described in Fig. 3. Note that GEOS-5 and CFSv2 only have 9 and 10 lead times, respectively, in comparison with the other models

exception is the case of CanCM3 and CanCM4, which exhibit a 'stepped' appearance, so the conditional biases increase (become redder) as lead time increases. These biases in the Canadian models tend to develop more rapidly in the earlier months of the year than in the later months (see CanCM4 conditional biases in the Southwest, for an example). Some of the other models, like GFDL2.1 and GEOS5, also reveal some seasonality in their conditional biases.

4.1.2 *Precipitation*

Precipitation forecasts generally have lower potential skill than temperature (Fig. 3b), as expected and found in other studies, due to the greater variability in rainfall patterns (e.g., Infanti and Kirtman 2016). The eight-model ensemble mean has better skill than any of the individual models (Fig. 3b vs. Fig. 5), and the regions with the highest eight-model potential skill reflect the ability of the most skillful models (e.g., CCSM4, CFSv2 in the Southeast). However, the individual models display relatively low potential skill, especially after the 0.5-month lead [consistent with results found by Mo and Lyon (2015)], and little spatial variation on the regional scale (Fig. 5). The models with the poorest forecasting ability (e.g., CCSM3 and FLORb01) do not even display potential skill at the 0.5-month lead. Other models (e.g., CCSM4, the Canadian models, GEOS5 and CFSv2) display some skill at longer lead times, but only for specific months, such as July in the Northwest (for CCSM4, GFDL2.1, CanCM4, and FLORB01), or May in the Southwest (e.g., CanCM4, GEOS5).

Similarly to temperature, the skill score for precipitation is mainly driven by unconditional biases in the models: the positive unconditional biases (red patterns) are mirrored by the negative skill score (blue patterns). Overall, however, the skill score for precipitation displays slightly less extreme (positive and negative) values than for temperature. This 'subdued' behavior may be caused by the greater variability in precipitation rates (i.e., lower agreement among forecast patterns) in space and time, for different months, lead times, and models. In other words, because of the small spatial scales of precipitation forecasts (compared to temperature), better results might be achieved by focusing on smaller spatial regions than the seven broad regions used here.

Interestingly, the seasonality of model skill also varies regionally for precipitation, but is different from the regional patterns for temperature. For the Northwest, Southwest, Great Plains North, Midwest, and Northeast regions, the highest unconditional biases in the precipitation forecasts tend to occur more frequently (lower skill) in the winter months (Fig. 3b). The Great Plains South and Southeast regions, on the contrary, display lower unconditional biases (higher skill) in the winter months. This finding is consistent with that of Infanti and Kirtman (2014) for the southeastern United States, and suggests that improved model skill in the winter months may well be related to the influence of ENSO (e.g., Mo and Lyon 2015; Roundy et al. 2015). In some regions, the unconditional biases tend to increase as the lead time of the forecast increases, so the color maps become progressively redder towards the right side of the plots (e.g., the Northwest region for CanCM3, FLORb01, or CFSv2) (Fig. 5). Elsewhere the biases decrease with increasing lead time (e.g., Great Plains South, FLORb01). All eight models display considerable biases, but CCSM3 displays the largest biases, specifically in the Great Plains North region.

The conditional biases are again much lower than the unconditional biases, and much more variable, displaying little regularity by month or by lead time. Some months display slightly higher conditional biases (e.g., April or July), but such patterns are infrequent. CCSM3 and CCSM4 have the largest conditional biases (red), followed by GFDL2.1, while the Canadian models, GEOS5 and CFSv2 tend to show lower conditional biases. Regionally, there seem to be slightly greater biases in the Southwest and Great Plains North.

4.2 Individual extreme events

4.2.1 Floods

We evaluate the skill of the eight NMME models in predicting four flood events (the 1993 July–August flood, the 1995 January–March flood, the 2008 June–August flood, and March 2010) by comparing the observed climatology (Fig. 2a–d) to the model precipitation forecasts (positive anomalies). As a caveat, it should first be conceded that we do not expect the models to reflect the observed historical precipitation anomalies perfectly over such broad spatial scales, even in the best-case scenarios, because of convection patterns that occur at local scales (and that cannot be captured in the same way as extreme temperature anomalies, which exhibit more spatially-consistent patterns). Overall, results indicate that the four flood events were relatively poorly predicted by all eight models (Fig. 6a–d). The 1993 Midwest flooding stands out as the least poorly forecast, since all models with the exception of CCSM3 predicted positive anomalies. CanCM4, CCSM4, FLORb01, CFSv2 and CanCM3 all forecast anomalies that were more than 2 times greater than their own average seasonal value (Fig. 6a). However, the actual historical anomaly was much greater than any of the predicted values, at 3.80. Generally speaking, skillful predictions tend to occur in regions that have strong air-sea coupling, so the initial condition of the atmosphere plays an important role in the forecast for several months (Materia et al. 2014). In the case of the 1993 flood, it is likely that the skill of the models resulted from the strength of the El Niño, which displaced the storm track over the central United States, with atmospheric rivers transporting large amounts of moisture from the Gulf of Mexico over the Mississippi River basin (Trenberth and Guillemot 1996; Lavers and Villarini 2013). The El Niño conditions also likely explain why the ability of the eight models to predict the 1993 flood visibly decreased here with initialization time (i.e., the further ahead of the event, the less able the models were to forecast the high rainfall).

The other three events were relatively less well forecast, although CFSv2 performed better than all other models in 2008 (Fig. 6c), as did FLORb01 in 2010 at the shortest lead time (Fig. 6d). The observed event anomalies (PRISM data) were of 2.34, 2.55, and 2.78 while the model forecasts, at best, attained 1.8 (GFDL2.1–1995 flood), 1.5

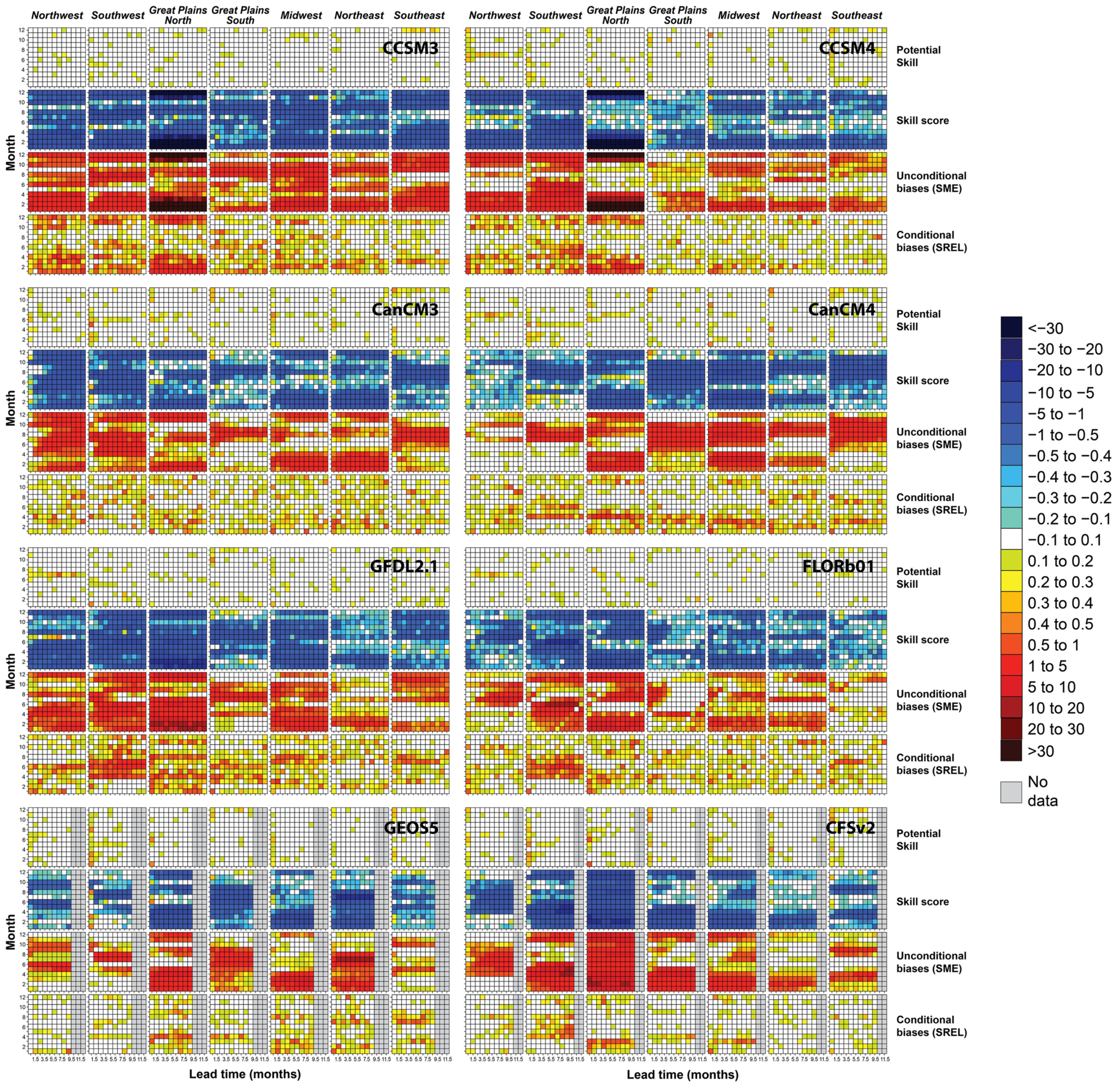

Fig. 5 Skill of the eight individual GCMs in forecasting precipitation (CCSM3, CCSM4, CanCM3, CanCM4, GFDL2.1, FLORb-01, GEOS5, and CFSv2). *Layout of the panels* is the same as described in Fig. 4

(CFSv2–2008 flood) and 2.3 (CFSv2–2010 flood), but somewhat fortuitously, since some of the highest anomalies were predicted many months ahead of the actual events. In fact, for all three of these flood events (Fig. 6b–d), the eight-model ensemble mean was near or below zero, and half of the individual model forecasts predicted a "drier-than-average" season. Figure 6b–d indicates that most models fluctuate between positive and negative anomalies, and in 2008 were mostly wrong, predicting a drier-than-average season overall; as for the other flood events, the predicted anomalies were as low as −1.5 (1995 flood–GEOS5), −2.5 (2008 flood–CanCM4), and −2.4 (CFSv2–2010 flood). Thus, no model consistently outperformed any of the others, and no single model was reliable in terms of consistently predicting these three flood events (Fig. 6b–d).

4.2.2 *Droughts*

Droughts tend to develop more slowly than floods, as it can take between five and eight months for the water deficit to

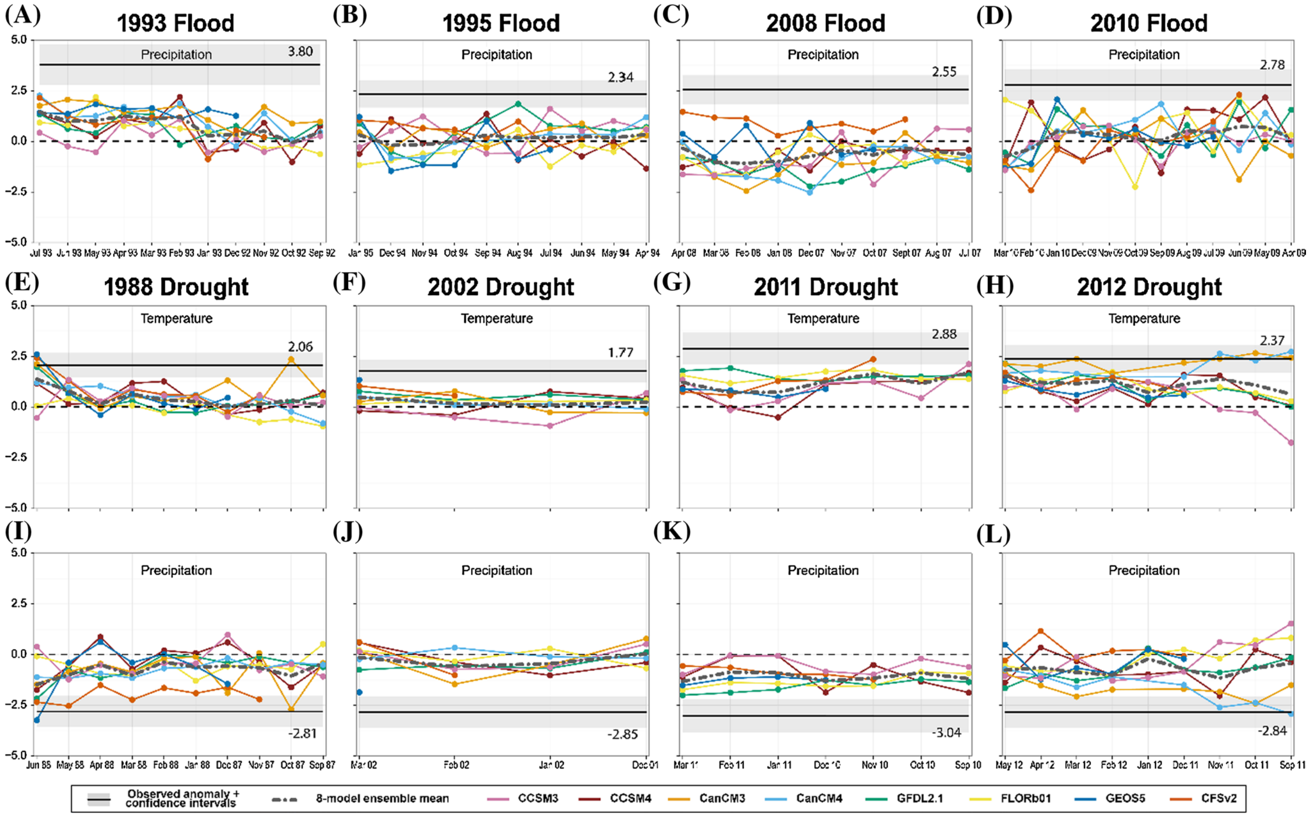

Fig. 6 Skill of the eight NMME models in predicting four flood and four drought events, in comparison with the observed climatology. Flood and drought events (**a–l**) are the same as in Fig. 2. Thick *horizontal black line* indicates the PRISM observed climatological anomaly, with 95 % confidence intervals indicated as *shaded grey rectangles* in the background. NMME anomalies are indicated as *colored lines*. *Long/short-dashed black* line indicates the eight-model ensemble mean. *Panels* **f** and **j:** note that GEOS5 only exhibits one lead time and CFSv2 two, because the event lasted for nine months and these models only issue nine- and 10-month lead times, respectively. *Panels* **g** and **k**: note that the two Canadian models have data gaps in 2011, so are not included in the evaluation of the 2011 March–August drought.

drop beneath a certain threshold and begin a drought (Mo 2011). Hence, skillful intraseasonal to interannual forecasts may prove particularly vital ahead of major drought events. Additionally, droughts also tend to be more predictable than floods because of the influence of the Pacific Decadal Oscillation (PDO) and the Atlantic Multi-decadal Oscillation (AMO) (McCabe et al. 2004) and the effects of land surface/atmosphere coupling (e.g., Koster et al. 2006, Seneviratne et al. 2010). Thus, droughts that are strongly influenced by initial conditions tend to be well-forecast (Roundy and Wood 2015).

Here we evaluate the ability of NMME models to predict droughts as high temperature anomalies (excess heat Fig. 2e–h) on the one hand, and low precipitation anomalies (lack of rainfall, Fig. 2i–l) on the other, in comparison with the observed climatology (red shades for excess temperature, blue shades for lack of rain). The comparison between temperature and precipitation predictions for drought events also allows us to assess whether the NMME models are more accurate in predicting excess heat or deficient rainfall, and to what extent temperature actually contributed to drought severity for each of these events. For instance, the 2014 California drought was driven principally by low precipitation, but intensified by high temperatures (Shukla et al. 2015).

The comparison between observed extreme temperature and observed extreme precipitation anomalies reveals a relatively good overlap in spatial extents (Fig. 2) with the exception of the 2002 March-November drought, which was also the least predictable of the four droughts (only small isolated parts of the south-east and south-west United States were affected by the positive temperature anomaly, Fig. 2f). During droughts, strong precipitation deficits and high heat anomalies tend to occur over the same regions, as was the case during the 1934, 1936, 2011 and 2012 events (Donat et al. 2016). The discrepancies between temperature and precipitation patterns tend to be relatively limited in space and are mainly caused by the noise associated with the precipitation signal; for instance, localized thunderstorms occurring in spring and summer may influence the rainfall anomalies computed for an entire season.

Of the four drought events, it appears that the 1988 drought was remarkably well predicted at the shortest initialization time by four models (GEOS5, CFSv2, CanCM3 and GFDL2.1) in terms of high temperature (Fig. 6e). The first two of those models actually exceeded the observed anomaly (PRISM = 2.1), with forecast values of 2.6 and 2.4. However, the skill of all models decreased rapidly with increasing lead time, indicating that they were unable to predict the event more than one month ahead of its actual occurrence. For the same event, the precipitation forecasts (lack of rainfall) were also relatively successful in June 1988 (anomaly values of −3.2 for GEOS5, −2.3 for CFSv2, −2.2 for GFDL2.1, in comparison with the observed −2.8) but the skill declined when predicted further ahead (Fig. 6e). CCSM3 performed the least well among all models, while CanCM3 predicted the drought successfully both in terms of temperature and precipitation eight months ahead of the actual event (Fig. 6e). Overall, the good predictability of the 1988 drought is likely a result of the strong La Niña conditions (e.g., Trenberth and Guillemot 1996) that occurred in conjunction with a cooling phase of the PDO and the warming phase of the AMO (McCabe et al. 2004).

The other three droughts were relatively less well predicted. For 2002, the eight-model ensemble mean was close to climatology (anomaly value around 0), and in the month preceding the event, only GEOS5 predicted a positive temperature anomaly of 1.3 versus 1.77 for the observed climatology, while half of the models actually predicted excess rainfall (Fig. 6f). In 2011, the March-August forecasts were slightly more accurate, likely because the drought resulted from a strong La Niña (Seager and Hoerling 2014) and the mean flow moisture divergence anomalies driven by the negative North Atlantic Oscillation of the previous winter (Seager et al. 2014). GFDL2.1 and FLORB01 both consistently predicted high positive temperature anomalies and low negative precipitation anomalies, even at the longer times before the event, and the eight-model ensemble mean correctly predicted positive/negative anomalies (Fig. 6g). Last, the 2012 drought was relatively well predicted, with slightly better results for temperature than precipitation. However, contrary to model forecasts, Pacific sea surface temperature (SST) did not play a major role in the drought (Kumar et al. 2013; Hoerling et al. 2014), so the skillful prediction of the drought was in fact "fortuitous, due to the erroneous coupling with pan-Pacific SSTs" (Kam et al. 2014). CanCM3 and CanCM4 display good results, but they become less skillful as one approaches the beginning of the event (Fig. 6h). As suggested by Roundy and Wood (2015), the varying skill of drought forecasts among years implies that they are driven by different mechanisms; atmospheric and land initial conditions, SST and radiative forcing may have varying influences to strengthen/weaken the predictability of events (Jia et al. 2016).

Overall, it is interesting to note that the precipitation and temperature forecasts are more similar than one might expect in terms of their ability to forecast the extreme events. In fact, comparing the positive temperature anomalies with the negative precipitation anomalies (Fig. 6e–l) indicates that seasonal precipitation and temperature forecasts do tend to reflect one another to a certain extent. When the temperature forecast is skillful, the precipitation forecast tends to be also (e.g., GEOS5 and CanCM3 in 1988, or GFDL in 2011, Fig. 6g). Likewise, the lack of skill is also mirrored for both temperature and precipitation (e.g., CCSM3 in 2011, Fig. 6g).

Comparing our results with historical ENSO forecasts suggests that when the land surface/atmosphere interaction is well represented, events tend to be better predicted; hence, the lack of land surface/atmosphere coupling in 2002 may explain why the drought was poorly predicted and why there was little consistency between temperature and precipitation patterns. Therefore, as different models have different abilities depending on seasonality and lead times, strategic multi-model averaging procedures may help increase the forecasting skill of these extreme flood and drought events (e.g., Luo and Wood 2008; Bradley et al. 2015), especially in locations with strong antecedent ENSO signal (e.g., Yuan and Wood 2013).

5 Summary and conclusions

By decomposing the skill score of the individual climate models into potential skill, unconditional and conditional biases, we have assessed the strengths and weaknesses of the eight GCM ensemble means and of the eight-model ensemble mean over a range of lead times and initialization months. Our findings provide a diagnostic tool that can provide model developers with feedback about the strengths and weaknesses of their models, and help develop model-averaging strategies.

The results can be summarized as follows:

1. The highest potential skill in temperature and precipitation forecasts is displayed at the shortest lead time (0.5 month) and declines rapidly thereafter. For both temperature and precipitation, the potential skill of the eight-model ensemble mean tends to surpass the skill of the best model within the ensemble. However, there is room for more sophisticated model averaging approaches (i.e., weighting individual models based on their strengths and weaknesses) to improve the model ensemble skill. Overall, the skill score is quite low for

all models. The eight-model ensemble displays positive values mostly at the shortest lead times, and there is not one model that clearly outperforms any of the others.

2. The biases in these eight models are predominantly unconditional (SME), with strong seasonal- and lead-dependent biases driving the negative skill scores (which are likely dependent on the initialization conditions in different regions and seasons). For temperature, in most regions, the unconditional biases tend to be the lowest in the winter/spring months, and to increase in the summer (while the reverse is true in the Northwest and Southwest). For precipitation, the unconditional biases tend to be the lowest in the summer and fall (while the reverse is true in the Great Plains South and Southeast). Thus, it appears that the skill of these forecasts could be improved by attenuating the unconditional biases that are specific to certain regions and seasons. The conditional biases (SREL) are generally about an order of magnitude smaller than the unconditional biases, and display much more variability across all regions, months, and lead times.
3. Overall, the skill of the eight NMME models in predicting four flood events and four drought events shows some inconsistencies. The droughts tend to be better forecast than the floods, even in terms of precipitation, likely because they are more tightly connected to SST-driven climate conditions (McCabe et al. 2004). However, air-sea coupling may also lead to fortuitous forecasts (Kam et al. 2014) as some of the best forecasts seem to occur randomly, sometimes many months ahead of the actual events. While some models are able to predict specific events well, and sometimes months in advance (e.g., CFSv2 for the 1988 drought, or CanCM3 for the 2012 drought), no model consistently outperforms any of the others, or is reliable in terms of consistently predicting events.
4. Perhaps more unexpectedly, although average temperature forecasts tend to outperform average precipitation forecasts, we find that the seasonal positive temperature anomalies for the droughts are not more accurately predicted than negative precipitation anomalies. In fact, the ability of the models to forecast drought is remarkably similar in terms of temperature and precipitation. Generally speaking, most forecast anomalies are at least one standard deviation beneath the observed anomaly, suggesting that the ensemble means of models cannot accurately forecast strongly deviating departures from the climatology over such broad spatial scales. Thus, in future work, extreme values may be better forecast by individual model members and over smaller regions, particularly in the case of precipitation, to avoid the influence of noise arising from localized convective events.

These findings highlight some of the strengths and weaknesses of the NMME models across all lead times, months, and for seven major regions of the United States. One of the remaining challenges is our ability to extend precipitation forecast skill beyond the shortest lead time, as is recognized in similar studies (Wood et al. 2015). The overall skill of the eight-model ensemble shows promise for multi-model averaging procedures (e.g., Luo et al. 2007; Bradley et al. 2015) that might enable more skillful forecasts at longer lead times. Moreover, future studies should examine whether it is possible to utilize these precipitation and temperature forecasts for impact studies including seasonal discharge forecasting.

Acknowledgments The authors thank the NMME program partners and acknowledge the help of NCEP, IRI and NCAR personnel in creating, updating and maintaining the NMME archive, with the support of NOAA, NSF, NASA and DOE. This study was supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections Program, Grant#NA15OAR4310073. Gabriele Villarini also acknowledges financial support from the USACE Institute for Water Resources and from Grant/Cooperative Agreement Number G11 AP20079 from the United States Geological Survey. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of NOAA, USACE or of the USGS.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Becker E, Van den Dool H, Zhang Q (2014) Predictability and forecast skill in NMME. J Clim 27(15):5891–5906. doi:10.1175/JCLI-D-13-00597.1

Boer GJ, Kharin VV, Merryfield WJ (2013) Decadal predictability and forecast skill. Clim Dyn 41(7–8):1817–1833. doi:10.1007/s00382-013-1705-0

Bradley AA, Schwartz SS (2011) Summary verification measures and their interpretation for ensemble forecasts. Mon Weather Rev 139(9):3075–3089. doi:10.1175/2010MWR3305.1

Bradley AA, Habib M, Schwartz SS (2015) Climate index weighting of ensemble streamflow forecasts using a simple bayesian approach. Water Res Res 51(9):1–49. doi:10.1002/2014WR016811

Daly C, Gibson WP, Taylor GH, Johnson GL, Pasteris P (2002) A knowledge-based approach to the statistical mapping of climate. Clim Res 22(9):99–113

DelSole T, Tippett MK (2014) Comparing Forecast Skill. Mon Weather Rev 142(12):4658–4678. doi:10.1175/MWR-D-14-00045.1

Delsole T, Nattala J, Tippett MK (2014) Skill improvement from increased ensemble size and model diversity. Geophys Res Lett 41(20):7331–7342. doi:10.1002/2014GL060133

Delworth TL, Broccoli AJ, Rosati A, Stouffer RJ, Balaji V, Beesley JA, Coke WF, Dixon KW, Dunne J, Dunne KA, Durachta JW (2006) GFDL's CM2 global coupled climate models. Part I:

formulation and simulation characteristics. J Clim 19(5):643–674. doi:10.1175/JCLI3629.1

Donat MG, King AD, Overpeck JT, Alexander LV, Durre I, Karoly DJ (2016) Extraordinary heat during the 1930s US Dust Bowl and associated large-scale conditions. Clim Dyn 46(1–2):413–426. doi:10.1007/s00382-015-2590-5

Hagedorn R, Doblas-Reyes FJ, Palmer TN (2005) The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept. Tellus A 57(3):219–233. doi:10.1111/j.1600-0870.2005.00103.x

Hashino T, Bradley AA, Schwartz SS (2007) Evaluation of bias-correction methods for ensemble streamflow volume forecasts. Hydrol Earth Syst Sci Discuss 3(2):561–594

Hijmans R (2015) Raster: geographic data analysis and modeling. R Package version 2.4-18. http://CRAN.R-project.org/package=raster

Hoerling M, Eischeid J, Kumar A, Leung R, Mariotti A, Mo K, Schubert S, Seagar R (2013) Causes and predictability of the 2012 great plains drought. Bull Am Meteorol Soc 95(2):269–282. doi:10.1175/BAMS-D-13-00055.1

Infanti JM, Kirtman BP (2014) Southeastern U.S. rainfall prediction in the North American multi-model ensemble. J Hydrometeorol 15(2):529–550. doi:10.1175/JHM-D-13-072.1

Infanti JM, Kirtman BP (2016) North American rainfall and temperature prediction response to the diversity of ENSO. Clim Dyn 46(9–10):3007–3023. doi:10.1007/s00382-015-2749-0

Jia L, Yang X, Vecchi GA, Gudgel RG, Delworth TL, Rosati A, Stern WF, Wittenberg AT, Krishnamurthy L, Zhang S, Msadek R, Kapnick S, Underwood S, Zeng Fanrong, Anderson WhitG, Balaji V, Dixon K (2015) Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. J Clim 28(5):2044–2062. doi:10.1175/JCLI-D-14-00112.1

Jia L, Vecchi GA, Yang X, Gudgel RG, Delworth TL, Stern WF, Paffendorf K, Underwood SD, Zeng F (2016) The roles of radiative forcing, sea surface temperatures, and atmospheric and land initial conditions in U.S. summer warming episodes. J Clim 29(11):4121–4135. doi:10.1175/JCLI-D-15-0471.1

Kam J, Sheffield J, Yuan X, Wood EF (2014) Did a skillful prediction of sea surface temperatures help or hinder forecasting of the 2012 Midwestern US drought? Environ Res Lett 9(3):1–9. doi:10.1088/1748-9326/9/3/034005

Karl TR, Melillo JM, Peterson TC (eds) (2009) Global climate change impacts in the United States. Cambridge University Press, Cambridge

Kirtman BP, Min D (2009) Multimodel ensemble ENSO prediction with CCSM and CFS. Mon Weather Rev 137(9):2908–2930. doi: 10.1175/2009MWR2672.1

Kirtman BP, Min Du, Infanti JM, Kinter JL III, Paolino DA, Zhang Q, van den Dool H, Saha S, Pena Mendez M, Becker E, Peng P, Tripp P, Huang J, DeWitt DG, Tippett MK, Barnston AG, Li S, Rosati A, Schubert SD, Rienecker M, Suarez M, Li ZE, Marshak J, Lim Y-K, Tribbia J, Pegion K, Merryfield WJ, Denis B, Wood EF (2014) The North American multi-model ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull Am Meteorol Soc 95(4):585–601. doi:10.1175/BAMS-D-12-00050.1

Koster RD, Sud YC, Guo Z, Dirmeyer PA, Bonan G, Oleson KW, Chan E, Verseghy D, Cox P, Davies H, Kowalczyk E (2006) GLACE: the global land–atmosphere coupling experiment. Part I: overview. J Hydrometeol 7(4):590–610. doi:10.1175/JHM510.1

Kumar A, Chen M, Hoerling M, Eischeid J (2013) Do extreme climate events require extreme forcings? Geophys Res Lett 40(13):3440–3445. doi:10.1002/grl.50657

Kumar A, Peng P, Chen M (2014) Is there a relationship between potential and actual skill? Mon Weather Rev 142(6):2220–2227. doi:10.1175/MWR-D-13-00287.1

Kunkel KE, Karl TR, Brooks H, Kossin J, Lawrimore JH, Arndt D, Bosart L, Changnon D, Cutter SL, Doesken N, Emanuel K, Groisman PY, Katz RW, Knutson T, O'brien J, Paciorek CJ, Peterson TC, Redmond K, Robinson D, Trapp J, Vose R, Weaver S, Wehner M, Wolter K, Wuebbles D (2013) Monitoring and understanding trends in extreme storms: state of knowledge. Bull Am Meteorol Soc 94(4):499–514. doi:10.1175/BAMS-D-11-00262.1

Lavers DA, Villarini G (2013) Atmospheric rivers and flooding over the central United States. J Clim 26(20):7829–7836

Lawrence DM, Oleson KW, Flanner MG, Fletcher CG, Lawrence PJ, Levis S, Swenson SC, Bonan GB (2012) The CCSM4 land simulation, 1850–2005: assessment of surface climate and new capabilities. J Clim 25(7):2240–2260

Luo LF, Wood EF (2008) Use of Bayesian merging techniques in a multimodel seasonal hydrologic ensemble prediction system for the eastern United States. J Hydrometeorol 9(5):866–884

Luo LF, Wood EF, Pan M (2007) Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. J Geophys Res Atmos. doi:10.1029/2006JD007655

Ma F, Ye A, Deng X, Zhou Z, Liu X, Duan Q, Xu J, Miao C, Di Z, Gong W (2015a) Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. Int J Climatol 36(1):132–144. doi:10.1002/joc.4333

Ma F, Yuan X, Ye A (2015b) Seasonal drought predictability and forecast skill over China. J Geophys Res Atmos 120(16):8264–8275. doi:10.1002/2015JD023185

Materia S, Borrelli A, Bellucci A et al (2014) Impact of atmosphere and land surface initial conditions on seasonal forecasts of global surface temperature. J Clim 27(24):9253–9271. doi:10.1175/JCLI-D-14-00163.1

McCabe GJ, Palecki MA, Betancourt JL (2004) Pacific and Atlantic Ocean influences on multidecadal drought frequency in the United States. Proc Natl Acad Sci 101(12):4136–4141. doi:10.1073/pnas.0306738101

Merryfield WJ, Lee W-S, Boer GJ, Kharin VV, Scinocca JF, Flato GM, Ajayamohan RS, Fyfe JC, Tang Y, Polavarapu S (2013) The Canadian seasonal to interannual prediction system. Part I: models and initialization. Mon Weather Rev 141(8):2910–2945. doi:10.1175/MWR-D-12-00216.1

Misra V, Li H (2014) The seasonal climate predictability of the Atlantic Warm Pool and its teleconnections. Geophys Res Lett 41(2):661–666. doi:10.1002/2013GL058740

Mo K (2011) Drought onset and recovery over the United States. J Geophys Res. doi:10.1029/2011JD016168

Mo KC, Lettenmaier DP (2014) Hydrologic prediction over the conterminous United States using the national multi-model ensemble. J Hydrometeorol 15(4):1457–1472. doi:10.1175/JHM-D-13-0197.1

Mo KC, Lyon B (2015) Global meteorological drought prediction using the North American multi-model ensemble. J Hydrometeorol 16(3):1409–1424. doi:10.1175/JHM-D-14-0192.1

Molod A, Takacs L, Suarez M, Bacmeister J, Song I-S, Eichmann A (2012) The GEOS-5 atmospheric general circulation model: mean climate and development from MERRA to Fortuna. In: Technical report series on global model data assimilation, vol 28. NASA Goddard Space Flight Cent., Greenbelt, p 175

Murphy AH, Winkler RL (1992) Diagnostic verification of probability forecasts. Int J Forecast 7(4):435–455. doi:10.1016/0169-2070(92)90028-8

National Research Council (US) (2010) Committee on assessment of intraseasonal to interannual climate prediction and predictability. In: Assessment of intraseasonal to interannual climate prediction and predictability. National Academies Press

Pierce D (2014) ncdf4: Interface to Unidata netCDF (version 4 or earlier) format data files. R Package Version 1.12. http://dwpierce.com/software

Roundy JK, Wood E (2015) The attribution of land-atmosphere interactions on the seasonal predictability of drought. J Hydrometeorol 16(2):793–810. doi:10.1175/JHM-D-14-0121.1

Roundy JK, Yuan X, Schaake J, Wood EF (2015) A framework for diagnosing seasonal prediction through Canonical event analysis. Mon Weather Rev 143(6):2404–2418. doi:10.1175/MWR-D-14-00190.1

Saha S, Moorthi S, Wu X, Wang J, Nadiga S, Tripp P, Behringer D, Hou Y-T, Chuang H-Y, Iredell M, Ek M, Meng J, Yang R, Peña Mendez M, van den Dool H, Zhang Q, Wang W, Chen M, Becker E (2014) The NCEP climate forecast system version 2. J Clim 27(6):2185–2208. doi:10.1175/JCLI-D-12-00823.1

Seager R, Hoerling M (2014) Atmosphere and ocean origins of North American droughts. J Clim 27(12):4581–4606. doi:10.1175/JCLI-D-13-00329.1

Seager R, Goddard L, Nakamura J, Henderson N, Lee DE (2014) Dynamical causes of the 2010/11 Texas-Northern Mexico drought. J Hydrometeorol 15(1):39–68. doi:10.1175/JHM-D-13-024.1

Seneviratne TC, Davin EL, Hirschi M, Jaeger EB, Lehner I, Orlowsky B, Teuling AJ (2010) Investigating soil moisture-climate interactions in a changing climate: a review. Earth Sci Rev 99(3):125–161

Shukla S, Safeeq M, AghaKouchak A, Guan K, Funk C (2015) Temperature impacts on the water year 2014 drought in California. Geophys Res Lett 42(11):4384–4393. doi:10.1002/2015GL063666

Stedinger JR, Vogel RM, Foufoula-Georgiou E (1993) Chapter 18, frequency analysis of extreme events. In: Maidment DR (ed) Handbook of Hydrology. McGrawHill Book Company, New York

Thober S, Kumar R, Sheffield J, Mai J, Schäfer D, Samaniego L (2015) Seasonal soil moisture drought prediction over Europe using the North American multi-model ensemble (NMME). J Hydrometeorol 16(6):2329–2344. doi:10.1175/JHM-D-15-0053.1

Tian D, Martinez CJ, Graham WD, Hwang S (2014) Statistical downscaling multi-model forecasts for seasonal precipitation and surface temperature over the Southeastern United States. J Clim 27(22):8384–8411. doi:10.1175/JCLI-D-13-00481.1

Trenberth, K.E., & Guillemot, C.J. (1996). Physical processes involved in the 1988 drought and 1993 floods in North America. *Journal of Climate*. *9*(6), 1288–1298. doi:10.1175/1520-0442(1996)009<1288:PPIITD>2.0.CO;2

Vecchi GA, Delworth T, Gudgel R, Kapnick S, Rosati A, Wittenberg A, Zeng F, Anderson W, Balaji V, Dixon K, Jia L, Kim H-S, Krishnamurthy L, Msadek R, Stern WF, Underwood SD, Villarini G, Yang X, Zhang S (2014) On the seasonal forecasting of regional tropical cyclone activity. J Clim 27(21):7994–8016. doi:10.1175/JCLI-D-14-00158.1

Vernieres G, Rienecker MM, Kovach R, Keppenne CL (2012) The GEOS-iODAS: description and evaluation. In: GEOS5 technical report NASA/TM-2012-104606, vol 30. 61 pp. [Available online at http://gmao.gsfc.nasa.gov/pubs/docs/Vernieres589.pdf.]

Wang H (2014) Evaluation of monthly precipitation forecasting skill of the National Multi-model Ensemble in the summer season. Hydrol Process 28(15):4472–4486. doi:10.1002/hyp.9957

Wolter K, Timlin MS (2011) El Nino/Southern Oscillation behaviour since 1871 as diagnosed in an extended multivariate ENSO index (MEI.ext). Int J Climatol 31(7):1074–1087. doi:10.1002/joc.2336

Wood EF, Schubert SD, Wood AW, Peters-Lidard CD, Mo KC, Mariotti A, Pulwarty RS (2015) Prospects for advancing drought understanding, monitoring, and prediction. J Hydrometeorol 16(4):1636–1657. doi:10.1175/JHM-D-14-0164.1

Younas W, Tang Y (2013) PNA predictability at various time scales. J Clim 26(22):9090–9114. doi:10.1175/JCLI-D-12-00609.1

Yuan X, Wood EF (2013) Multimodel seasonal forecasting of global drought onset. Geophys Res Lett 40(18):4900–4905. doi:10.1002/grl.50949

Zhang S, Harrison MJ, Rosati A, Wittenberg A (2007) System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. Mon Weather Rev 135(10):3541–3564. doi:10.1175/MWR3466.1