Evaluation of Statistical Downscaling of North American Multimodel Ensemble Forecasts over the Western United States

RENAUD BARBERO

School of Civil Engineering and Geosciences, Newcastle University, Newcastle, United Kingdom

JOHN T. ABATZOGLOU AND KATHERINE C. HEGEWISCH

Department of Geography, University of Idaho, Moscow, Idaho

(Manuscript received 17 June 2016, in final form 22 November 2016)

ABSTRACT

The skill of two statistical downscaled seasonal temperature and precipitation forecasts from the North American Multimodel Ensemble (NMME) was evaluated across the western United States at spatial scales relevant to local decision-making. Both statistical downscaling approaches, spatial disaggregation (SD) and bias correction spatial disaggregation (BCSD), exhibited similar correlative skill measures; however, the BCSD method showed superior tercile-based skill measures since it corrects for variance deflation in NMME ensemble averages. Geographic and seasonal variations in downscaled forecast skill revealed patterns across the complex topography of the western United States not evident using coarse-scale skill assessments, particularly in regions subject to inversions and variability in orographic precipitation ratios. Similarly, differences in the skill of cool-season temperature and precipitation forecasts issued when the fall El Niño–Southern Oscillation (ENSO) signal was strong versus ENSO-neutral years were evident across topographic gradients in the northwestern United States.

1. Introduction

Seasonal climate forecasts have the potential to help mitigate detrimental climate impacts and capitalize on beneficial climate impacts to society and the environment (e.g., Steinemann 2006; Fraisse et al. 2006). Just as skillful forecasts can be used to mitigate losses associated with climate variability (Troccoli 2010), the misuse of unskillful forecasts can be costly and hinder our subsequent ability to use forecasts (Hartmann et al. 2002). The use of seasonal forecasts is currently limited by not only forecast quality, but also the spatial mismatch between model output and user needs, which results in a lack of knowledge of how skillful these forecasts are at local scales (e.g., Doblas-Reyes et al. 2013).

Seasonal climate predictability is imparted by slower-evolving components of the climate system including sea surface temperature, snow, and soil moisture (Koster et al. 2010; Slingo and Palmer 2011). Seasonal climate forecasts include both empirical and dynamic approaches, and associated derivatives thereof, combined with expert judgment. Whereas empirical forecasts rely on observed relationships, dynamical models utilize either general circulation models (GCMs) or high-resolution regional models (Doblas-Reyes et al. 2013). The accuracy of seasonal climate forecasts varies across multiple dimensions including the modeling approach, variable of interest, region, season, and forecast lead time (e.g., Tian et al. 2014; Ma et al. 2016). Seasonal forecast skill is typically higher in the tropics (e.g., Graham et al. 2000; Goddard and Mason 2002; Lavers et al. 2009), typically higher for temperature than precipitation (e.g., Lavers et al. 2009), and typically attenuates with longer lead times. Forecast skill may also be contingent on large-scale modes of climate variability (Scaife et al. 2014). For example, improved seasonal forecast skill during ENSO years (Frías et al. 2010; Kim et al. 2012; Manzanas et al. 2014) may provide windows of opportunity for forecast users (Troccoli 2010).

Supplemental information related to this paper is available at the Journals Online website: http://dx.doi.org/10.1175/WAF-D-16-0117.s1.

Corresponding author e-mail: Dr. Renaud Barbero, renaud.barbero@ncl.ac.uk

DOI: 10.1175/WAF-D-16-0117.1

TABLE 1. Seasonal hindcast models from the NMME project used in this study. The $\overline{\text{MM}}$ was taken to be the average hindcast from the six-model ensemble average outputs listed in the table.

Model	No. of ensemble runs	Organization	Reference
CMC1	10	Canadian Meteorological Centre (CMC)	Merryfield et al. (2013)
CMC2	10	CMC	Merryfield et al. (2013)
CFSv2	24	National Oceanic and Atmospheric Administration/National Centers for Environmental Prediction (NOAA/NCEP)	Saha et al. (2014)
GFDL	10	Geophysical Fluid Dynamics Laboratory (GFDL)	Zhang et al. (2007)
GFDL- Forecast Version Low Ocean Resolution (GFDL-FLOR)	10	GFDL	Jia et al. (2015)
NASA	10	National Aeronautics and Space Administration (NASA)	Molod et al. (2012)

Whereas the assessment of seasonal climate forecasts at coarse resolution provides valuable information on the climate predictability at macroscales, such information may be obscured at spatial scales where this information is used for decision-making (Rayner et al. 2005) and for applications that need more refined information (e.g., Wood and Lettenmaier 2006). Heterogeneous skill across the western United States was found in previous assessments of seasonal climate forecasts at their native resolution (e.g., Peng et al. 2012; Roundy et al. 2015). In addition, the confluence of topography and large-scale climate can result in subregional variations in climate across parts of the western United States (e.g., Abatzoglou et al. 2009) that may alter model skill and the usability of forecasts. In this study, we evaluate the skill of statistically downscaled seasonal climate forecasts across the western United States given the potential utility of such forecasts for water resources and wildfire management (e.g., Hartmann et al. 2002). Recent studies have demonstrated the utility of statistical downscaling over raw GCM output for reproducing monthly climatic statistics and extremes for both climate scenarios (Ning et al. 2012; Ahmed et al. 2013; Ning et al. 2015) and seasonal climate forecasting (Yoon et al. 2012).

To address this gap, we assess in this study the skill of North American Multimodel Ensemble (NMME) seasonal forecasts of temperature and precipitation statistically downscaled to 4-km (1/24°) across the western United States from 1982 to 2010. Two simple statistical downscaling methods (e.g., Fowler et al. 2007) were compared: 1) spatial disaggregation (SD), which interpolates model anomalies to observed climatologies, and 2) bias correction spatial disaggregation (BCSD), which quantile maps model anomalies to observations and then interpolates these data to observed climatologies. A suite of models participating in NMME and two downscaling approaches are used to address three primary research questions: 1) How does forecast skill compare between the two simple statistical downscaling approaches? 2) What added value do downscaled forecasts provide over coarser-resolution forecasts across the complex terrain of the western United States? 3) How does ENSO phase alter the skill of seasonal forecasts?

2. Data and methods

a. Data

Retrospective forecasts (hindcasts) on a 1° grid were acquired from six models participating in the NMME system (Table 1). Model hindcast runs were initialized over the first several days each month. Rather than considering each ensemble member, we use the ensemble average of the 10–24 deterministic hindcasts for each model over the 29-yr period (1982–2010). We examined monthly temperature and precipitation across the western United States (west of 103°W) for lead times from 1 to 6 months. In addition to individual ensemble means for each model, we computed the multimodel mean (denoted $\overline{\text{MM}}$) as a simple average of ensemble means from individual models. Although the use of an ensemble averaging breaks down the consistency of physical processes in each model, this approach has proven to, on average, produce better skill than any single model (e.g., Kirtman et al. 2014). Notice that the $\overline{\text{MM}}$ is the mean of all downscaled models (i.e., we first downscaled the models and then computed the $\overline{\text{MM}}$).

Historical observations of monthly mean temperature and precipitation on a 1/24°-resolution (~4 km) grid were acquired from the surface meteorological dataset of Abatzoglou (2013) for 1979–2010. This dataset was also aggregated to 1° resolution in order to evaluate the skill of NMME hindcasts at their native resolution (see the online supplement to this article) in order to compare the differences that arise at local scales as a result of the downscaling.

b. Methods

Two statistical downscaling methods were used to bridge the scale between the output of NMME models and the 4-km resolution of the observations: SD and BCSD. We focus here on these statistical downscaling approaches as they are widely used and relatively simple to apply (e.g., Ning et al. 2012; Abatzoglou and Brown 2012; Ahmed et al. 2013). Both methods use monthly hindcast anomalies, which are computed relative to the model climatology (1982–2010). The SD method, analogous to the delta approach (e.g., Fowler et al. 2007), interpolates these anomalies to the 4-km grid, whereas the BCSD method first bias corrects the anomalies using empirical quantile mapping (e.g., Wood et al. 2002) thereby adjusting these data to the statistical distribution of the coarse-resolution observed distribution (1982–2010) before interpolating the output to the 4-km grid. As a final step for both methods, anomalies are converted to raw values using monthly climatology from observations. Note that resultant time series of SD and BCSD are highly correlated to the raw model outputs on monthly time scales, but may differ when aggregated over multiple months because of differences in the background climatology.

For each model we considered a three-dimension hindcast matrix for monthly temperature and precipitation that considers 1) each calendar month of the year, 2) lead times of 1–6 months, and 3) data aggregated over a period from one to six consecutive months contingent on lead time. For example, a seasonal hindcast made in October includes hindcasts for individual months from October to March and all possible combinations of consecutive months. A 2-month lead time is herein referred to as a hindcast of conditions for next month made from initial conditions at the beginning of the present month. A hindcast made at the beginning of March and covering the months of March–May can be evaluated for each month separately or in the aggregate. Climate anomalies may be more predictable on seasonal rather than monthly time scales (e.g., Luo et al. 2007) as temporal averaging can increase the signal-to-noise ratio (Fricker et al. 2013; Roundy et al. 2015). A comprehensive evaluation of hindcast skill over various temporal permutations is needed to cover the range of seasonal applications among decision-makers (e.g., Steinemann 2006). Wildland fire managers, for instance, begin to prioritize suppression resource allocations in the spring and are interested in seasonal forecasts issued in March–May (Corringham et al. 2008).

Hindcast verification is a complex and multidimensional problem lacking a universal approach (e.g., Willmott et al. 2012). To cover a wide range of potential applications, the quality (or relative accuracy) of seasonal hindcasts was evaluated through deterministic (correlation) and categorical [Heidke skill score (HSS)] means. Correlations were computed at each grid point separately for each time period (e.g., March, March–May) and lead time (e.g., 1 or 2 months). Since correlation represents a widely used summary measure of association between hindcasts and observations but yet fails to evaluate the magnitude of the errors (Barnston 1992), we use HSS to assess the categorical forecast skill for terciles (usually referred to as below, near, and above normal or three class) given their widespread use in operational seasonal outlooks (e.g., Peng et al. 2012). The HSS (expressed as a percentage) indicates the hindcast accuracy for each tercile relative to that expected by chance and is defined as

$$\mathrm{HSS} = \frac{(c - e) \times 100}{(t - e)},$$

where c is the number of cases (i.e., years) with correct forecasts (i.e., the hindcast falls within the same tercile as the observation), t is the total number of years (i.e., 29 in our period) in the outlook, and e is number of years expected to be correct by chance (i.e., $t/3$ for tercile-based categorization). The HSS ranges from -50 (for no hits) to 100 (all hits), where HSS $= 0$ is the expected value for a completely random hindcast. HSS was evaluated collectively for all categories, as well as separately for below-normal and above-normal conditions separately given the potentially greater value in utilization of seasonal hindcasts when the signal is particularly profound (e.g., Thomson et al. 2006). To assess the significance of HSS skill measures, we performed a bootstrapping test by generating 1000 samples of random hindcasts with replacement for $N = 29$ yr and defined the level of significance with the HSS value corresponding to the 95th percentile of the resulting 1000 HSS values. HSS and correlative skill values observed to fall below the 95% confidence level were masked out. We report the percentage of the western United States with statistically significant skill (HSS or correlation) at the 95% confidence level.

An additional analysis was conducted to assess the conditional hindcast skill based on the ENSO phase. Monthly Niño-3.4 SSTs from 1982 to 2010 were obtained from the NOAA/Climate Prediction Center (available online at http://www.cpc.ncep.noaa.gov/data/indices/). We evaluated the skill of hindcasts initialized in October contingent on ENSO phase given its strong teleconnections to the western United States climate in winter (e.g., Redmond and Koch 1991). We qualified ENSO years (either an El Niño or a La Niña event)

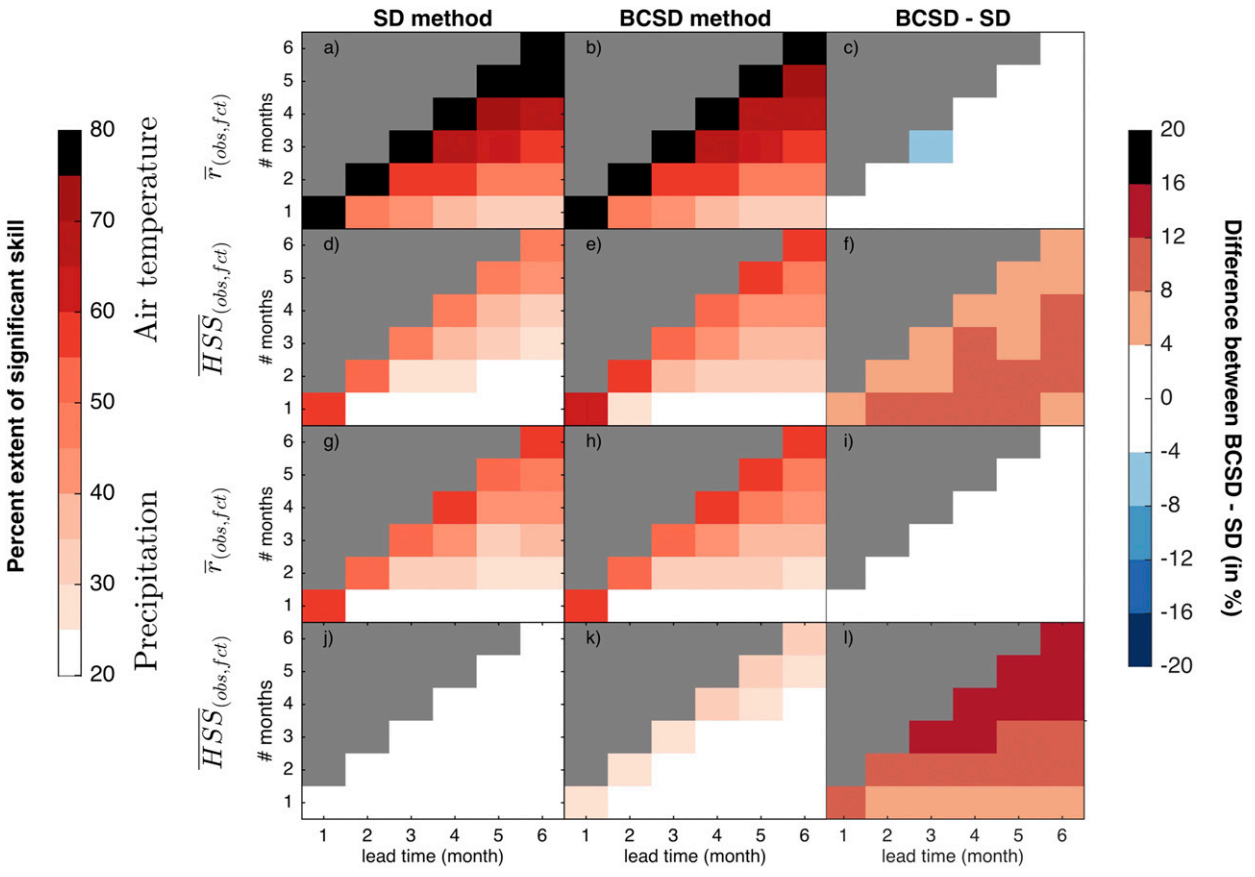

FIG. 1. Percent of grid points of the western United States with statistically significant skill for both (a)–(f) air temperature and (g)–(l) precipitation as a function of lead times (x axis) and the number of cumulative months over which hindcasts and observations are aggregated (y axis). The significance of HSS was assessed from a bootstrapping test by generating 1000 samples of random hindcasts with replacement ($N = 29$ yr). The HSS values corresponding to a confidence level of 95% for all categories were used to define the level of significance. Skill scores were computed from the $\overline{\text{MM}}$ ensemble for each month separately and then averaged across the year. The first column indicates scores from the SD method, the second column indicates scores from the BCSD method, while the third column gives the difference between BCSD and SD. Time periods for which no hindcasts were made above the diagonal are shown in gray.

when the mean August–October Niño-3.4 value was exceeded by one standard deviation. This resulted in a total of four El Niño winters (1982/83, 1987/88, 1997/98, and 2002/03) and four La Niña winters (1988/89, 1998/99, 1999/2000, and 2007/08). We compare the skill of the hindcasts initialized during years with a strong ENSO signal (El Niño or La Niña) versus a weak ENSO signal.

3. Results and discussion

a. Comparison of downscaling approaches

Figure 1 shows the percent of grid cells across the western United States with statistically significant correlations and HSS scores as a function of lead time and number of cumulative months for the $\overline{\text{MM}}$ hindcast. We illustrate these statistics averaged for all months of the year. Similar results were obtained using individual models from the NMME. Correlative skill measures were comparable between the two downscaling methods. Seasonal hindcasts of temperature showed significant correlative skill for more than 80% of the domain along the diagonal (e.g., 3-month hindcast aggregated for the next 3 months, 6-month hindcast aggregated for the next 6 months) (Figs. 1a,b). Hindcast skill decreased with lead time given the decaying influence of the initial conditions (e.g., Lavers et al. 2009); however, significant model skill was present for longer lead times that considered multiple months (e.g., 3-month hindcast ending with a lead time of 5 months). Seasonal precipitation hindcasts were less skillful than those for temperature (Figs. 1g,h). The maximum extent of the correlative skill (55%) was found for 1-month lead-time hindcasts, but approximately 30% of the domain had skill for 3-month hindcasts ending with a lead time of 5 months. Similar patterns were found for HSS; however, the BCSD approach showed significant skill

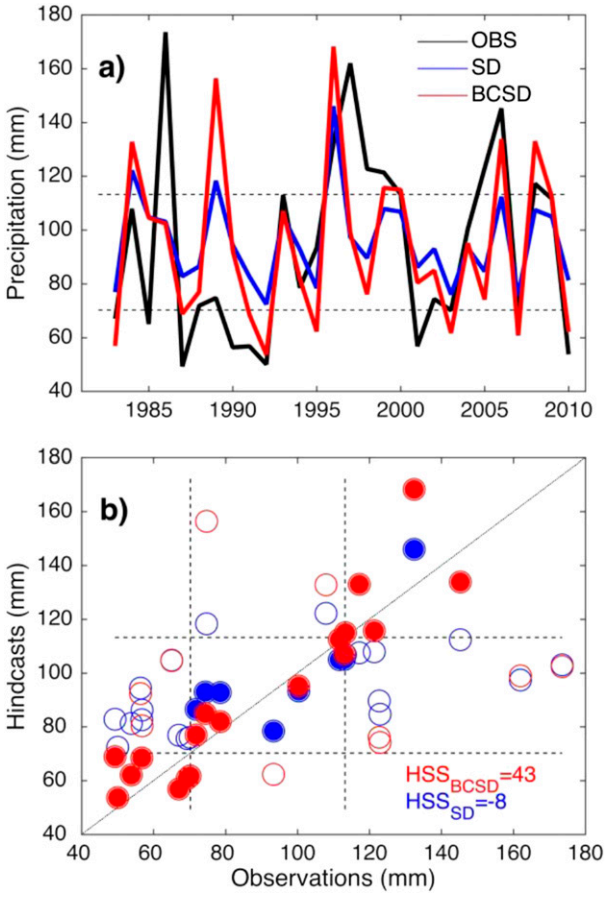

FIG. 2. Illustration of the impact of the variance reinflation after BCSD on the HSS for December–February precipitation hindcasts issued in early December in southeastern Idaho (42.23°N, 111.15°W). (a) Time series of observed precipitation (black) and the hindcasts from the $\overline{\text{MM}}$ after SD (blue) and BCSD (red). Dashed lines indicate tercile boundaries. (b) Scatterplot of observations against hindcasts. Red (blue) filled circles indicate correct hindcasts (i.e., hindcast falls within the observed tercile) from the $\overline{\text{MM}}$ after BCSD (SD). The overall HSS obtained for all categories from both BCSD and SD is also indicated.

over a much larger spatial extent of the domain than SD (Figs. 1f,l). As significant autocorrelations of monthly anomalies are often observed because of the persistence of the low-frequency mode of variability such as ENSO, the skill of the $\overline{\text{MM}}$ hindcast was also compared with the skill of the persistence hindcasts based on the observed climate from the month preceding the forecast. While the $\overline{\text{MM}}$ shows significant skill over more than 80% of the domain along the diagonal, persistence forecasts display significant skill over less than 30% of the domain, with a strong decline in the skill beyond 1-month lead time (not shown). This suggests that the $\overline{\text{MM}}$ is skillful relative to persistence forecasts.

The higher skill obtained using BCSD downscaling for categorical hindcasts is a consequence of quantile mapping the model output to the distribution of aggregated (1° resolution) observations. The use of ensemble means for individual models and the $\overline{\text{MM}}$ results in a deflation of variance. As the SD method strictly uses anomalies from these models, it will underestimate the variance and potentially have less utility for a place-based hindcast for tercile-based skill measures. A comparison of SD and BCSD hindcasts for a December–February forecast made in early December in southeastern Idaho is shown in Fig. 2. While both hindcasts had similar correlations with observations, SD tends to produce hindcasts close to the observed climatology (Fig. 2a) while BCSD, by correcting for the variance deflation, increases the proportion of categorical matches (Fig. 2b).

Tables 2 and 3 summarize the overall mean annual skill scores of hindcasts for each downscaling method and model. Although CFSv2 uses a larger number of runs relative to other models (which may lead to a stronger variance deflation), differences in hindcast skill between CFSv2 and other models were not evident. The BCSD method substantially outperforms the SD method for all HSS metrics, while comparable skill is seen between SD and BCSD for correlative metrics, although the skill of SD tends to slightly exceed that of BCSD for temperature. Additionally, the HSS skill is notably higher for upper- and lower-tercile hindcasts than for hindcasts including all categories, consistent with previous studies that have shown forecast skill for ''normal'' conditions to be limited (e.g., Van den Dool and Toth 1991). Hindcast skill varies across models, with CFSv2 and NASA generally outperforming the other models. However, the $\overline{\text{MM}}$ demonstrated skill exceeding any single model's skill, confirming the results of many other studies that are a consequence of combining complementary predictive skills in multimodel ensembles (e.g., Kirtman et al. 2014). Hereafter, we constrain our results to $\overline{\text{MM}}$ output downscaled using the BCSD approach for conciseness.

b. *Geographic and seasonal variability in hindcast skill*

Figure 3 shows the geographical distribution of the seasonal correlation and HSS scores between observed temperature and 3-month seasonal hindcasts initialized at the beginning of each season. While the spatial distribution of the correlation and HSS skill in winter is highest across the northwest United States, in agreement with previous findings (e.g., Arribas et al. 2011; Becker et al. 2014), our results indicate that the $\overline{\text{MM}}$ is most skillful in northwest Washington, the Idaho panhandle, and parts of eastern Oregon and western Idaho, while the skill tends to weaken across the lower

TABLE 2. Overall mean of temperature hindcast skill for the NMME models averaged over the western United States for different lead times (3/3 = lead time/cumulative month) and different metrics (ALL = all categories, WARM = above normal, COLD = below normal). The highest skill scores across both models and downscaling methods are set in boldface.

	SD/BCSD						
	CFSv2	CMC1	CMC2	GFDL	GFDL-FLOR	NASA	$\overline{\text{MM}}$
$\bar{r}_{3/3}$	0.45/0.46	0.46/0.44	0.44/0.42	0.47/0.45	0.27/0.27	0.48/0.47	**0.52**/0.48
$\bar{r}_{6/6}$	0.45/0.44	0.37/0.36	0.40/0.38	0.41/0.40	0.38/0.34	0.48/0.47	**0.49**/0.47
$\overline{\text{HSS} - \text{ALL}}_{3/3}$	11/**17**	12/14	12/12	12/**17**	2/7	12/**17**	12/**17**
$\overline{\text{HSS} - \text{ALL}}_{6/6}$	12/15	9/12	12/12	12/12	7/12	12/**17**	12/**17**
$\overline{\text{HSS} - \text{WARM}}_{3/3}$	23/**30**	26/28	27/26	26/28	13/18	28/**30**	26/**30**
$\overline{\text{HSS} - \text{WARM}}_{6/6}$	22/26	21/24	25/27	22/26	19/28	28/32	22/**36**
$\overline{\text{HSS} - \text{COLD}}_{3/3}$	29/33	24/29	29/32	27/28	20/18	29/33	31/**34**
$\overline{\text{HSS} - \text{COLD}}_{6/6}$	31/30	26/29	31/32	30/31	25/25	29/36	32/**35**

elevations of the Columbia Plateau in Washington State. Similar results were found for cold and warm terciles, albeit with noisier patterns. In spring, high-skill areas are found in the Northwest and for Arizona and New Mexico, while the lack of skill evident during the winter persists in California, Nevada, and Utah, in agreement with previous findings based on coarse-resolution NMME models (e.g., Becker et al. 2014). Higher skill in spring temperature is also found in the Snake River valley across southern Idaho relative to the surrounding mountainous areas. Widespread skill in summer temperature hindcasts is evident across the western United States, with the primary exception seen in coastal California and parts of the Sacramento valley (Shukla et al. 2015). This substantial decrease in the skill in the region is likely due to the strong influence of the maritime airflow from the ocean, while the topographic barrier south and north of the Sacramento delta prevents the flow of maritime air inland (Abatzoglou et al. 2009). The lowest forecast skill was found for fall, as seen in previous efforts (e.g., Roundy et al. 2015). However, some skill persists across New Mexico, Arizona, and Colorado. More details about the seasonal dependency of the coarse-resolution skill across the United States can be found in Peng et al. (2012).

Seasonal hindcasts of precipitation show regional skill (Fig. 4), although typically less than that for temperature. The greatest skill in the winter is found across southern Arizona and New Mexico, much of Idaho, the northern Great Basin, and much of California, with more pronounced spatial gradients in skill than in coarse-scale studies (e.g., Peng et al. 2012). There is also correlative skill in the Cascades of Oregon and central Washington State, despite a notable absence in adjacent low-lying areas, potentially tied either to fluctuations in orographic precipitation enhancement (e.g., Dettinger et al. 2004; Luce et al. 2013) or inadequacies in how GCMs simulate moisture transport pathways that affect the low-lying regions (Mo and Lettenmaier 2014). There is a clear asymmetry for dry and wet terciles, with large geographic regions of skill for wet terciles including parts of the Rocky Mountains and the Sierra Nevada during winter while no skill is found for dry conditions in these regions. High correlative skill for spring precipitation is found across much of the southwestern United States. Hindcast skill is notably absent across the Southwest in summer across a region generally affected by the North American monsoon (Higgins et al. 1998), whereas widespread skill is evident across the northern half of the western United States, as seen in Roundy et al. (2015). Summer precipitation hindcasts initialized in early June may therefore be useful for fire management planning across parts of the northern and middle

TABLE 3. As in Table 2, but for precipitation.

	SD/BCSD						
	CFSv2	CMC1	CMC2	GFDL	GFDL-FLOR	NASA	$\overline{\text{MM}}$
$\bar{r}_{3/3}$	0.30/0.30	0.29/0.26	0.34/0.32	0.29/0.29	0.19/0.21	0.35/0.33	**0.38**/0.37
$\bar{r}_{6/6}$	0.30/0.29	0.29/0.27	0.32/0.32	0.27/0.28	0.23/0.22	0.34/0.31	**0.37**/0.36
$\overline{\text{HSS} - \text{ALL}}_{3/3}$	1–7	7/12	9/**17**	7/12	2/7	9/**17**	7/**17**
$\overline{\text{HSS} - \text{ALL}}_{6/6}$	−1/7	2/12	12/**17**	7/12	2/12	7/12	7/**17**
$\overline{\text{HSS} - \text{WET}}_{3/3}$	16/19	14/16	18/21	16/19	10/13	20/23	17/**25**
$\overline{\text{HSS} - \text{WET}}_{6/6}$	14/22	13/18	19/21	17/21	13/18	20/22	14/**28**
$\overline{\text{HSS} - \text{DRY}}_{3/3}$	3/18	5/17	12/18	13/18	4/13	12/21	0/**23**
$\overline{\text{HSS} - \text{DRY}}_{6/6}$	0/18	5/18	14/24	12/20	4/17	10/19	0/**28**

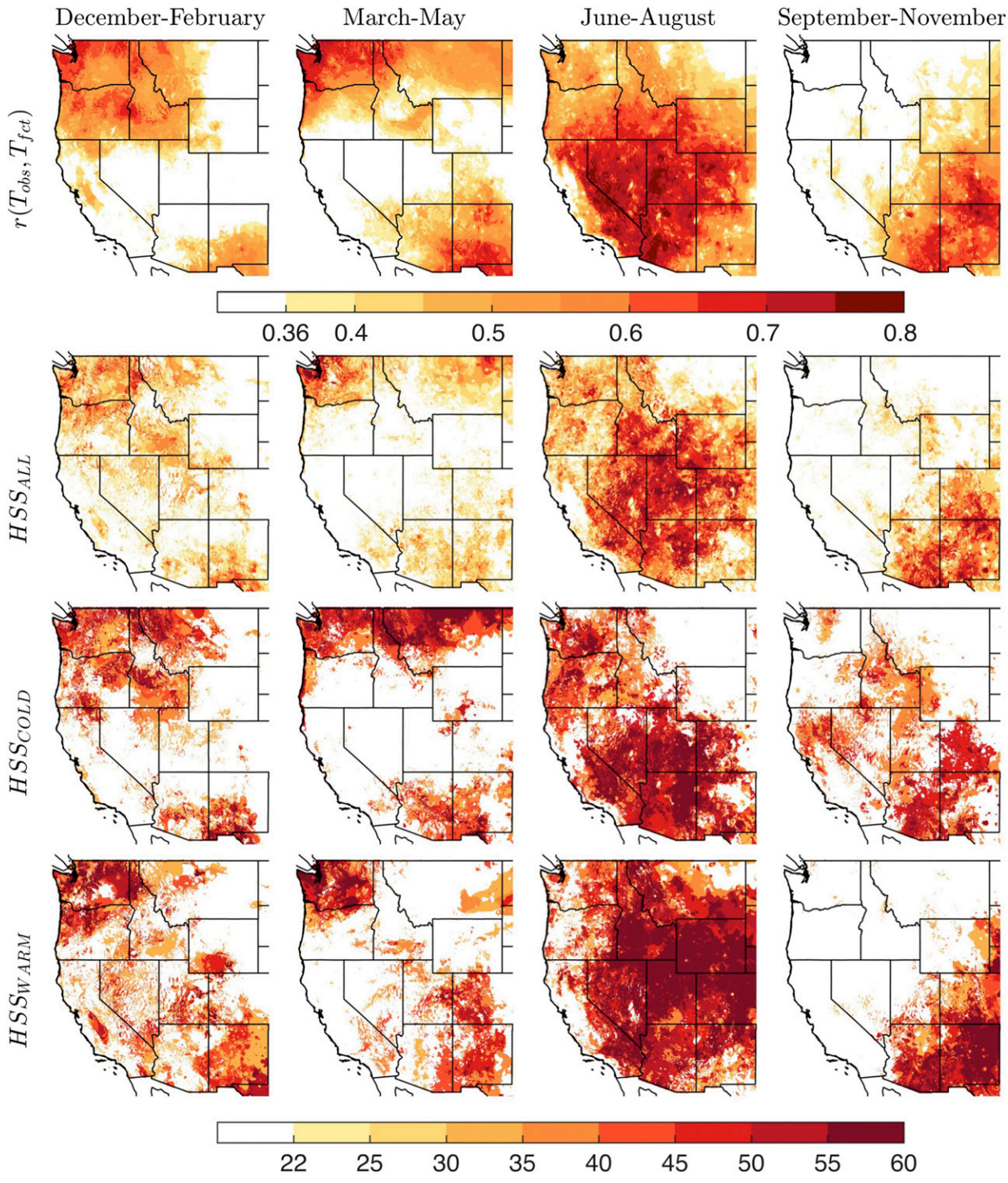

FIG. 3. Spatial distribution of seasonal skill scores for air temperature computed from the $\overline{\text{MM}}$ ensemble with 1-, 2-, and 3-month lead times (e.g., the first column shows the skill of the hindcasts made in early December for the December–February period) using the (first row) coefficient correlation, (second row) HSS for all categories (including below, normal, and above normal), and HSS for only (third row) cold and (fourth row) warm conditions. Nonsignificant values at the 95% confidence level were masked out.

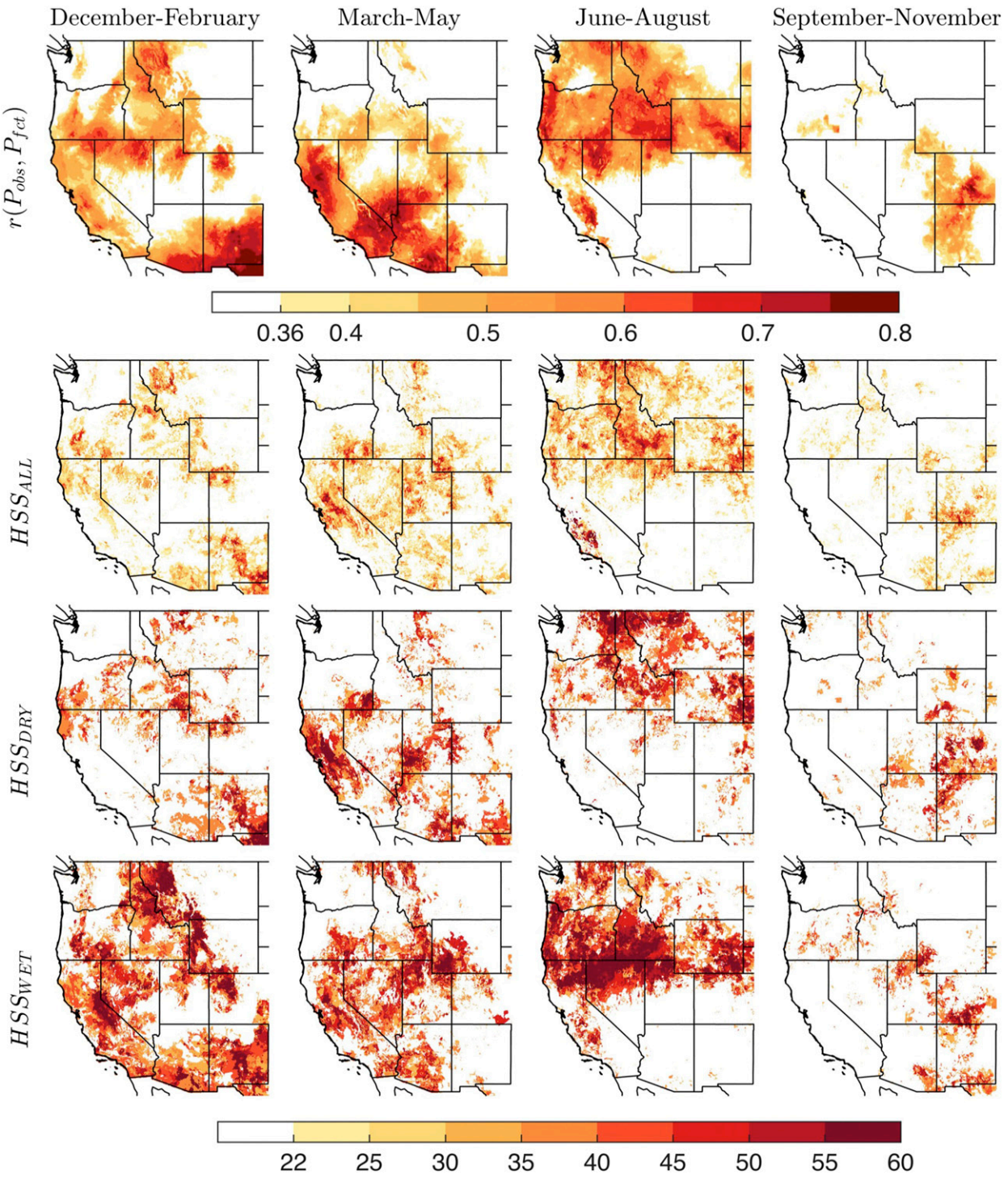

FIG. 4. As in Fig. 3, but for precipitation, where the HSS for only (third row) dry and (fourth row) wet conditions is shown.

Rocky Mountains and Cascades, where summer precipitation is well correlated with burned area (e.g., Littell et al. 2009; Abatzoglou and Kolden 2013). The ability of the $\overline{\text{MM}}$ to accurately predict wet conditions in this area may be related to the impact of the warm phase of ENSO during summer that brings an excess of moisture across the interior Northwest (Barbero et al. 2015). Hindcast skill during fall is notably weaker outside of the region across western Colorado and New Mexico. As the patterns for correlative scores and HSS

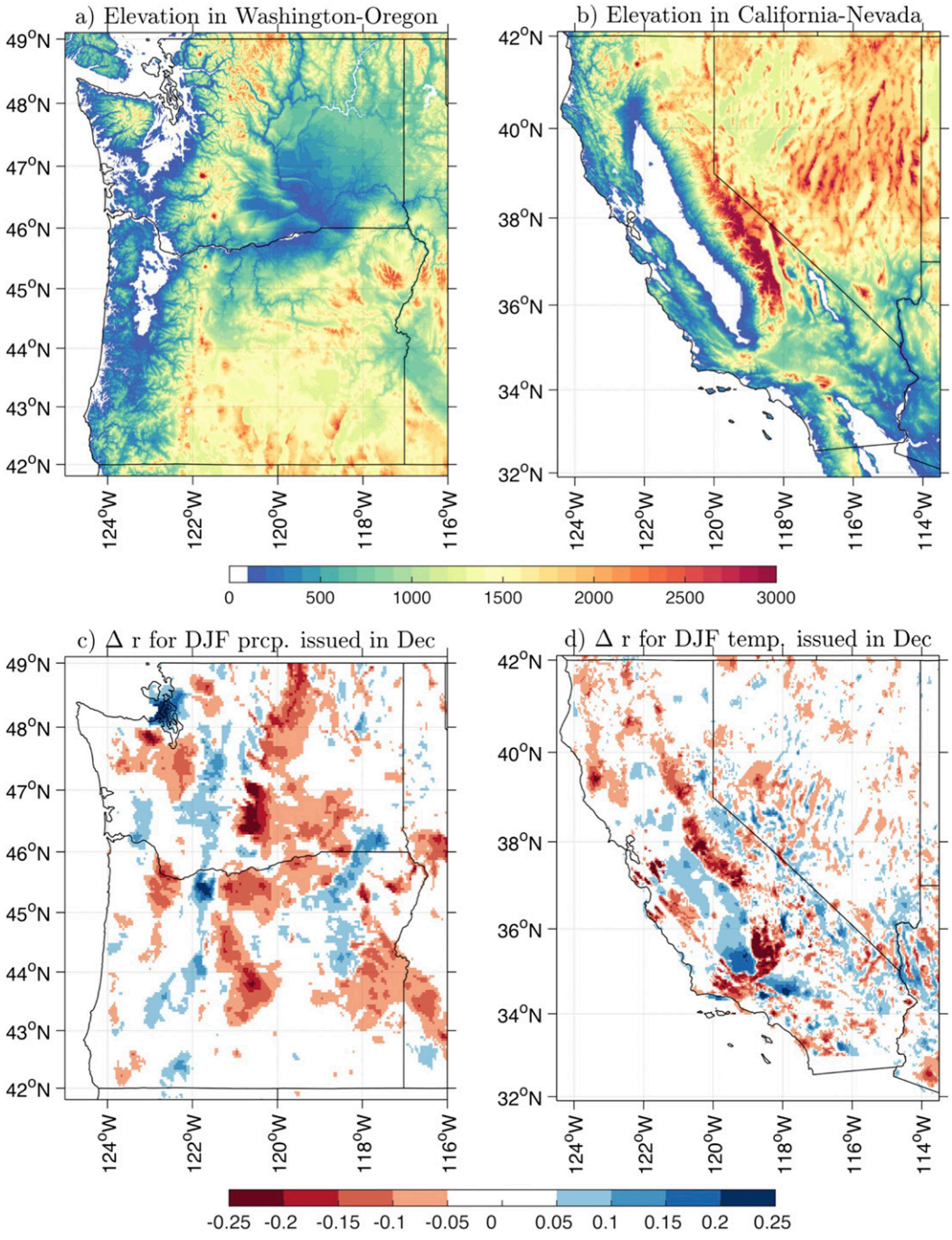

FIG. 5. Elevation (in m) in (a) WA–OR and (b) CA–NV. Difference in correlative scores from the $\overline{MM}$ ensemble between BCSD hindcasts at 4 km and BC hindcasts at the native resolution of NMME outputs (1°) for December–February (c) precipitation in WA–OR and (d) temperature in CA–NV hindcasts issued in early December. Blue (red) indicates regions where the downscaled hindcasts improved (degraded) the correlative skill relative to the coarse-resolution skill. Notice that differences in correlative scores between fine- and coarse-scale hindcasts are shown only where both datasets overlap in space, which is not the case, for instance, along the U.S.–Mexico border.

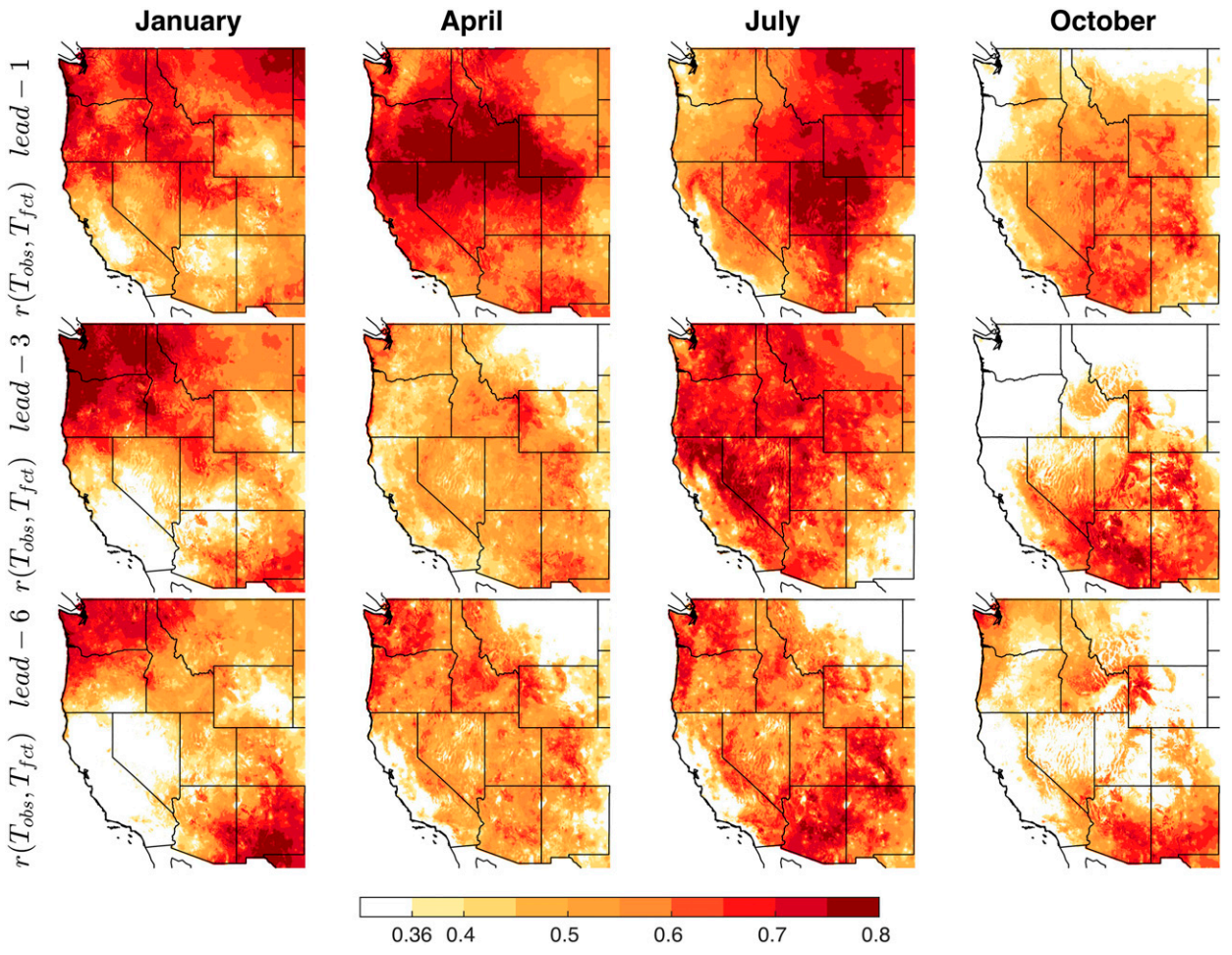

FIG. 6. Spatial distribution of seasonal correlative scores for air temperature computed from the $\overline{\text{MM}}$ ensemble for (top) the current month, (middle) the next 3 months, and (bottom) the next 6 months. Nonsignificant values at the 95% confidence level were masked out.

were in general agreement, we hereafter constrain our analysis to correlative scores.

The correlative scores of downscaled hindcasts showed distinct differences from those obtained using coarse resolution that highlight the role of complex topography and affect the usability of seasonal climate forecasts for local decision-making. The skill of downscaled estimates is generally lower than the coarse-resolution skill due to the additional variability in climate at finer spatial scales (e.g., Gangopadhyay et al. 2004). However, downscaled hindcasts showed higher correlative skill than coarse-resolution hindcasts along the windward slopes of the Cascade Range (Figs. 5a,c) and in the northern Rockies (see Fig. S3 in the online supplement to this paper). Conversely, reduced skill was present in the lee of topographic barriers, where seasonal precipitation totals occur in a reduced number of precipitation events, and was more likely to be associated with individual synoptic systems (e.g., Abatzoglou 2016) rather than large-scale climate patterns (e.g., Wise 2010). Figures 5b and 5d provide a comparison of downscaled versus coarse-resolution hindcast skill results for December–February temperature in California and Nevada for forecasts made in early December. Results show that downscaled hindcasts had higher skill than coarse-scale hindcasts in the Central Valley of California but showed less skill in the Sierra Nevada.

The spatial distribution of hindcast skill for air temperature for 1-month (current month), 3-month (next 3 months), and 6-month (next 6 months) time spans is shown in Fig. 6 for hindcasts initialized in January, April, July, and October. Although the spatial extent of the significant correlations globally generally decreases with lead time, increased correlative skill is seen for some regions. For example, a strengthening of the skill between 1- and 3-month hindcasts is seen in Arizona, the Wasatch Mountains in Utah, and across parts of the Rocky Mountains in Colorado for forecasts initialized in

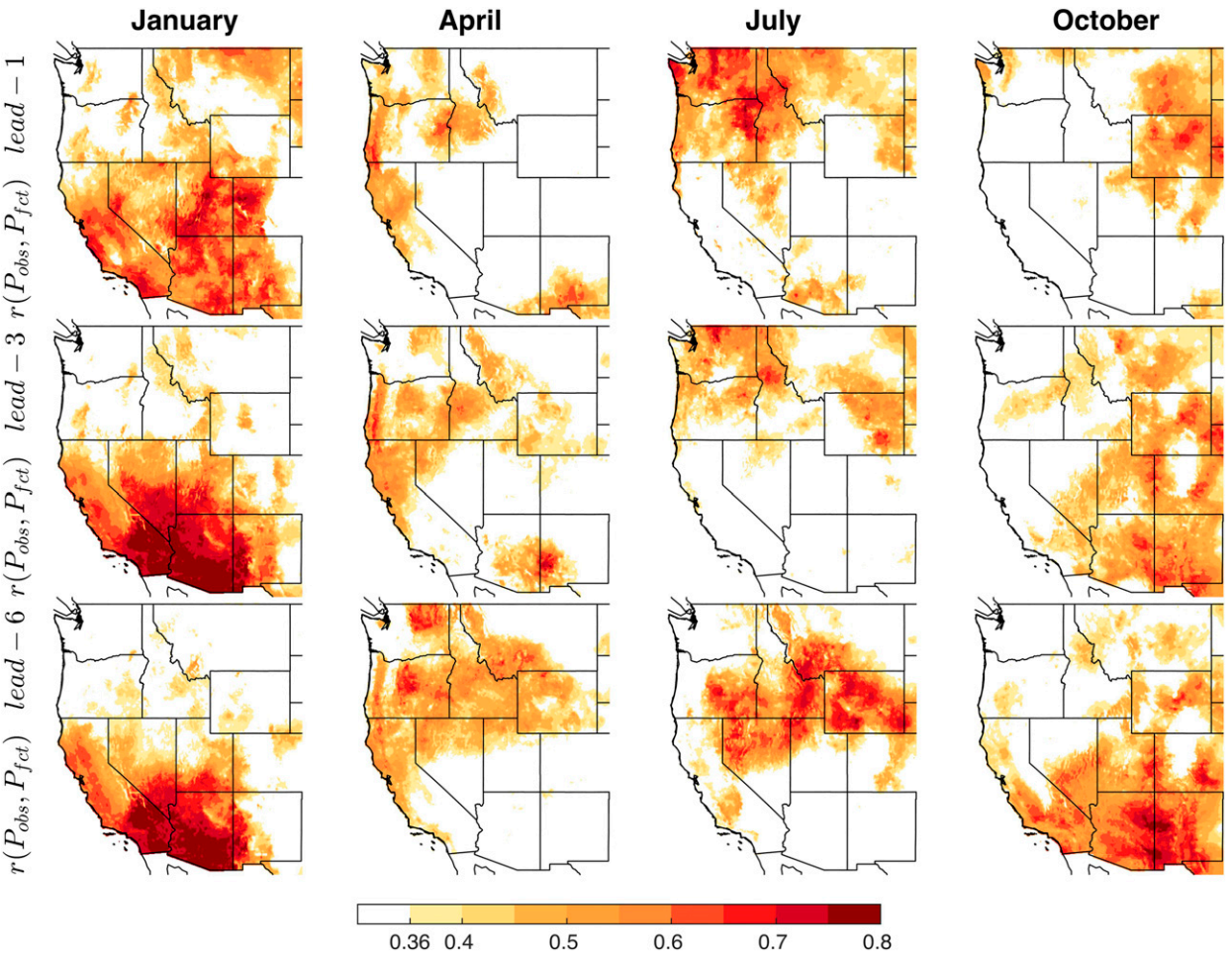

FIG. 7. As in Fig. 6, but for precipitation.

October. Likewise, correlation skill for 6-month hindcasts initialized in October is higher than 1-month hindcasts across high-elevation areas including the Sawtooth Range in central Idaho, and the Wind River Range in western Wyoming. However, a sharp decrease in correlations can be seen across the Snake River plain, the eastern slopes of the northern Rockies, and the valleys in the Great Basin relative to their surrounding locales during the wintertime. These regions are particularly prone to winter radiation inversions that decouple lower-elevation temperatures from regional air temperatures.

The hindcast skill for precipitation exhibited variability across seasons, lead times, and geography (Fig. 7). Overall, skillful precipitation hindcasts initialized in January precipitation for 1-month lead time were found across much of the southern half of the domain, and in Arizona and Southern California for 3- and 6-month time spans where coarse-resolution precipitation outlooks are generally skillful (Mo et al. 2012). Hindcasts initialized in October suggest low predictability across the western United States at short lead times whereas 6-month hindcasts that could support water management decisions showed skill over the Colorado River basin.

c. *ENSO-based conditional hindcast skill*

A comparison of hindcast skill for October–March temperatures for forecasts made in early October for ENSO versus ENSO-neutral years shows common regions of significant skill across the Desert Southwest and the northwestern United States (Figs. 8a,b). For the northwestern United States, an additional 20% of the temperature variance was explained for forecasts issued during active ENSO years (i.e., absolute value of August–October Niño-3.4 exceeded one standard deviation) across portions of the lower elevations versus during ENSO-neutral years.

Six-month precipitation hindcasts initialized in October for ENSO-neutral years showed high correlations ($r > 0.8$) near the Four Corners region and moderate

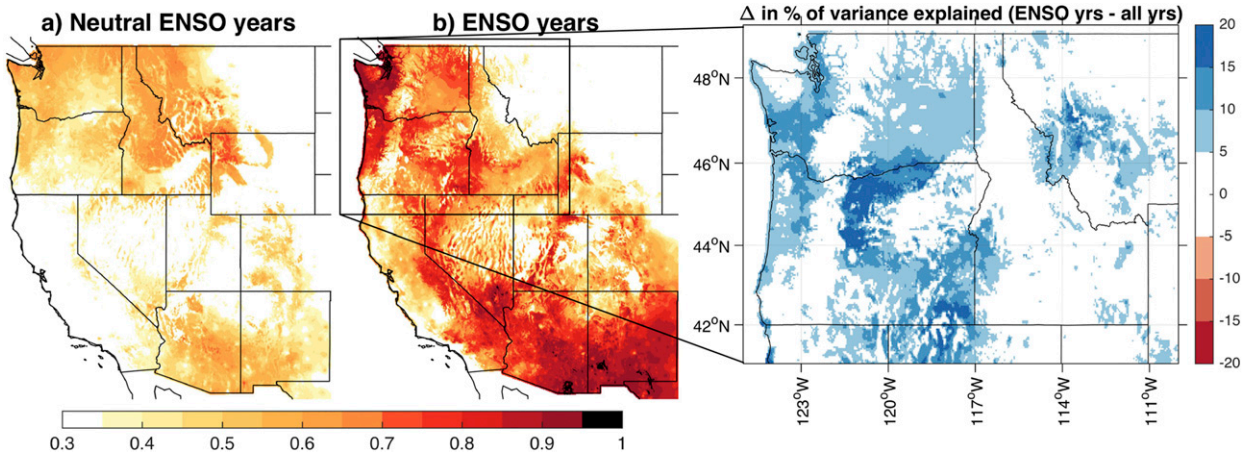

FIG. 8. Spatial distribution of correlations between mean air temperature conditions from October to March and the $\overline{\text{MM}}$ ensemble hindcasts made in October during (a) non-ENSO years and (b) ENSO years. We consider ENSO years for the period 1982–2010 when the August–October Niño-3.4 exceeded one standard deviation of its August–October values. Difference in the percentage of variance explained between ENSO years and all years is indicated on the right for the Pacific Northwest. Note that correlations in (a) and (b) were computed from different sample sizes. The critical values at the 95% confidence level are 0.44 and 0.71 in (a) and (b), respectively.

correlations ($r > 0.4$) across parts of interior Oregon and the southern two-thirds of Idaho (Fig. 9a). Stark differences in regions of strong skill were found for ENSO years (Fig. 9b) compared with ENSO-neutral years, including across much of California, most notably in the southwest portion of the state in agreement with previous efforts (Gershunov 1998; Peng et al. 2012; Roundy et al. 2015). For the northwestern United States, an additional 20% of the precipitation variance was explained for forecasts issued during active ENSO years versus during ENSO-neutral years along the windward side of the Cascade Range and northern Rockies, likely a consequence of orographic precipitation controls that arise in response to latitudinal variations in the storm track (e.g., Mass et al. 2015). Likewise, this may arise as a result of the limited predictability of cool precipitation on the lee side of the Cascade Range. We caution that the assessed skill for ENSO years is based on a small sample size ($N = 8$ yr) that may introduce uncertainty (Kumar 2009); further analyses based on longer records are needed to confirm these findings.

4. Conclusions

While previous efforts investigated the skill of seasonal hindcasts over the United States using coarse-scale GCMs, this analysis provides a useful guide to determining the skill of statistically downscaled seasonal temperature and precipitation forecasts over the western United States. We demonstrated the advantages of implementing quantile mapping in BCSD versus SD in categorical skill measures as a means to improving the

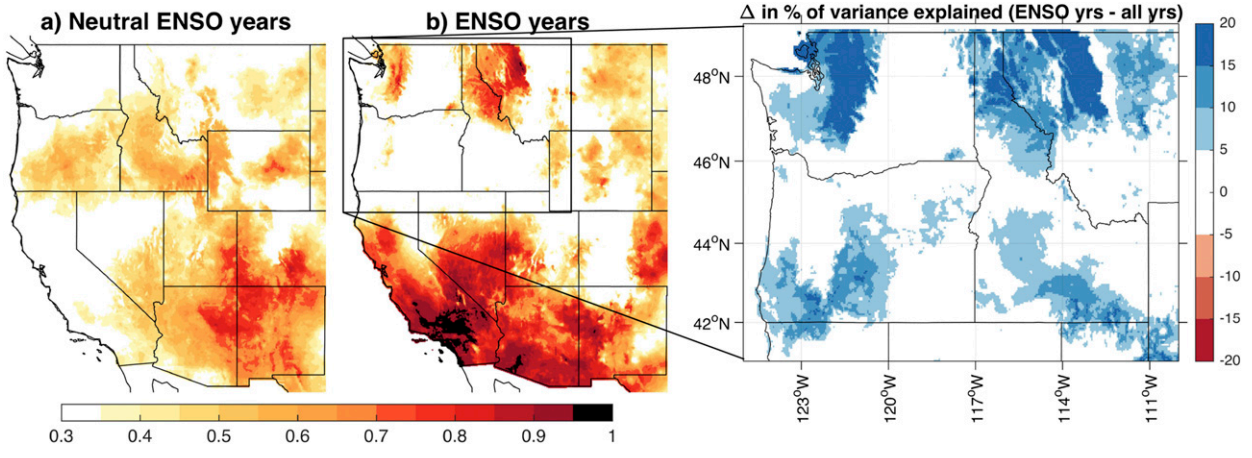

FIG. 9. As in Fig. 8, but for precipitation.

variance deflation of ensemble means. While these results are not unexpected given the nature of ensemble model output such as from NMME, they can be informative for developing place-based seasonal climate forecasts. While this work only considered simple statistical downscaling approaches, more sophisticated methods might be able to better capture climate features observed in complex terrain such as inversions and orographic precipitation (Ning et al. 2012; Abatzoglou and Brown 2012; Bürger et al. 2012, 2013).

Our results demonstrate differences in skill of downscaled seasonal forecasts versus at their native resolution. These differences were most pronounced in regions of complex topography, where inversions and orographic processes can result in substantial differences in monthly and seasonal climate variability over short distances. Statistically downscaled seasonal climate forecasts result in place-based predictions that are not only more usable but that also better elucidate finescale information about model credibility as compared with forecasts issued at the native resolution of the model. Finally, our results suggest that forecast skill is contingent upon whether there is a significant ENSO signal or not. This can be of value for water supply management agencies that begin to project seasonal water supplies for winter and spring in October (Hartmann et al. 2002).

Understanding the skill of these seasonal climate forecasts across the multiple dimensions at actionable scales can provide valuable information for decision-making in water and natural resources, agriculture, and infrastructure planning. The economic benefits of seasonal forecasts may be substantial (Steinemann 2006), but the utility of seasonal outlooks in the decision-making process may be limited by the coarse-scale nature of the available information. Downscaled seasonal climate forecasts may help improve the utility of seasonal hindcasts for consumers of such information.

Acknowledgments. The authors appreciate the constructive reviews by three anonymous reviewers who helped improve the quality of this manuscript. We thank the climate modeling groups for producing and making available their model outputs. This research was supported by USDA-NIFA Awards 2011-68002-30191 and AFRI-60-5360-2-832, as well as NOAA RISA under Award NA15OAR4310145.

REFERENCES

Abatzoglou, J. T., 2013: Development of gridded surface meteorological data for ecological applications and modeling. *Int. J. Climatol.*, **33**, 121–131, doi:10.1002/joc.3413.

——, 2016: Contribution of cutoff lows to precipitation across the United States. *J. Appl. Meteor. Climatol.*, **55**, 893–899, doi:10.1175/JAMC-D-15-0255.1.

——, and T. J. Brown, 2012: A comparison of statistical downscaling methods suited for wildfire applications. *Int. J. Climatol.*, **32**, 772–780, doi:10.1002/joc.2312.

——, and C. A. Kolden, 2013: Relationships between climate and macroscale area burned in the western United States. *Int. J. Wildland Fire*, **22**, 1003–1020, doi:10.1071/WF13019.

——, K. T. Redmond, and L. M. Edwards, 2009: Classification of regional climate variability in the state of California. *J. Appl. Meteor. Climatol.*, **48**, 1527–1541, doi:10.1175/2009JAMC2062.1.

Ahmed, K. F., G. Wang, J. Silander, A. M. Wilson, J. M. Allen, R. Horton, and R. Anyah, 2013: Statistical downscaling and bias correction of climate model outputs for climate change impact assessment in the U.S. northeast. *Global Planet. Change*, **100**, 320–332, doi:10.1016/j.gloplacha.2012.11.003.

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

Barbero, R., J. T. Abatzoglou, and T. J. Brown, 2015: Seasonal reversal of the influence of El Niño–Southern Oscillation on very large wildfire occurrence in the interior western United States. *Geophys. Res. Lett.*, **42**, 3538–3545, doi:10.1002/2015GL063428.

Barnston, A. G., 1992: Correspondence among the correlation, RMSE, and Heidke forecast verification measures; refinement of the Heidke score. *Wea. Forecasting*, **7**, 699–709, doi:10.1175/1520-0434(1992)007<0699:CATCRA>2.0.CO;2.

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

Bürger, G., T. Q. Murdock, A. T. Werner, S. R. Sobie, and A. J. Cannon, 2012: Downscaling extremes: An intercomparison of multiple statistical methods for present climate. *J. Climate*, **25**, 4366–4388, doi:10.1175/JCLI-D-11-00408.1.

——, S. R. Sobie, A. J. Cannon, A. T. Werner, and T. Q. Murdock, 2013: Downscaling extremes: An intercomparison of multiple methods for future climate. *J. Climate*, **26**, 3429–3449, doi:10.1175/JCLI-D-12-00249.1.

Corringham, T., A. L. Westerling, and B. Morehouse, 2008: Exploring use of climate information in wildland fire management: A decision calendar study. *J. For.*, **106** (2), 71–77.

Dettinger, M., K. Redmond, and D. Cayan, 2004: Winter orographic precipitation ratios in the Sierra Nevada—Large-scale atmospheric circulations and hydrologic consequences. *J. Hydrometeor.*, **5**, 1102–1116, doi:10.1175/JHM-390.1.

Doblas-Reyes, F. J., J. García-Serrano, F. Lienert, A. P. Biescas, and L. R. L. Rodrigues, 2013: Seasonal climate predictability and forecasting: Status and prospects. *Wiley Interdiscip. Rev.: Climate Change*, **4**, 245–268, doi:10.1002/wcc.217.

Fowler, H. J., S. Blenkinsop, and C. Tebaldi, 2007: Linking climate change modelling to impacts studies: Recent advances in downscaling techniques for hydrological modelling. *Int. J. Climatol.*, **27**, 1547–1578, doi:10.1002/joc.1556.

Fraisse, C. W., N. E. Breuer, D. Zierden, J. G. Bellow, J. Paz, V. E. Cabrera, and J. J. O'Brien, 2006: AgClimate: A climate forecast information system for agricultural risk management in the southeastern USA. *Comput. Electron. Agric.*, **53**, 13–27, doi:10.1016/j.compag.2006.03.002.

Frías, M. D., S. Herrera, A. S. Cofiño, and J. M. Gutiérrez, 2010: Assessing the skill of precipitation and temperature seasonal forecasts in Spain: Windows of opportunity related to ENSO events. *J. Climate*, **23**, 209–220, doi:10.1175/2009JCLI2824.1.

Fricker, T. E., C. A. T. Ferro, and D. B. Stephenson, 2013: Three recommendations for evaluating climate predictions. *Meteor. Appl.*, **20**, 246–255, doi:10.1002/met.1409.

Gangopadhyay, S., M. Clark, K. Werner, D. Brandon, and B. Rajagopalan, 2004: Effects of spatial and temporal aggregation on the accuracy of statistically downscaled precipitation estimates in the upper Colorado River basin. *J. Hydrometeor.*, **5**, 1192–1206, doi:10.1175/JHM-391.1.

Gershunov, A., 1998: ENSO influence on intraseasonal extreme rainfall and temperature frequencies in the contiguous United States: Implications for long-range predictability. *J. Climate*, **11**, 3192–3203, doi:10.1175/1520-0442(1998)011<3192:EIOIER>2.0.CO;2.

Goddard, L., and S. J. Mason, 2002: Sensitivity of seasonal climate forecasts to persisted SST anomalies. *Climate Dyn.*, **19**, 619–632, doi:10.1007/s00382-002-0251-y.

Graham, R. J., A. D. L. Evans, K. R. Mylne, M. S. J. Harrison, and K. B. Robertson, 2000: An assessment of seasonal predictability using atmospheric general circulation models. *Quart. J. Roy. Meteor. Soc.*, **126**, 2211–2240, doi:10.1256/smsqj.56711.

Hartmann, H., T. C. Pagano, S. Sorooshian, and R. Bales, 2002: Confidence builders: Evaluating seasonal climate forecasts from user perspectives. *Bull. Amer. Meteor. Soc.*, **83**, 683–698, doi:10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Higgins, R. W., K. C. Mo, and Y. Yao, 1998: Interannual variability of the U.S. summer precipitation regime with emphasis on the southwestern monsoon. *J. Climate*, **11**, 2582–2606, doi:10.1175/1520-0442(1998)011<2582:IVOTUS>2.0.CO;2.

Jia, L., and Coauthors, 2015: Improved seasonal prediction of temperature and precipitation over land in a high-resolution GFDL climate model. *J. Climate*, **28**, 2044–2062, doi:10.1175/JCLI-D-14-00112.1.

Kim, H. M., P. Webster, and J. Curry, 2012: Seasonal prediction skill of ECMWF System 4 and NCEP CFSv2 retrospective forecast for the Northern Hemisphere winter. *Climate Dyn.*, **39**, 2957–2973, doi:10.1007/s00382-012-1364-6.

Kirtman, B. P., and Coauthors, 2014: The North American Multimodel Ensemble: Phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. *Bull. Amer. Meteor. Soc.*, **95**, 585–601, doi:10.1175/BAMS-D-12-00050.1.

Koster, R. D., and Coauthors, 2010: Contribution of land surface initialization to subseasonal forecast skill: First results from a multi-model experiment. *Geophys. Res. Lett.*, **37**, L02402, doi:10.1029/2009GL041677.

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

Lavers, D., L. Luo, and E. F. Wood, 2009: A multiple model assessment of seasonal climate forecast skill for applications. *J. Geophys. Res.*, **36**, L23711, doi:10.1029/2009GL041365.

Littell, J. S., D. McKenzie, D. L. Peterson, and A. L. Westerling, 2009: Climate and wildfire area burned in western U.S. ecoprovinces, 1916–2003. *Ecol. Appl.*, **19**, 1003–1021, doi:10.1890/07-1183.1.

Luce, C. H., J. T. Abatzoglou, and Z. A. Holden, 2013: The missing mountain water: Slower westerlies decrease orographic enhancement in the Pacific Northwest USA. *Science*, **342**, 1360–1364, doi:10.1126/science.1242335.

Luo, L. F., E. F. Wood, and M. Pan, 2007: Bayesian merging of multiple climate model forecasts for seasonal hydrological predictions. *J. Geophys. Res.*, **112**, D10102, doi:10.1029/2006JD007655.

Ma, F., and Coauthors, 2016: Evaluating the skill of NMME seasonal precipitation ensemble predictions for 17 hydroclimatic regions in continental China. *Int. J. Climatol.*, **36**, 132–144, doi:10.1002/joc.4333.

Manzanas, R., M. D. Frías, A. S. Cofiño, and J. M. Gutiérrez, 2014: Validation of 40 year multimodel seasonal precipitation forecasts: The role of ENSO on the global skill. *J. Geophys. Res. Atmos.*, **119**, 1708–1719, doi:10.1002/2013JD020680.

Mass, C., N. Johnson, M. Warner, and R. Vargas, 2015: Synoptic control of cross-barrier precipitation ratios for the Cascade Mountains. *J. Hydrometeor.*, **16**, 1014–1028, doi:10.1175/JHM-D-14-0149.1.

Merryfield, W. J., and Coauthors, 2013: The Canadian Seasonal to Interannual Prediction System. Part I: Models and initialization. *Mon. Wea. Rev.*, **141**, 2910–2945, doi:10.1175/MWR-D-12-00216.1.

Mo, K. C., and D. P. Lettenmaier, 2014: Hydrologic prediction over the conterminous United States using the National Multi-Model Ensemble. *J. Hydrometeor.*, **15**, 1457–1472, doi:10.1175/JHM-D-13-0197.1.

——, S. Shukla, D. P. Lettenmaier, and L. C. Chen, 2012: Do Climate Forecast System (CFSv2) forecasts improve seasonal soil moisture prediction? *Geophys. Res. Lett.*, **39**, L23703, doi:10.1029/2012GL053598.

Molod, A., L. Takacs, M. Suarez, J. Bacmeister, I. S. Song, and A. Eichmann, 2012: The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to Fortuna. NASA Tech. Memo. NASA/TM–2012-104606, NASA Tech. Rep. Series on Global Modeling and Data Assimilation, Vol. 28, 175 pp. [Available online at https://gmao.gsfc.nasa.gov/pubs/docs/tm28.pdf.]

Ning, L., M. E. Mann, R. Crane, and T. Wagener, 2012: Probabilistic projections of climate change for the mid-Atlantic region of the United States: Validation of precipitation downscaling during the historical era. *J. Climate*, **25**, 509–526, doi:10.1175/2011JCLI4091.1.

——, E. E. Riddle, and R. S. Bradley, 2015: Projected changes in climate extremes over the northeastern United States. *J. Climate*, **28**, 3289–3310, doi:10.1175/JCLI-D-14-00150.1.

Peng, P., A. Kumar, M. S. Halpert, and A. G. Barnston, 2012: An analysis of CPC's operational 0.5-month lead seasonal outlooks. *Wea. Forecasting*, **27**, 898–917, doi:10.1175/WAF-D-11-00143.1.

Rayner, S., D. Lach, and H. Ingram, 2005: Weather forecasts are for wimps: Why water resource managers do not use climate forecasts. *Climatic Change*, **69**, 197–227, doi:10.1007/s10584-005-3148-z.

Redmond, K. T., and R. W. Koch, 1991: Surface climate and streamflow variability in the western United States and their relationship to large-scale circulation indices. *Water Resour. Res.*, **27**, 2381–2399, doi:10.1029/91WR00690.

Roundy, J. K., X. Yuan, J. Schaake, and E. F. Wood, 2015: A framework for diagnosing seasonal prediction through canonical event analysis. *Mon. Wea. Rev.*, **143**, 2404–2418, doi:10.1175/MWR-D-14-00190.1.

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

Scaife, A. A., and Coauthors, 2014: Skillful long range prediction of European and North American winters. *Geophys. Res. Lett.*, **41**, 2514–2519, doi:10.1002/2014GL059637.

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

Slingo, J., and T. Palmer, 2011: Uncertainty in weather and climate prediction. *Philos. Trans. Roy. Soc.*, **369A**, 4751–4767, doi:10.1098/rsta.2011.0161.

Steinemann, A. C., 2006: Using climate forecasts for drought management. *J. Appl. Meteor. Climatol.*, **45**, 1353–1361, doi:10.1175/JAM2401.1.

Thomson, M. C., F. J. Doblas-Reyes, S. J. Mason, R. Hagedorn, S. J. Connor, T. Phindela, A. P. Morse, and T. N. Palmer, 2006: Malaria early warnings based on seasonal climate forecasts from multi-model ensembles. *Nature*, **439**, 576–579, doi:10.1038/nature04503.

Tian, D., C. J. Martinez, W. D. Graham, and S. Hwang, 2014: Statistical downscaling multimodel forecasts for seasonal precipitation and surface temperature over the southeastern United States. *J. Climate*, **27**, 8384–8411, doi:10.1175/JCLI-D-13-00481.1.

Troccoli, A., 2010: Seasonal climate forecasting. *Meteor. Appl.*, **17**, 251–268.

Van den Dool, H. M., and Z. Toth, 1991: Why do forecasts for "near normal" often fail? *Wea. Forecasting*, **6**, 76–85, doi:10.1175/1520-0434(1991)006<0076:WDFFNO>2.0.CO;2.

Willmott, C. J., S. M. Robeson, and K. Matsuura, 2012: A refined index of model performance. *Int. J. Climatol.*, **32**, 2088–2094, doi:10.1002/joc.2419.

Wise, E. K., 2010: Spatiotemporal variability of the precipitation dipole transition zone in the western United States. *Geophys. Res. Lett.*, **37**, L07706, doi:10.1029/2009GL042193.

Wood, A. W., and D. P. Lettenmaier, 2006: A test bed for new seasonal hydrologic forecasting approaches in the western United States. *Bull. Amer. Meteor. Soc.*, **87**, 1699–1712, doi:10.1175/BAMS-87-12-1699.

——, E. P. Maurer, A. Kumar, and D. P. Lettenmaier, 2002: Long-range experimental hydrologic forecasting for the eastern United States. *J. Geophys. Res.*, **107**, 4429, doi:10.1029/2001JD000659.

Yoon, J. H., L. R. Leung, and J. Correia Jr., 2012: Comparison of dynamically and statistically downscaled seasonal climate forecasts for the cold season over the United States. *J. Geophys. Res.*, **117**, D21109, doi:10.1029/2012JD017650.

Zhang, S., M. J. Harrison, A. Rosati, and A. T. Wittenberg, 2007: System design and evaluation of coupled ensemble data assimilation for global oceanic climate studies. *Mon. Wea. Rev.*, **135**, 3541–3564, doi:10.1175/MWR3466.1.