Global Evaluation of Seasonal Precipitation and Temperature Forecasts from NMME[✉]

TIRTHANKAR ROY,^a XIAOGANG HE,^b PEIRONG LIN, AND HYLKE E. BECK

Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

CHRISTOPHER CASTRO

Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, Arizona

ERIC F. WOOD

Civil and Environmental Engineering, Princeton University, Princeton, New Jersey

(Manuscript received 25 April 2019, in final form 16 June 2020)

ABSTRACT

We present a comprehensive global evaluation of monthly precipitation and temperature forecasts from 16 seasonal forecasting models within the NMME Phase-1 system, using Multi-Source Weighted-Ensemble Precipitation version 2 (MSWEP-V2; precipitation) and Climate Research Unit TS4.01 (CRU-TS4.01; temperature) data as reference. We first assessed the forecast skill for lead times of 1–8 months using Kling–Gupta efficiency (KGE), an objective performance metric combining correlation, bias, and variability. Next, we carried out an empirical orthogonal function (EOF) analysis to compare the spatiotemporal variability structures of the forecasts. We found that, in most cases, precipitation skill was highest during the first lead time (i.e., forecast in the month of initialization) and rapidly dropped thereafter, while temperature skill was much higher overall and better retained at higher lead times, which is indicative of stronger temporal persistence. Based on a comprehensive assessment over 21 regions and four seasons, we found that the skill showed strong regional and seasonal dependencies. Some tropical regions, such as the Amazon and Southeast Asia, showed high skill even at longer lead times for both precipitation and temperature. Rainy seasons were generally associated with high precipitation skill, while during winter, temperature skill was low. Overall, precipitation forecast skill was highest for the NASA, NCEP, CMC, and GFDL models, and for temperature, the NASA, CFSv2, COLA, and CMC models performed the best. The spatiotemporal variability structures were better captured for precipitation than temperature. The simple forecast averaging did not produce noticeably better results, emphasizing the need for more advanced weight-based averaging schemes.

KEYWORDS: Ensembles; Forecast verification/skill; Forecasting; Seasonal forecasting

1. Introduction

The North American Multi-Model Ensemble (NMME; Kirtman et al. 2014) system incorporates seasonal forecasts of different hydroclimatic variables from multiple U.S. and

Canadian models. These forecasts are invaluable for a plethora of scientific and operational applications, including precipitation and temperature forecasting (Setiawan et al. 2017; Wang 2014; Krakauer 2017; Cash et al. 2019; Wanders et al. 2017), prediction of extremes (Yuan et al. 2015), atmospheric rivers (Zhou and Kim 2018), drought (Hao et al. 2017; Yao and Yuan 2018), hurricane activity (Harnos et al. 2017), and tropical cyclones (Finan et al. 2016).

As in any other multimodel framework, models within the NMME system show mixed performance in terms of seasonal and subseasonal predictions. Wang (2014) evaluated seven models from the NMME system for the CONUS and found that the uncertainty of precipitation forecasts is higher in the western United States during summer, which the authors attributed to the overall low

[✉] Supplemental information related to this paper is available at the Journals Online website: <https://doi.org/10.1175/JHM-D-19-0095.s1>.

^a Current affiliation: Civil and Environmental Engineering, University of Nebraska–Lincoln, Lincoln, Nebraska.

^b Current affiliation: Water in the West, Woods Institute for the Environment, Stanford University, Stanford, California.

Corresponding author: Tirthankar Roy, roy@unl.edu

TABLE 1. The 16 NMME models considered in this study.

Model	Hindcast period	Ensemble size	Lead time (months)	Provider
CMC1-CanCM3	1982–2010	10	12	Environment Canada
CMC2-CanCM4	1982–2010	10	12	Environment Canada
COLA-RSMAS-CCSM3	1982–2010	6	12	George Mason University
COLA-RSMAS-CCSM4	1982–2010	10	12	George Mason University
GFDL-CM2p1	1982–2010	10	12	NAA
GFDL-CM2p1-aer04	1982–2010	10	12	NOAA
GFDL-CM2p5-FLOR-A06	1982–2010	12	12	NOAA
GFDL-CM2p5-FLOR-B01	1982–2010	12	12	NOAA
IRI-ECHAM4p5-AnomalyCoupled	1982–2010	12	8	Columbia University
IRI-ECHAM4p5-DirectCoupled	1982–2010	12	8	Columbia University
NASA-GEOSS2S	1982–2010	4	9	NASA
NASA-GMAO	1982–2010	10	9	NASA
NASA-GMAO-062012	1982–2010	12	9	NASA
NCAR-CESM1	1982–2010	10	12	NCAR
NCEP-CFSv1	1982–2009	15	9	NCEP
NCEP-CFSv2	1982–2010	24	10	NCEP

precipitation in the region. The performance of the forecasts increased for below normal (33rd percentile of the observations) and above normal (67th percentile of the observations) conditions. [Thober et al. \(2015\)](#) studied NMME forecasts within the context of drought prediction over Europe and showed that the soil moisture skill of the forecasts can show variability up to 40% in space and time. [Ma et al. \(2016\)](#) compared precipitation forecasts from 11 models within the NMME system over 17 hydroclimatic regions in continental China and found overall higher skill in autumn and spring and lower skill in summer. In their case, the CFSv2 model (see details in [Table 1](#)) performed the best, whereas the IRI and CCSM3 models performed poorly. They attributed the performance differences among the models to local climatology and climate variability, ocean–land–atmosphere interactions, and predictability of atmospheric circulations. In their case, model averaging, especially the Bayesian model averaging, significantly improved the forecast skill. [Slater et al. \(2016\)](#) compared the precipitation and temperature forecasts from eight NMME models against their respective climatology over seven major regions in the United States and found that the models perform better in predicting droughts than floods on a seasonal scale. The drought prediction was comparable in terms of both high temperature and low precipitation. In addition, they found that the skill drops rapidly after the shortest lead time. They found that the unconditional biases (i.e., standardized mean error) in the models, which also show strong season and lead dependencies, dictate the forecasting skill. [Shukla et al. \(2016\)](#) studied the precipitation and temperature forecasts of eight NMME models over East Africa and found that temperature forecasts are more skillful than precipitation forecasts,

with the latter exhibiting some skill only for a small portion of the domain. Precipitation forecasts failed to capture the interannual variability, but the predictability of precipitation was higher during ENSO. The ensemble means, as they found, performed as good as or better than the individual forecasts. [Setiawan et al. \(2017\)](#) investigated the precipitation hindcasts from seven NMME models over South Sulawesi and found that the hindcasts were more skillful during June–November and less so during December–May. The hindcasts were overall better during the dry periods. They argued that the climatological monthly precipitation variability played a key role in determining the skill for the hindcasts. [Zhao et al. \(2018\)](#) carried out principal component analysis on the anomaly correlations of 10 sets of global precipitation forecasts from the NMME system and detected spatial patterns over regions where high (low) anomaly correlation for any given initialization time coincides with high (low) anomaly correlation at other initialization times. The temporal patterns from their analysis showed improvement of anomaly correlation with the initialization time, which they argued, was due to the availability of more information (observations and model simulations) for later initialization times. The effectiveness of the temporal patterns was also conditioned upon the efficiency of the data assimilation algorithm.

Although the studies discussed above provide an overview of the performance of different models within the NMME system, several limitations persist. First, most of the above studies are regional assessments, which makes it difficult to generalize their results to other regions. Second, none of the studies considered all the models currently available within the NMME system (there are 16 in total that has both

precipitation and temperature forecasts), raising concerns about the performance of the excluded models. Third, not all of them considered both precipitation and temperature in the skill assessment. Fourth, none of them provided an assessment of the spatio-temporal variability of the precipitation or temperature fields across different models.

Here, we overcome these limitations by carrying out a comprehensive global-scale assessment of the complete set of models currently available within the NMME system (16 models with 178 ensembles in total). Our assessment comprises two main parts: 1) assessment of the forecasting skill across different locations and seasons for different lead times and 2) comparison of the spatiotemporal variability structures of the model forecasts and references. The overarching objective of this study is to evaluate how different models perform across the globe in terms of seasonal precipitation and temperature forecasts, which is beneficial for 1) model developers to improve their models and 2) users of seasonal forecasts to know which models to use.

2. Data and methods

a. Data

We assessed the performance of 16 models within the NMME system, as shown in Table 1. The number of ensembles within each of these models varied from 4 to 24, and the common lead time was 8 months (maximum 10 months). Note that for the first lead, the forecast month corresponds to the month of initialization in this case. Two of these models were Canadian (from Environment Canada) and the rest were from various U.S. organizations (NASA, NOAA, NCAR, NCEP, George Mason University, and Columbia University). In this study, we assessed the hindcasts, which had a common time frame of 29 years (1982–2010). There were some data available for 2011 and onward (termed as “forecasts”); however, several models were missing in that dataset. Since our objective was to include as many models as possible, we focused only on the hindcast datasets. One of the models (NASA-GMAO) had some missing months throughout the 29-yr period across all the ensembles. We discarded these missing values while calculating the error statistics for this model. This model was not used in the EOF analysis to avoid any potentially biased assessment. One of the other models (CFSv1) had results available until 2009. Therefore, we ignored the year 2010 in this case while analyzing this particular model.

The state-of-the-art gauge, reanalysis, and satellite-based merged precipitation dataset Multi-Source Weighted-

Ensemble Precipitation version 2 (MSWEP-V2; Beck et al. 2019b) was used as the reference for evaluating precipitation (available for 1979–2017). MSWEP (Beck et al. 2017b) has been shown to perform superior to several other precipitation products in recent evaluation studies covering the globe (Beck et al. 2017a) and the CONUS (Beck et al. 2019a). For temperature, we used the Climate Research Unit TS4.01 (CRU-TS4.01; Harris et al. 2014) gridded observational dataset covering the period 1901–2016. All the gridded analyses in this study were carried out consistently at 1° spatial resolution on a monthly time scale from 1982 to 2010. The reference datasets were upscaled to 1° using bilinear averaging.

b. Assessment of forecasting skill

The first part of the study focuses on the assessment of the forecasting skill. In particular, we investigated the effects of lead times on precipitation and temperature forecasts. We identified models that were skillful across different lead times and checked if models with high precipitation skill were also associated with a high skill for temperature. We further explored how the skill varied based on the location and timing of the year, and whether or not, these variations were consistent for both precipitation and temperature. We further divided the year into four periods to carry out seasonal analyses: March–May, June–August, September–November, and December–February. We divided the entire globe into 21 distinct regions (Tian et al. 2018; Giorgi and Francisco 2000) (Fig. 1). These regions are large enough to include multiple climate model grids, and also correspond to distinct climatic regimes and physiographic settings (Giorgi and Francisco 2000).

The skill of the forecasts was assessed using the Kling–Gupta efficiency (KGE; Gupta et al. 2009; Kling et al. 2012), bias error, variability error, and correlation coefficient. KGE can be written as

$$\text{KGE} = 1 - \sqrt{(r - 1)^2 + (\beta - 1)^2 + (\gamma - 1)^2}, \quad (1)$$

where r is the linear correlation coefficient of the estimated and reference data, β is the ratio of the means of the estimated and reference data, and γ is the ratio of the coefficients of variation (ratio of standard deviation to mean) of the estimated and reference data. The KGE and its three components have their optimum at unity. KGE provides information about the bias, variability, and correlation, which helps identify the dominant error sources. In this study, instead of looking at the bias ratio directly, we considered the bias error. For precipitation, we calculated the normalized bias error (NBE) as

FIG. 1. The 21 regions across the globe considered in this study. Adapted from Giorgi and Francisco (2000).

$$\text{NBE} = \frac{\sum_{i=1}^N X_i - \sum_{i=1}^N Y_i}{\sum_{i=1}^N Y_i}, \quad (2)$$

where X and Y are the estimated and reference data, respectively. NBE was appropriate for precipitation since all values were positive. For temperature, we calculated the bias error (BE) in absolute terms [i.e., $\text{BE} = (1/N)\sum_{i=1}^N X_i - (1/N)\sum_{i=1}^N Y_i$] to avoid the effects of the unit. For example, the bias ratio between 20° and 10°C is 2, however, in the kelvin unit, this ratio becomes 1.03. The absolute bias remains the same in both units. For variability, we calculated the normalized standard deviation error (NSDE) in both cases:

$$\text{NSDE} = \frac{\sqrt{\sum_{i=1}^N (X_i - \bar{X})^2} - \sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}{\sqrt{\sum_{i=1}^N (Y_i - \bar{Y})^2}}, \quad (3)$$

where $\bar{X}$ and $\bar{Y}$ are the means of the estimated and reference data, respectively.

c. Assessment of spatiotemporal variability

Finally, we carried out the empirical orthogonal function (EOF) analysis of the reference data as well as the ensemble mean precipitation and temperature fields to compare the spatiotemporal variability structures

of both. Additionally, in all cases, we also evaluated the arithmetic means of the forecasts to see how useful such a simple averaging scheme was for improving the forecast performance.

The spatiotemporal variabilities of the reference and the model data were studied using the EOF analysis (Wallace and Dickinson 1972; Wilks 1995), which yields the dominant modes of spatial variabilities along with their temporal evolution patterns, i.e., the principal component (PC) time series. If the data are stored in form of a 2D matrix such that one dimension represents space and the other time, the singular value decomposition (SVD) of the data matrix would generate the eigenvectors of space, the eigenvalues, and the eigenvectors of time. The eigenvectors of space are the EOFs and the eigenvectors of time represent the PCs. The eigenvalues represent the variability explained. For example, if the data matrix $\mathbf{X}$ has dimensions $m \times n$ where m represents space and n time, the SVD of $\mathbf{X}$ would produce:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, \quad (4)$$

where the columns of $\mathbf{U}$ ($m \times m$) are the EOFs and the columns of $\mathbf{V}$ ($n \times n$) are the PCs. Thus, the EOF analysis explores the structure and temporal evolution of the dominant modes of variabilities within the data. Note that in regular EOF analysis, the EOFs are orthogonal and so are the PCs. There are some variations to this, where the orthogonality constraint is relaxed in order to explore more realistic variability structures

(e.g., rotated EOFs; Kawamura 1994; Chen and Tung 2018). In addition to that, for large spatial datasets with regular latitude–longitude gridding, some weighting is usually applied (Chung and Nigam 1999) in order to take into account the unequal-area grid issue (grid cell area decreases at higher latitudes). Since the objective of this study was to compare the variability structures across different models, and not necessarily to explain any physical phenomena per se, we resorted to the unrotated and unweighted EOF analysis.

3. Results and discussion

In this model intercomparison study, we investigated several crucial aspects that define the performance of the models in terms of the forecasts. These included the effects of lead time, seasonal and regional variations of skill, spatiotemporal variabilities of precipitation and temperature fields, and the effects of simple multimodel averaging on forecast performance.

a. Assessment of forecasting skill

From the heat maps for different lead times (Fig. 2), it can be observed that the skill of the forecasts, as expressed by the KGE score, was highest for the first lead time and dropped thereafter. However, there were regional variations. For example, some regions showed higher skill even at longer leads (e.g., the Amazon and Southeast Asia) for both precipitation and temperature. The drop in the skill from first to second lead times has already been shown in some previous studies [e.g., Slater et al. (2016) with eight NMME models over the United States], but regional dependencies have not been systematically addressed. Overall, the temperature skill was much higher than the precipitation skill, which also supports the findings of Shukla et al. (2016), who used eight NMME models over East Africa to assess the skill of the precipitation and temperature forecasts. We also see that this high skill was carried along even until the highest lead (up to 8 months in this study) in some regions (e.g., Southeast Asia, the Amazon, East Africa, southern Africa, and the Sahara).

The precipitation forecasting skill was higher overall for the NASA, NCEP, CMC, and GFDL models. The newer generation models performed better both in the case of CanCM and CFS (i.e., CanCM4 was better than CanCM3 and CFSv2 was better than CFSv1). Among the NASA models, GMAO-062012 showed the highest skill for precipitation followed by GEOSS2S. The performance of the IRI models was poor for precipitation. The skill results across some of the models are in agreement with Ma et al. (2016), where the authors found that (out of 11 NMME models) CFSv2,

GFDL-CM2p1 models, and NASA-GMAO exhibited the highest precipitation skill, while the IRI models and CCSM3 showed poor skill, over 17 hydroclimatic regions in China. For temperature, the skill was high for the NASA models, CFSv2, the COLA models, and the CMC models during the first lead time. From the second lead time and beyond, the regional dependence of the skill became more apparent. The NASA models, CFSv2, the GFDL models, CCM4, and the CMC models retained high skills of temperature for longer lead times. The IRI models, CESM1, CFSv1, and CCSM3 showed the lowest overall skill for temperature.

Regionally, the precipitation skill of the models for the first lead time was higher in the Amazon, Southeast Asia, West Africa, South Asia, and Australia. For longer leads, the Amazon, Southeast Asia, and Australia retained high skill. Except for Australia, these are also the regions that receive high overall precipitation. For temperature, the skill was higher in the Amazon, East and West Africa, Southeast Asia, southern Africa, and the Sahara, which was also carried along into the higher leads. Low precipitation skill was evident in Alaska, the Sahara, southern Africa, Tibet, and Central America. These results reflect higher uncertainty of the forecasts in high latitude regions (e.g., Alaska) and also in capturing low precipitation (e.g., the Sahara). For temperature, north Europe, north Asia, western North America, southern South America, eastern and central North America, and Alaska were associated with poor skill of the models, in general.

The KGE score, both in the case of precipitation and temperature, was largely impacted by the variability error (NSDE) and correlation (R) (Figs. S2–S4 in the online supplemental material). Stronger correlations could be seen in Southeast Asia, followed by Amazon, Australia, and Central America in the case of precipitation. For temperature, high correlations were evident in the Amazon, East Africa, the Sahara, and Southeast Asia. The mean bias error was present but did not change much over different lead times, indicating that the forecasts represented the long-term water balance consistently across different leads. The precipitation bias was consistently high in the Sahara, southern Africa, and Tibet, where the average precipitation amount was low. For temperature, consistent high bias was seen in Southeast Asia. For both precipitation and temperature, the models, in general, underestimated the variability during the longer leads. For precipitation, the highest underestimation of variability was seen in Greenland, while Amazon showed a high overestimation of variability for the temperature at all lead times.

FIG. 2. KGE statistics of precipitation and temperature forecasts for different models (y axis) across different regions (x axis). For clear demonstration, only the first four lead months are shown here, while the complete results are provided in the supplemental material (Fig. S1). Other statistics (bias error, variability error, and correlation) are provided in Figs. S2–S4. Note that we first calculated the skill for each month and then calculated the average skill to create this figure. A similar method was followed also for the seasonal skill plots (Figs. S5 and S6).

FIG. 3. KGE skill of the precipitation forecasts during four different time periods within a year. This figure shows the results for Southeast Asia where the models, in general, show higher skill. Two different lead times (1 and 3 months) are presented to show the drop in the skill.

The skill also showed seasonal variations (Fig. 3 shows an example for Southeast Asia; also see Figs. S5 and S6). For precipitation, the skill generally dropped after the first lead time for all four seasons, however, June–August and September–November showed higher skill persistence in some regions, such as Southeast Asia, West Africa, Australia, etc. These regions retained high skill even at longer forecast lead times. The rainy seasons showed higher skill of precipitation forecasts for all the models across all the regions. Overall, the seasonal precipitation skill was lower in the Sahara, southern Africa, and Tibet. Southeast and South Asia showed high skill at longer leads during September–November, which corresponds to the rainy season in those regions (October–December). Some of these results are also in line with Setiawan et al. (2017), where the authors found more skillful precipitation forecasts from seven NMME models over the South Sulawesi region (Indonesia). In the Amazon, most of the rainfall occurs between December and March, during which, the precipitation skill of all the models was generally higher (see December–February skill in Fig. S5). For temperature, June–August forecasts retained higher skill at longer lead times. The midlatitude regions in the Northern Hemisphere showed negative temperature skill, especially for December–February and March–May, which corresponds to the winter season

in the Northern Hemisphere. This implies that, in general, the temperature forecasting performance of the models was worse during the winter. We see that the precipitation skill was higher during June–August in the western North America region, which supports the findings of Wang (2014), who found higher precipitation skill during summer in the western United States; however, the study was limited by the use of only seven models.

Note that the method of calculating the error metric can impact the results and their interpretation. For example, the ensemble mean will likely show lower variability than the individual ensemble members. Ensemble mean calculated from a large ensemble will show lower variability due to the damping of noise (Becker et al. 2014). To rule out this possibility, we compared the mean of the KGE of the individual ensemble member forecast against the KGE of the ensemble mean (see Fig. S7). For both precipitation and temperature, they generally agreed quite well, which gave us confidence that the ensemble mean is a good representative of the overall model forecasts in this case, and the same has been used in the analyses carried out in this study.

b. Assessment of spatiotemporal variability

In this study, we also looked at the similarities between the modeled and reference spatiotemporal variability

structures for both precipitation and temperature, as derived from the EOF analysis. This analysis gives an idea about how close the forecasts are to the corresponding reference in terms of their dominant modes of variability, thereby complementing the skill analysis (previous section). The idea here is to look at the overall patterns and see how similar they are to the reference.

EOF modes can reveal certain data characteristics that cannot otherwise be seen directly in the original dataset. For spatial comparison, we looked at the first mode of variability of all the forecasts and the reference data, since it explains the maximum variability while compared with other modes (see the second and third modes in the Figs. S8 and S9). The first mode of global precipitation reveals ENSO patterns (Dai et al. 1997). As can be seen in Fig. 4, precipitation from the reference data does demonstrate ENSO-like patterns in the lower latitudes, and the amount of variability explained (i.e., 7%) is also comparable to the previous studies (e.g., Dai et al. 1997). All the models captured the first mode of variability pattern quite well for precipitation. Some models, such as the IRI, showed some inconsistencies (e.g., compare the Indian Ocean region), however, the results generally did not vary significantly. The differences were more apparent in the percentage of variability explained, where the multimodel mean explained almost 5 times larger variability. This could be a result of reduced random variability in the multimodel average, which helps explain more variability by the first mode. Note that the sign of the variability is irrelevant for EOF patterns in this case. In other words, if the sign is reversed within a particular EOF pattern, it does not alter the pattern per se. The variability signs were consistent across different models for precipitation.

The differences in the variability structures became more apparent in the case of the first mode of temperature. Although the amount of variability explained was comparable, the patterns for many models were different. For example, CCSM3 and IRI models showed strong dipole patterns for South America, which was not seen in the reference mode. The IRI, as well as the FLOR models, showed strong dipole patterns over North America, which were not evident in the reference EOFs. GFDL-CM2p1 models showed the closest match in terms of the temperature EOFs. Similar to precipitation, the first mode of the multimodel average temperature also explained the maximum variability. Clearly, more research is needed to understand the causes of the differences in the variability patterns of temperature.

The PC time series provide the amplitudes of the dominant modes of variability over time period of analysis (Fig. 5). The temporal analysis part of the EOF

analysis looks into how the dominant modes of variability of the reference and model data are correlated with each other in time. The first and the second PC time series were highly correlated for all the models in the case of precipitation. For the third PC time series, some models (e.g., CFS, GMAO, GEOSS, FLOR-A06), as well as the multimodel mean showed negative correlations. FLOR-B01 and CCSM4 showed strong positive correlations for all three PC time series. For temperature, CanCM3, NCEP models, some GFDL models, and the multimodel mean showed strong correlations for all three PC time series. The performance was poor in the case of the IRI models, FLOR models, CESM1, and CCSM3, as can be seen by the low values of the correlation coefficient.

c. Advantages of multimodel mean

In this study, we assessed the performance of a simple multimodel mean, i.e., the arithmetic average, within the context of forecasting skill and spatiotemporal variability analyses. As can be seen, the advantage of using the mean was not always apparent. In some cases (e.g., precipitation in the Amazon, temperature in the Sahara), the mean showed high skill, however, the best forecasts still outperformed the mean forecasts. For the EOF analysis of precipitation, the mean did not perform significantly better. For temperature, the mean performed well in capturing the PC time series corresponding to the first three modes of variability, however, the spatial patterns showed significant differences as compared to the reference. Some previous studies (e.g., Ma et al. 2016) showed that simple arithmetic averages significantly improved the forecast performance over China. Our results indicate that the simple average did have high skill over Southeast Asia, however, the skill was highest in the case of CFSv2 (see Fig. 2). Thus, there is a clear need for a more comprehensive assessment of the skill of the merged forecasts on a global scale. This also leads to an important question: can more advanced merging methods improve forecasting skill? Studies in the past have shown that for the linear case, the equal weighting performs quite well (DelSole et al. 2013), however, the situation could be different for nonlinear weights. The types of forecasts are also important. It has been shown that, in general, the multimodel mean will be more skillful than the best model if the models produce very different forecasts (Winter and Nychka 2010). The community is familiar with the benefits of multimodel averaging (Hagedorn et al. 2005; Doblas-Reyes et al. 2005), and a plethora of model averaging schemes have already been proposed (e.g., Wood 1978; Raftery et al. 1997; Wanders and Wood 2016; Rings et al. 2012), however, there is no standard procedure for the implementation of these averaging schemes. Each scheme has its own set of pros and cons and is not

FIG. 4. Comparison of the first EOF mode of variability from different models and the reference for both precipitation and temperature. The percentage value following the title in each subplot corresponds to the variability explained.

FIG. 5. Comparison of the first three principal component time series of precipitation and temperature. The figure shows the linear correlation coefficients between the first three modes of variability of the models and the reference.

necessarily equally efficient for all types of problems, i.e., the performance can vary based on the season, location, forecast variable, lead time, etc. Thus, a more rational approach would be to test multiple model averaging schemes and select the ones that are suitable for different aspects of the problem at hand. Moreover, the model averaging schemes are also purely statistical and often ignore the underlying physics. Thus, there is also a need for more physically based model averaging schemes.

4. Overall outlook

In this study, we found that the forecast skill, in general, starts dropping after the first lead time, however, for some regions across the globe (e.g., the Amazon and Southeast Asia) high skill is retained at much higher lead times. The regional dependence of skill becomes more apparent from the second lead time since the forecasts during the first lead time are usually skillful for all the models. High precipitation skill is retained at higher leads in regions that have high precipitation. For low precipitation regions, the models, in general, show high bias. Temperature forecasts are comparatively less skillful during the winter. Temperature forecasts are more skillful than precipitation forecasts across the globe, which is likely due to the longer

temporal persistence in the latter. Forecast uncertainty increases in the high latitude regions. The mean bias error does not change much over the lead times, which means that the bias could be present but consistently so across different leads. For both precipitation and temperature, the models underestimated the variability during the longer leads.

Skill assessments are generally carried out using temporally aggregated metrics, which lack spatiotemporal information. This is the first study to carry out a systematic assessment of the spatiotemporal variability structures of precipitation and temperature forecasts across globe to gain additional insights on the performance of the forecasts. Four different forecast characteristics can be studied based on this analysis, 1) similarities in the spatial patterns of the dominant modes, 2) similarities in the temporal patterns of the dominant modes, 3) percentage variability explained by the dominant modes, and 4) the ability of the dominant modes to explain large-scale climate patterns (rotated EOFs may be better for this). Thus, the EOF analysis presented in this study complements traditional skill analysis in several different ways. Our results show that all the models were able to capture the first mode of variability reasonably well. The percentage of precipitation variability explained showed some differences. The analysis revealed large

inconsistencies in the spatial variability patterns among models for temperature, even though the percentage of variability explained was comparable. Several models showed strong correlation between the principle component time series of the forecasts and reference, which is a desirable property of the forecast.

Collectively, the performance of the model forecasts can be improved by a better understanding of the predictability of different climate processes. The predictability of these processes is reduced significantly at shorter time scales, since the natural variability tends to obscure the predictable signals (Greene et al. 2012; Roy et al. 2018). Local and large-scale natural climate variabilities, atmospheric circulations, moisture advection and recycling, ocean–land–atmosphere interactions—all these phenomena are, therefore, difficult to predict at seasonal time scales. If an event has low predictability, the forecast skill will be affected. Therefore, it is crucial to develop clear understanding of how predictability changes as a function of space and time, which can guide us to improved model performance. For example, if the initial states are corrected in regions and months with high predictability, the forecasts will likely show stronger persistence, i.e., high skill will be retained at higher leads.

In this study, we showed that in terms of skill, the NASA, NCEP, CMC, and GFDL models showed superior performance for precipitation, while the IRI models performed poorly. The NASA models, CFSv2, COLA models, GFDL models, CCM4, and CMC models showed high skill for temperature, while the IRI models, CESM1, CFSv1, and CCSM3 performed poorly. Due to the differences in the structure and configuration of the models in the NMME system, the model errors and deficiencies are also specific to the models. While it is beyond the scope of this study to address the issues in the models individually, some common strategies can be applied across the spectrum to improve the model performance. One of these strategies could be more realistic representation of the initial states. For example, realistic soil moisture initialization has proven to improve temperature forecast in the CanCM3 model (Drewitt et al. 2012). Ideally, multiple initialization schemes should be tested to select the one that performs the best. For example, Liu and Wang (2015) showed that initialization based on CFSR ocean data assimilation outperformed SST-based initialization to produce more accurate surface weather elements, geopotential height fields, and SST anomalies in the CCSM3 model. Another strategy would be to test how well the models capture the relationships between different climate processes and their teleconnections. For example, Sabeerali et al. (2019) showed that CFSv2

does not properly capture the relationship between the Indian summer monsoon rainfall (ISMR) and its teleconnection, Atlantic zonal mode (AZM), which causes forecasts initialized in February to perform poorly in forecasting AZM, ultimately affecting the predictive skill of ISMR. The same study also showed that ISMR prediction skill is also affected by the misrepresentation of the relationship between ISMR and ENSO, which affect the forecasts initialized in May. Controlled experiments need to be carried out to pinpoint the exact issues in the models. In this regard, studies like the current one can help the model developers decide where to look at for the issues. For example, we have shown that the uncertainty in the forecasts is higher in the upper latitudes. Controlled experiments can help study the attributions. The EOF analysis conducted in this study revealed model discrepancies which are otherwise not possible to find from aggregated skill metrics. For example, we have shown that the IRI models have different variability patterns for temperature compared to the reference data and the other models. Model diagnostic studies need to be conducted to narrow down the possible explanations (e.g., model fails to capture the dominant climate processes or the teleconnections thereof). Each model has its own set of modeling issues, related to the model bias, model structural inadequacy, model configuration, uncertainties in inputs, boundary conditions (for regional models), lack of explicit representations of feedbacks, improper model calibration, etc. A better understanding of the predictability issues and proper diagnostics and accounting of the model deficiencies can lead to significant improvement in the seasonal forecasts.

Statistical postprocessing techniques are the alternative ways to improve the forecasting skill. We already discussed about model averaging and its caveats in section 3c. In our case, the arithmetic mean showed good skill in many cases, but they did not necessarily outperform all the individual forecasts. Although we found that some models are outright erroneous in certain aspects, it is unlikely that a model would be poor-performing ubiquitously. On the other hand, any given model hardly performs consistently well in all cases. Each model in the group adds some type of information, either good or bad. Thus, instead of trying to select the best model, which is not necessarily the best from all aspects, it is arguably more useful and productive to look at how to merge the model forecasts in a way that combines the strengths of the individual models. In this regard, the results of this study serve two purposes: first, they provide information about the skill of the forecasts from different aspects, which can be incorporated into

the merging scheme; second, they set a benchmark for the performance of the merged forecasts through the simple arithmetic averaging. Our study also highlights the need for a comprehensive assessment of different model averaging techniques within the context of NMME forecasts on a global scale.

Reconstructed forecasts have also been proposed as a postprocessing technique. For example, Jia et al. (2015) identified the predictable components of precipitation and temperature over lands and applied the high resolution GFDL-FLOR model to find that the reconstructed model forecasts from the predictable components had superior skill compared to the raw forecasts.

Bias correction is one of the other popular statistical postprocessing techniques (Roy et al. 2017a,b, 2020). For example, Khajehi et al. (2017) compared a copula-based Bayesian ensemble post processing approach and the quantile mapping technique to correct large precipitation bias detected in the Great Plains and central United States from 11 NMME models, and showed that both methods effectively correct the bias while the former is more reliable and accurate. Bias correction, however, is not an ultimate solution either. For example, Xu et al. (2019) carried out wavelet-based bias correction and statistical downscaling on eight NMME model forecasts over 518 stations in China and found that although the methods improved the correlation and reduced the root-mean-square error, the uncertainty in the precipitation forecasts was still very high for summer and in cases with extreme wetness. Therefore, there is a clear need for a thorough assessment of different bias correction techniques for NMME forecasts on a global scale.

Limitations of this study

This study is comprehensive in the sense that it looks at several different aspects of forecast performance (seasonal and regional variations of skill, spatiotemporal variabilities of precipitation and temperature fields, and the effects of simple multimodel averaging) in addition to just assessing the forecasting skill across different lead times. However, there are still limitations to it. For example, we only look at the precipitation and temperature forecasts, while many of the models presented here also produce several other variables, which can show differences in the skill. It should be noted that some of these variables are also difficult to evaluate on a global scale due to the lack of reference observations. We also acknowledge that the results presented here are sensitive to the choice of the error metric, which is common for any skill assessment for that matter. However, we believe that it is more useful to look at

different components of the forecast error, which is why we calculate the aggregated KGE statistic, along with three other metrics, i.e., correlation coefficient, error in mean, and error in standard deviation. This evaluation study is also subject to the reliability of the reference data. While we used the best available resources for this purpose, these datasets are not devoid of errors.

5. Concluding remarks

In this study, we carried out a detailed global-scale assessment of the NMME precipitation and temperature forecasts using two state-of-the-art reference datasets, MSWEP-V2 and CRU-TS4.01. To the best of our knowledge, this is the most comprehensive assessment of NMME precipitation and temperature forecasts to date, in terms of the overall scope, the analysis methods, number of models used, and number of regions considered. We focused on several crucial aspects, including the forecasting skill at different lead times, seasonal and regional variations of skill, and spatiotemporal variabilities of precipitation and temperature.

Many of our skill assessment results are in good agreement with relevant past studies. We provide a thorough assessment of the forecast skill of all 16 models over 21 regions and the seasonal characteristics of the forecasts. The results also compared the dominant modes of variability of the forecasts to that of the reference, which is an alternative way of looking at the forecast performance, complementing the skill assessment. The assessment of the spatiotemporal variability structures revealed inconsistencies in the models which was otherwise not possible to observe through temporally aggregated skill metrics. We identified the models that performed well in skill assessment and spatiotemporal variability analysis, and also the ones that showed poor performance. We discussed the potential ways forward through more model diagnostics (i.e., analysis of different initialization schemes, models' ability to capture teleconnection processes, and controlled experiments to identify the source of errors in the model) and comparative assessment of different postprocessing techniques on a global scale (i.e., model averaging, reconstructed forecasts, and bias correction schemes). It is crucial to conduct systematic assessment of the predictability of different climate processes across different regions and time of year. Model intercomparison efforts where the models are run with consistent inputs can collectively address many of these issues.

Acknowledgments. Funding for this research was made available through the NOAA Grant NA14OAR4310236 (Understanding the Role of Land-Atmospheric Coupling

in Drought Forecast Skill for the 2011 and 2012 U.S. Droughts). NMME datasets can be downloaded from <https://iridl.ldeo.columbia.edu>. Websites for downloading the climatic indices are given in Table S1. The reference data MSWEP-V2 and CRU-TS4.01 are available on princetonclimate.com and crudata.uea.ac.uk, respectively. For the EOF analysis, we used the “eof.m” function written by Chad A. Greene, which is based on the “caleof.m” function of Guillaume Maze. The authors express no conflicts of interest.

REFERENCES

- Beck, H. E., and Coauthors, 2017a: Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol. Earth Syst. Sci.*, **21**, 6201–6217, <https://doi.org/10.5194/hess-21-6201-2017>.
- , A. I. J. M. Van Dijk, V. Levizzani, J. Schellekens, D. G. Miralles, B. Martens, and A. De Roo, 2017b: MSWEP: 3-hourly 0.25° global gridded precipitation (1979–2015) by merging gauge, satellite, and reanalysis data. *Hydrol. Earth Syst. Sci.*, **21**, 589–615, <https://doi.org/10.5194/hess-21-589-2017>.
- , and Coauthors, 2019a: Daily evaluation of 26 precipitation datasets using Stage-IV gauge-radar data for the CONUS. *Hydrol. Earth Syst. Sci.*, **23**, 207–224, <https://doi.org/10.5194/hess-23-207-2019>.
- , E. F. Wood, M. Pan, C. K. Fisher, D. G. Miralles, A. I. J. M. van Dijk, T. R. McVicar, and R. F. Adler, 2019b: MSWEP V2 global 3-hourly 0.1° precipitation: Methodology and quantitative assessment. *Bull. Amer. Meteor. Soc.*, **100**, 473–500, <https://doi.org/10.1175/BAMS-D-17-0138.1>.
- Becker, E., H. V. den Dool, and Q. Zhang, 2014: Predictability and forecast skill in NMME. *J. Climate*, **27**, 5891–5906, <https://doi.org/10.1175/JCLI-D-13-00597.1>.
- Cash, B. A., J. V. Manganello, and J. L. Kinter, 2019: Evaluation of NMME temperature and precipitation bias and forecast skill for South Asia. *Climate Dyn.*, **53**, 7363–7380, <https://doi.org/10.1007/s00382-017-3841-4>.
- Chen, X., and K. K. Tung, 2018: Global-mean surface temperature variability: Space-time perspective from rotated EOFs. *Climate Dyn.*, **51**, 1719–1732, <https://doi.org/10.1007/s00382-017-3979-0>.
- Chung, C., and S. Nigam, 1999: Weighting of geophysical data in principle component analysis. *J. Geophys. Res.*, **104**, 16 925–16 928, <https://doi.org/10.1029/1999JD900234>.
- Dai, A., I. Y. Fung, A. D. Del Genio, A. Dai, I. Y. Fung, and A. D. Del Genio, 1997: Surface observed global land precipitation variations during 1900–88. *J. Climate*, **10**, 2943–2962, [https://doi.org/10.1175/1520-0442\(1997\)010<2943:SOGLPV>2.0.CO;2](https://doi.org/10.1175/1520-0442(1997)010<2943:SOGLPV>2.0.CO;2).
- DelSole, T., X. Yang, and M. K. Tippett, 2013: Is unequal weighting significantly better than equal weighting for multi-model forecasting? *Quart. J. Roy. Meteor. Soc.*, **139**, 176–183, <https://doi.org/10.1002/qj.1961>.
- Doblas-Reyes, F. J., R. Hagedorn, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting — II. Calibration and combination. *Tellus*, **57A**, 234–252, <https://doi.org/10.3402/tellusa.v57i3.14658>.
- Drewitt, G., A. A. Berg, W. J. Merryfield, and W. S. Lee, 2012: Effect of realistic soil moisture initialization on the Canadian CanCM3 seasonal forecast model. *Atmos.–Ocean*, **50**, 466–474, <https://doi.org/10.1080/07055900.2012.722910>.
- Finan, C., H. Wang, and J. Schemm, 2016: Evaluation of an NMME-based hybrid prediction system for Eastern North Pacific basin tropical cyclones. *41st NOAA Annual Climate Diagnostics and Prediction Workshop*, Orono, ME, NOAA/NWS, 3 pp., <https://www.nws.noaa.gov/ost/climate/STIP/41CDPW/41cdpw-CFinan.pdf>.
- Giorgi, F., and R. Francisco, 2000: Uncertainties in regional climate change prediction: A regional analysis of ensemble simulations with the HADCM2 coupled AOGCM. *Climate Dyn.*, **16**, 169–182, <https://doi.org/10.1007/PL00013733>.
- Greene, A. M., M. Hellmuth, and T. Lumsden, 2012: Stochastic decadal climate simulations for the Berg and Breede water management areas, Western Cape province, South Africa. *Water Resour. Res.*, **48**, W06504, <https://doi.org/10.1029/2011WR011152>.
- Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. *J. Hydrol.*, **377**, 80–91, <https://doi.org/10.1016/j.jhydrol.2009.08.003>.
- Hagedorn, R., F. J. Doblas-Reyes, and T. N. Palmer, 2005: The rationale behind the success of multi-model ensembles in seasonal forecasting - I. Basic concept. *Tellus*, **57A**, 219–233, <https://doi.org/10.3402/tellusa.v57i3.14657>.
- Hao, Z., X. Yuan, Y. Xia, F. Hao, and V. P. Singh, 2017: An overview of drought monitoring and prediction systems at regional and global scales. *Bull. Amer. Meteor. Soc.*, **98**, 1879–1896, <https://doi.org/10.1175/BAMS-D-15-00149.1>.
- Harnos, D. S., J.-K. E. Schemm, H. Wang, and C. A. Finan, 2017: NMME-based hybrid prediction of Atlantic hurricane season activity. *Climate Dyn.*, **53**, 7267–7285, <https://doi.org/10.1007/s00382-017-3891-7>.
- Harris, I., P. D. Jones, T. J. Osborn, and D. H. Lister, 2014: Updated high-resolution grids of monthly climatic observations - The CRU TS3.10 dataset. *Int. J. Climatol.*, **34**, 623–642, <https://doi.org/10.1002/joc.3711>.
- 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, <https://doi.org/10.1175/JCLI-D-14-00112.1>.
- Kawamura, R., 1994: A rotated EOF analysis of global sea surface temperature variability with interannual and interdecadal scales. *J. Phys. Oceanogr.*, **24**, 707–715, [https://doi.org/10.1175/1520-0485\(1994\)024<0707:AREAOG>2.0.CO;2](https://doi.org/10.1175/1520-0485(1994)024<0707:AREAOG>2.0.CO;2).
- Khajehi, S., A. Ahmadi, and H. Moradkhani, 2017: An effective post-processing of the North American multi-model ensemble (NMME) precipitation forecasts over the continental US. *Climate Dyn.*, **51**, 457–472, <https://doi.org/10.1007/s00382-017-3934-0>.
- 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, <https://doi.org/10.1175/BAMS-D-12-00050.1>.
- Kling, H., M. Fuchs, and M. Paulin, 2012: Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. *J. Hydrol.*, **424–425**, 264–277, <https://doi.org/10.1016/j.jhydrol.2012.01.011>.
- Krakauer, N. Y., 2017: Temperature trends and prediction skill in NMME seasonal forecasts. *Climate Dyn.*, **53**, 7201–7213, <https://doi.org/10.1007/s00382-017-3657-2>.

- Liu, S., and H. Wang, 2015: Seasonal prediction systems based on CCSM3 and their evaluation. *Int. J. Climatol.*, **35**, 4681–4694, <https://doi.org/10.1002/joc.4316>.
- 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, <https://doi.org/10.1002/joc.4333>.
- Raftery, A. E., D. Madigan, and J. A. Hoeting, 1997: Bayesian model averaging for linear regression models. *J. Amer. Stat. Assoc.*, **92**, 179–191, <https://doi.org/10.1080/01621459.1997.10473615>.
- Rings, J., J. A. Vrugt, G. Schoups, J. A. Huisman, and H. Vereecken, 2012: Bayesian model averaging using particle filtering and Gaussian mixture modeling: Theory, concepts, and simulation experiments. *Water Resour. Res.*, **48**, W05520, <https://doi.org/10.1029/2011WR011607>.
- Roy, T., A. Serrat-Capdevila, H. Gupta, and J. Valdes, 2017a: A platform for probabilistic Multimodel and Multiproduct Streamflow Forecasting, *Water Resour. Res.*, **53**, 376–399, <https://doi.org/10.1002/2016WR019752>.
- , —, J. Valdes, M. Durcik, and H. Gupta, 2017b: Design and implementation of an operational multimodel multiproduct real-time probabilistic streamflow forecasting platform. *J. Hydrometeorol.*, **19**, 911–919, <https://doi.org/10.2166/hydro.2017.111>.
- , J. B. Valdés, B. Lyon, E. M. C. Demaria, A. Serrat-Capdevila, H. V. Gupta, R. Valdés-Pineda, and M. Durcik, 2018: Assessing hydrological impacts of short-term climate change in the Mara River basin of East Africa. *J. Hydrol.*, **566**, 818–829, <https://doi.org/10.1016/j.jhydrol.2018.08.051>.
- , —, A. Serrat-Capdevila, M. Durcik, E. Demaria, R. Valdés-Pineda, and H. Gupta, 2020: Detailed Overview of the multi-model multiproduct streamflow forecasting platform. *J. Appl. Water Eng. Res.*, <https://doi.org/10.1080/23249676.2020.1799442>, in press.
- Sabeerali, C. T., R. S. Ajayamohan, and S. A. Rao, 2019: Loss of predictive skill of Indian summer monsoon rainfall in NCEP CFSv2 due to misrepresentation of Atlantic zonal mode. *Climate Dyn.*, **52**, 4599–4619, <https://doi.org/10.1007/s00382-018-4390-1>.
- Setiawan, A. M., Y. Koesmaryono, A. Faqih, and D. Gunawan, 2017: North American Multi Model Ensemble (NMME) performance of monthly precipitation forecast over South Sulawesi, Indonesia. *IOP Conf. Ser. Earth Environ. Sci.*, **58**, 012035, <https://doi.org/10.1088/1755-1315/58/1/012035>.
- Shukla, S., J. Roberts, A. Hoell, C. C. Funk, F. Robertson, and B. Kirtman, 2016: Assessing North American Multimodel Ensemble (NMME) seasonal forecast skill to assist in the early warning of anomalous hydrometeorological events over East Africa. *Climate Dyn.*, **53**, 7411–7427, <https://doi.org/10.1007/s00382-016-3296-z>.
- Slater, L. J., G. Villarini, and A. A. Bradley, 2016: 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. *Climate Dyn.*, **53**, 7381–7396, <https://doi.org/10.1007/s00382-016-3286-1>.
- Thober, S., R. Kumar, J. Sheffield, J. Mai, D. Schäfer, and L. Samaniego, 2015: Seasonal soil moisture drought prediction over Europe using the North American Multi-Model Ensemble (NMME). *J. Hydrometeorol.*, **16**, 2329–2344, <https://doi.org/10.1175/JHM-D-15-0053.1>.
- Tian, D., M. Pan, and E. F. Wood, 2018: Assessment of a high-resolution climate model for surface water and energy flux simulations over global land: An intercomparison with reanalyses. *J. Hydrometeorol.*, **19**, 1115–1129, <https://doi.org/10.1175/JHM-D-17-0156.1>.
- Wallace, J. M., and R. E. Dickinson, 1972: Empirical orthogonal representation of time series in the frequency domain. Part I: Theoretical considerations. *J. Appl. Meteor.*, **11**, 887–892, [https://doi.org/10.1175/1520-0450\(1972\)011<0887:EOROTS>2.0.CO;2](https://doi.org/10.1175/1520-0450(1972)011<0887:EOROTS>2.0.CO;2).
- Wanders, N., and E. F. Wood, 2016: Improved sub-seasonal meteorological forecast skill using weighted multi-model ensemble simulations. *Environ. Res. Lett.*, **11**, 094007, <https://doi.org/10.1088/1748-9326/11/9/094007>.
- , and Coauthors, 2017: Forecasting the hydroclimatic signature of the 2015/16 El Niño event on the western United States. *J. Hydrometeorol.*, **18**, 177–186, <https://doi.org/10.1175/JHM-D-16-0230.1>.
- Wang, H., 2014: Evaluation of monthly precipitation forecasting skill of the National Multi-model Ensemble in the summer season. *Hydrol. Processes*, **28**, 4472–4486, <https://doi.org/10.1002/hyp.9957>.
- Wilks, D. S., 1995: *Statistical Methods in the Atmospheric Sciences: An Introduction*. Academic Press, 467 pp.
- Winter, C. L., and D. Nychka, 2010: Forecasting skill of model averages. *Stochastic Environ. Res. Risk Assess.*, **24**, 633–638, <https://doi.org/10.1007/s00477-009-0350-y>.
- Wood, E. F., 1978: Analyzing hydrologic uncertainty and its impact upon decision making in water resources. *Adv. Water Resour.*, **1**, 299–305, [https://doi.org/10.1016/0309-1708\(78\)90043-X](https://doi.org/10.1016/0309-1708(78)90043-X).
- Xu, L., N. Chen, X. Zhang, Z. Chen, C. Hu, and C. Wang, 2019: Improving the North American Multi-Model Ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. *Climate Dyn.*, **53**, 601–615, <https://doi.org/10.1007/s00382-018-04605-z>.
- Yao, M.-N., and X. Yuan, 2018: Evaluation of summer drought ensemble prediction over the Yellow River basin. *Atmos. Ocean. Sci. Lett.*, **11**, 314–321, <https://doi.org/10.1080/16742834.2018.1484253>.
- Yuan, X., J. K. Roundy, E. F. Wood, and J. Sheffield, 2015: Seasonal forecasting of global hydrologic extremes: System development and evaluation over GEWEX basins. *Bull. Amer. Meteor. Soc.*, **96**, 1895–1912, <https://doi.org/10.1175/BAMS-D-14-00003.1>.
- Zhao, T., Y. Zhang, and I. Chen, 2018: Predictive performance of NMME seasonal forecasts of global precipitation: A spatial-temporal perspective. *J. Hydrol.*, **570**, 17–25, <https://doi.org/10.1016/j.jhydrol.2018.12.036>.
- Zhou, Y., and H.-M. Kim, 2018: Prediction of atmospheric rivers over the North Pacific and its connection to ENSO in the North American Multi-Model Ensemble (NMME). *Climate Dyn.*, **51**, 1623–1637, <https://doi.org/10.1007/s00382-017-3973-6>.