Application of Postprocessing to Watershed-Scale Subseasonal Climate Forecasts over the Contiguous United States

SARAH A. BAKER

Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, and Bureau of Reclamation, Boulder, Colorado

ANDREW W. WOOD

Climate and Global Dynamics Laboratory, and Research Applications Laboratory, National Center for Atmospheric Research, Boulder, Colorado

BALAJI RAJAGOPALAN

Department of Civil, Environmental, and Architectural Engineering, University of Colorado Boulder, Boulder, Colorado

(Manuscript received 18 July 2019, in final form 24 March 2020)

ABSTRACT

Subseasonal to seasonal (S2S) climate forecasting has become a central component of climate services aimed at improving water management. In some cases, operational S2S climate predictions are translated into inputs for follow-on analyses or models, whereas the S2S predictions on their own may provide for qualitative situational awareness. At the spatial scales of water management, however, S2S climate forecasts often suffer from systematic biases, and low skill and reliability. This study assesses the potential to improve S2S forecast skill and salience for watershed applications through the use of postprocessing to harness skills in large-scale fields from the global climate model forecast outputs. To this end, the components-based technique—partial least squares regression (PLSR)—is used to improve the skill of biweekly temperature and precipitation forecasts from the Climate Forecast System version 2 (CFSv2). The PLSR method forms predictor components based on a cross-validated analysis of hindcasts from CFSv2 climate and land surface fields, and the results are benchmarked against raw CFSv2 forecasts, remapped to intermediate-scale watershed areas. We find that postprocessing affords marginal to moderate gains in skill in many watersheds, raising climate forecast skill above a usability threshold over the four seasons analyzed. In other locations, however, postprocessing fails to improve skill, particularly for extreme events, and can lead to unreliably narrow forecast ranges. This work presents evidence that the statistical postprocessing of climate forecast system outputs has potential to improve forecast skill, but that more thorough study of alternative approaches and predictors may be needed to achieve comprehensively positive outcomes.

1. Introduction and background

Subseasonal to seasonal (S2S) climate forecast skill has received greater attention in recent years due to the potential applications of climate forecasts. Many sectors including public health, disaster preparedness, energy, agriculture, and water management would benefit by applying S2S climate forecasts to their specific needs (White et al. 2017). In the public health sector, S2S forecasts could help predict the probability of floods and droughts at longer lead times, which in turn could inform disaster responses and warnings for mitigating such extreme events. Skillful forecasts would help the energy sector anticipate energy demands and could inform the production of renewable energy sources, such as wind or solar power. Seasonal climate outlooks are presently used in the agricultural sector to make operational decisions on crop management, planting, irrigation scheduling, fertilizer application, and commodity pricing.

In the water management sector, skillful forecasts of precipitation and temperature could improve the skills

Corresponding author: Sarah A. Baker, sarah.a.baker@colorado.edu

DOI: 10.1175/JHM-D-19-0155.1

of streamflow forecasts informing projections of runoff volume, water levels in rivers and reservoirs, and water supply availability (Raff et al. 2013). Academic studies have indicated that water managers are reluctant to use climate forecasts due to perceived poor forecast skill, inadequate or misaligned temporal or spatial scale, institutional hurdles such as mandated decision workflows, organizational restraints, and risk aversion (Callahan et al. 1999; Kirchhoff et al. 2013; Rayner et al. 2005; White et al. 2017). Baker et al. (2019) sought to address some of these hurdles by translating and bias-correcting S2S climate forecasts to a watershed spatial unit—U.S. Geological Survey (USGS) hydrologic unit code 4 (HUC4) watersheds—for biweekly, monthly, and seasonal prediction periods. This aggregated forecast product was made available in real time on the S2S Climate Outlooks for Watersheds web-based tool (http://hydro.rap.ucar.edu/s2s/). Baker et al. (2019) found that bias-correction to watershed climatologies improved forecast relevance through tailoring forecast outputs, and reduced bias, but did not improve S2S forecast performance for skill metrics other than bias (e.g., for correlation).

The increased demand from various sectors for S2S climate forecast information motivates an exploration of the potential for multivariate postprocessing methods to increase the skill of forecasts. The S2S time scale (from 2 weeks to 2 months) is a challenging period for climate forecast skill because it falls between shorter and longer, more aggregated time scales when weather forecasts and seasonal climate projections, respectively, exhibit skill (Vitart et al. 2017). In weather forecasting, skill comes from initial atmospheric and land surface conditions, which tend to have less influence with increasing lead time. Seasonal prediction is influenced by land and ocean conditions such as sea surface temperature (SST) and to a lesser extent soil moisture, and their influence via large-scale ocean–climate teleconnection patterns such as El Niño–Southern Oscillation (ENSO), North Atlantic Oscillation (NAO), Pacific decadal oscillation (PDO), and Pacific North American (PNA) pattern. The S2S time scale falls in the gap between when initial conditions dominate forecast skill and when coupled climate system dynamics provide sources of atmospheric predictability.

Many studies have investigated the predictability of this time scale, with an increasing recent emphasis on the weeks 3–4 period. DelSole et al. (2017) explored the predictability of raw Climate Forecast System version 2 (CFSv2) precipitation and temperature forecasts during January and July, and found that winter exhibited more predictability than summer and that predictability was linked to large-scale climate features such as ENSO and the Madden–Julian oscillation (MJO). Their analyses suggested that precipitation and temperature alone exhibit some predictability, but other climate and land surface fields (e.g., SST) could be used to improve week 3–4 forecasts.

There are several strategies to improve S2S climate prediction skill. One approach to improving climate forecast skill is through enhancements to the coupled dynamical climate or Earth system models used to generate climate forecasts. This effort is steadily pursued by the centers that maintain and develop these large-scale dynamical models. For instance, NOAA's operational dynamical model, CFSv2 improved upon its predecessor, CFSv1, through upgrades to nearly all aspects of the prediction system, including data assimilation systems, the model physics and parameterizations, dynamical core, resolution and coupling strategies, which resulted in major improvements to forecast skill (Saha et al. 2014).

A second strategy to improve climate forecast skill lies in the statistical postprocessing of dynamical forecast model outputs. Postprocessing is applied through statistically translating raw, large-scale dynamical model outputs to a regional scale that is useful for local applications, in this case regional water managers (Sansom et al. 2016; Li et al. 2017). Raw dynamical model output typically requires postprocessing or downscaling (a form of postprocessing) to be used in follow-on applications due to systematic biases, unreliable ensemble spread, and/or forecasts' lack of skill. Common statistical postprocessing methods include bias correction, and different forms of regression that uses large-scale circulation features as predictors. In weather prediction, techniques such as model output statistics (Glahn and Lowry 1972) that regress atmospheric predictors from numerical weather prediction (NWP) onto surface meteorological variables have been common for decades. Recently, Hamill and Whitaker (2006) popularized hindcast or reforecast datasets by showing analog techniques applied to weather to medium range climate precipitation can significantly raise the skill of NWP predictions. In the climate forecast context, Tian et al. (2014) found that a locally weighted polynomial regression method showed higher skill than direct spatial disaggregation and bias correction for North American Multi-Model Ensemble (NMME) precipitation and temperature forecasts for Alabama, Georgia, and Florida. Zhao et al. (2017) corroborates this view in showing that bias correction methods alone are insufficient for postprocessing seasonal forecasts because they merely apply a climatological correction without considering forecast skill. The authors demonstrate that forecast calibration techniques (e.g., the Bayesian joint probability method) are needed to account for skill in the

course of adjusting both forecast mean and forecast spread.

Some statistical postprocessing techniques employ additional information from large-scale climate fields to improve dynamical model forecasts. Many studies have focused on improving seasonal precipitation and temperature forecasts (DelSole and Banerjee 2017; Madadgar et al. 2016; Schepen et al. 2014; Ward and Folland 1991; Xing et al. 2016). Methods include analog-year models, regression methods, and empirical orthogonal function (EOF) mode techniques. Madadgar et al. (2016) explored forecasting seasonal precipitation over the southwestern United States using a hybrid statistical–dynamical approach. The statistical approach used an analog-year technique based on copula functions informed by teleconnections such as the PDO, multivariate ENSO index, and Atlantic multidecadal oscillation, and generated weighted NMME model combinations that showed improvements over the raw NMME ensemble mean seasonal precipitation.

Other studies have found value in using model-predicted SSTs instead of empirical climate indices or atmospheric fields. Xing et al. (2016) used partial least squares regression (PLSR; Wold 1966) to predict the principal component (PC) and consequently, forecast summer rainfall over China using winter SSTs and temperature over land. They found that the summer rainfall prediction skill of the PLSR-based method at 4-month lead was significantly higher compared to 1-month lead dynamical model prediction. Another study by McIntosh et al. (2005), explored using PLSR to predict plant growth days using global SSTs and showed higher skills compared to predictions of rainfall.

PLSR has been used in a wide variety of fields, from early applications in economics (Wold 1966) to recent applications in the physical sciences to predict streamflow (Abudu et al. 2010; Mendoza et al. 2017; Tootle et al. 2007), teleconnections (Black et al. 2017), precipitation (Xing et al. 2016), and climate variability (Smoliak et al. 2015). Many of these studies have used climate fields to develop empirical forecasts of variables of interest. Black et al. (2017) employed PLSR with predictor fields of outgoing longwave radiation (OLR), 300-hPa geopotential height, and 50-hPa geopotential height to predict Northern Hemisphere teleconnection patterns at leads of weeks 3–4. Tootle et al. (2007) showed improvements to long lead streamflow forecasts at gauges in the United States using PLSR with previous spring and summer's SSTs.

In this study, we assess whether a linear component-based forecast postprocessing approach can lead to improvements in the skill of S2S forecasts. We apply the aforementioned PLSR to postprocess subseasonal CFSv2 forecasts of weeks 2–3 and 3–4 surface precipitation and temperature at watershed scales. Because tropical and subtropical SSTs have long been identified as a primary source of seasonal temperature and precipitation forecast skill for the CONUS domain (e.g., Quan et al. 2006), we assess whether it is beneficial to incorporate the conditioning influence of a widely used climate system variable, SSTs, in the postprocessing approach. This experiment is presented as a first cut assessment of whether multivariate predictor approaches may outperform the postprocessing of a single predictor, such as precipitation. We then further investigate whether additional climate system variables beyond SSTs may be a useful source of predictability in postprocessed predictions tailored to specific watersheds.

This paper is organized as follows. The second section describes the data used in the study and preliminary data processing. In section 3, we provide a description of the PLSR method and verification metrics. We then summarize results for the forecast assessment with SST conditioning, followed by findings for the predictability associated with a broader range of climate system predictors. We conclude with a discussion of the potential use and hurdles associated with postprocessing approaches in this context.

2. Data

a. *Precipitation and temperature analysis at watershed scales*

The observational dataset used in this study is phase 2 of the near-real-time North American Land Data Assimilation System (NLDAS-2; Xia et al. 2012). NLDAS-2 is an analysis product generated in near–real time that includes hourly precipitation and temperature. The precipitation field is a temporal disaggregation of a gauge-only analysis (from the NOAA/NCEP Climate Prediction Center) of daily precipitation, which includes an orographic adjustment using the 1/8° PRISM climatology. The NLDAS-2 temperature fields are derived from the North American Regional Reanalysis (NARR). NLDAS-2 data are available from 1979 to present at an hourly temporal resolution at a 1/8° grid spacing. Precipitation and temperature fields from NLDAS-2 are spatially and temporally aggregated to biweekly periods at a USGS HUC4 watershed scale over the CONUS domain (Baker et al. 2019). NLDAS-2 fields are translated to a 0.5° grid and temporally averaged to a daily time step. The fields are then areally aggregated to 202 USGS HUC4 watersheds through spatially conservative remapping, and temporally averaged to biweekly periods (weeks 2–3 and 3–4 lead times).

TABLE 1. CFSv2 predictor fields and spatial extents used in this study.

Predictor name	Variable name	Spatial extent
Sea surface temperature	sst	20°S–70°N × 100°E–360°
Geopotential height (500 hPa)	hgt	25°–80°N × 100°–340°E
Specific humidity (2 m)	q2m	20°S–70°N × 100°–340°E
Surface pressure	prs	20°S–30°N × 100°–340°E
Sea level pressure	slp	20°S–30°N × 100°–340°E
Precipitable water	pwt	20°S–70°N × 100°–340°E
Zonal winds (850 hPa)	uwnd	0°–80°N × 100°–340°E
Meridional winds (850 hPa)	vwnd	0°–80°N × 100°–340°E
Outgoing longwave radiation	olr	20°S–20°N × 100°–340°E
Surface temperature	tmp	24°–53°N × 235°–293°E
Surface precipitation rate	prt	24°–53°N × 235°–293°E

b. *CFSv2 climate and surface variable forecasts*

The dynamical climate forecasts used in this study are from the operational fully coupled atmosphere–ocean–land model CFSv2 (Saha et al. 2014). CFSv2 forecasts a variety of climate and land surface variables, including temperature and precipitation rate (hereafter referred to as precipitation), on a 6-h time step with a ~100-km grid resolution. Reforecasts are available from 1999 to 2010 with four initializations each day at synoptic times 0000, 0006, 0012, and 0018 UTC. Reforecast lead times extend to 45 days or to 9 months depending on the forecast initialization time. The CFSv2 precipitation and temperature reforecasts are spatially and temporally aggregated in the same fashion as the NLDAS-2 fields, yielding CFSv2-based HUC4 watershed forecasts over the CONUS domain. We pooled forecasts over a 2-day period (creating 8-member ensembles) to smooth variability in forecast ensemble means from one day to the next. These aggregated HUC4 watershed temperature and precipitation reforecasts are referred to as raw CFSv2 forecasts.

In the first part of this study, we used the predicted CFSv2 SST field, and went further to assess whether addition fields may represent a source of predictability. These fields and the associated spatial extents (domain) selected for use in this study are summarized in Table 1 and visualized in Fig. 1. The CFSv2 predictor fields were spatially aggregated to a 2° grid resolution to reduce computational processing time, and like the CFSv2 precipitation and temperature reforecasts, were then aggregated to biweekly periods and pooled into 8-member lagged ensemble means. The fields and their domains were identified due to their linkages to North American atmospheric circulation and surface climate, and/or their use in prior postprocessing studies (e.g., Koster et al. 2017; Doblas-Reyes et al. 2013; Grantz et al. 2005). Scaife et al. (2014), for instance, found sources of predictability for North American winters in large-scale climate circulation patterns such as NAO, jet stream winds, and sea level pressures.

3. Methods

a. *Partial least squares regression*

Statistical postprocessing can be achieved through a wide range of techniques, and indeed, recent interest in climate forecast postprocessing has delved increasing into nonlinear machine learning approaches (e.g., Hwang et al. 2018). We used the linear PLSR approach alluded to earlier—a components-based regression method similar to principal component regression (PCR) that combines features of principal component analysis (PCA) and multiple linear regression (Abdi 2010). PLSR forms predictor components that are ordered to explain the maximum covariance of the predictors and a single-valued predictand, while principal component analysis forms components that are ordered to maximize only the explained variance of the predictors. PLSR provides for dimension reduction and avoids multicollinearity in analyses with large sets of cross-correlated and/or dependent predictors, such as is common in gridded model fields.

The PLSR method is detailed in papers such as Abdi (2010) and Smoliak et al. (2010), and is summarized here. The predictors $\mathbf{X}$ (specified as a two-dimensional matrix in which the rows define time records and the columns define the spatial elements—the grid cells of the CFSv2 predictor space) can be decomposed through the following relationship:

$$\mathbf{X} = \mathbf{Z}\mathbf{P}^{\mathrm{T}} \quad \text{with} \quad \mathbf{Z}^{\mathrm{T}}\mathbf{Z} = \mathbf{I},$$

where $\mathbf{Z}$ is the latent vectors or scores (sometimes referred to as partial least squares, or PLS, predictors). The term $\mathbf{P}$ contains their loadings (weights in space), and $\mathbf{I}$ is the identity vector. Similarly, the predictand $\mathbf{Y}$

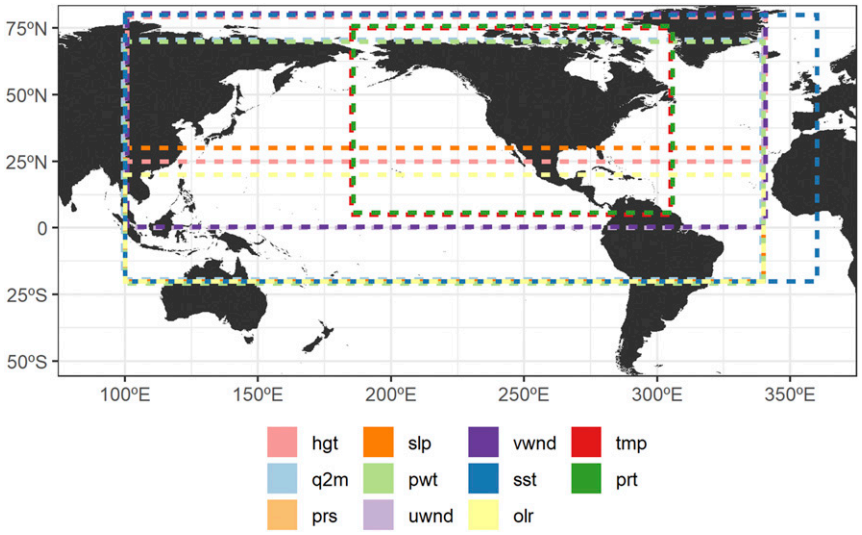

FIG. 1. Spatial extent of CFSv2 predictors corresponding to variables in Table 1.

(which in our application is a vector of the watershed predictand, precipitation or temperature, for time records matching those in **X**) can be estimated through the relationship:

$$\hat{\mathbf{Y}} = \mathbf{ZBC}^{\mathrm{T}},$$

where $\hat{\mathbf{Y}}$ is the deterministic estimate of **Y**, **B** is the regression weights in a diagonal matrix, and **C** is the weights matrix of the predictand. This system of equations does not have enough information to be solved; additional conditions are required to solve for the latent vectors **Z**. To find the latent vectors, two sets of weights, **w** and **c**, are found that form linear combinations of **X** and **Y** that maximize the covariance and find structures that explain the most variance in the predictor field:

$$\mathbf{z} = \mathbf{Xw} \quad \text{and} \quad \mathbf{u} = \mathbf{Yc},$$

with the following constraints

$$\mathbf{w}^{\mathrm{T}}\mathbf{W} = 1, \quad \mathbf{z}^{\mathrm{T}}\mathbf{z} = 1, \quad \text{and} \quad \mathbf{z}^{\mathrm{T}}\mathbf{u} = \text{maximal}.$$

This estimation is performed iteratively to obtain all of the necessary vectors. Once the first latent vector **Z** is solved such that $\mathbf{z}^{\mathrm{T}}\mathbf{u}$ is maximized, it is subtracted from **X** and **Y** through an ordinary least squares regression to form a matrix composed of residuals. The processes are then reiterated to solve for the resulting predictor from this residual matrix. This process can be done using algorithms such as the SIMPLS (de Jong 1993) and ensures that the latent vectors are mutually orthogonal components with respect to the predictors and predictand. Further details on how specifically PLSR was applied in this application is given in the appendix.

Smoliak et al. (2015) investigated PLSR performance related to Northern Hemisphere air temperature variability. The predictand types tested were 1) point-wise where the predictand is a time series for a single grid point or an area average, 2) PC-wise where the target is a PC time series, and 3) field-wise where the predictand is an entire field. They found that point-wise and PC-wise PLSR methods explained more variance in the predictand with a lower number of predictor vectors, as expected, and that all performed slightly better than PCR. In this analysis, we apply the point-wise predictand approach where we predict individual watershed-averaged NLDAS-2 precipitation and temperature at biweekly periods of 2–3 and 3–4 weeks. Both the predictors and predictands are standardized with a mean of 0 and a standard deviation of 1 to remove emphasis on predictor regions with relatively large amplitudes of variation. Standardization alone does not correct distributional issues (e.g., skewness, intermittency) that would undermine the use of a predictor such as precipitation in regression-based methods. As noted earlier, the use of biweekly aggregations improves intermittency and to some extent improves normality, but the application of normality transforms would almost certainly improve upon the results presented here.

A separate PLSR analysis is performed for each watershed with one model variable for each month. PLSR models are trained using data from the adjacent months meaning each year of data has 3 months of data available to train the model. For example, the PLSR model for a

forecast of 1 January for the week 2–3 predictand would be trained using CFSv2 predictors and NLDAS-2 analyses from all forecast–analysis pairs in December, January, and February. The PLSR models are completely cross validated by separating the 12-yr reforecast period into training and verification periods—in this case, by dropping the year in which forecasts are verified from the dataset training period. The nominal training and test sample sizes are approximately 1001 (11 years × 91 days or 3 months) and 31, respectively, although the use of lagged ensembles reduces the effective sample sizes due to serial autocorrelation (dependence). The analysis utilizes the R statistical software package "pls" (Mevik and Wehrens 2019) to perform PLSR. Although the pls function does offer train and predict modes, we further separate the training and test data before applying the pls function as an additional measure to enforce that test period data cannot influence the component training, and to allow for analyses on cross-validation samples not possible through using the internal pls cross-validation function.

b. *Verification metrics*

Verification metrics are applied to compare the performance of ensemble-mean precipitation and temperature forecasts from PLSR-based postprocessing with raw watershed-scale forecasts from CFSv2. The main verification metric presented in this paper is the anomaly correlation (ACC), which is commonly used in the climate prediction community to measure the association of forecast and observed anomalies (avoiding the boost in correlation arising from a correspondence of space–time forecast and observed climatologies). A score of 1 indicates a perfect forecast and a score of 0 or below represents a forecast that is not skillful. Other deterministic forecast verification metrics calculated for this study include mean absolute error (MAE) and bias (not shown). The metrics are calculated separately for all forecasts in each 3-month seasonal basis to show seasonal variability in forecast performance. To translate forecasts and observations into anomalies, the precipitation and temperature climatologies for each watershed, lead, and the day of year were estimated based on averaging across a 15-day window (±7 days from forecast date).

c. *Additional CFSv2 predictor analysis*

The primary analyses of this paper assess whether a multivariate postprocessing approach that combines SSTs and primary forecast variables, precipitation and temperature, can be used to boost the skill of S2S forecasts at watershed scales. The recognized influence of SSTs (even as indexed by tropical Pacific regions such as Niño-3.4) on North American climate is a key motivation for attempting to incorporate SST information (Wang et al. 2013; Scaife et al. 2014; Barnston et al. 2005). This is a relatively conservative postprocessing strategy, given the ability of postprocessing schemes to mine an extensive suite of climate forecast model variables. In the practice of S2S empirical prediction, however, forecasters and stakeholders note discomfort with forecast models that are overwhelmingly data driven—that is, in which selected predictors are allowed to vary significantly in space and time depending on prediction model fitting—due to the risk that predictor selection is spuriously driven by training sample noise, which is particularly a challenge for S2S forecast contexts. Moreover, if predictors vary from prediction to prediction (e.g., from watershed to watershed, and from initialization date to initialization date), it can be difficult to attribute changes in prediction outcomes to the evolution of individual predictor values, which can be an important narrative for stakeholders.

On the other hand, it is likely that climate dynamics do vary in space and by season, such that an optimal predictor set will also vary. Aside from predictor variables, another choice that must be made is the number of components or predictors to include. There exist quantitative metrics for predictor adoption and component acceptance [e.g., the Bayesian information criterion (BIC); Schwarz 1978] or regularization approaches to reduce the risk of overfitting [e.g., least absolute shrinkage and selection operator (LASSO); Santosa and Symes 1986]. To lessen the risk of overfitting, and based on exploratory data analysis suggesting minimal useful variance explained after the second component, we limit the number of components applied in our forecast models to two. As will be shown, even using only two components appears to lead to overfitting for many watersheds.

After assessing the SST-conditioned PLSR models, we explore whether expanding the consideration of potential CFSv2 circulation-related predictors could further improve postprocessing forecast skill. This exploration is not exhaustive, as the goal of the study is not to optimize an empirical postprocessing model but rather to present a general outlook for the potential enhancement of raw climate model forecast outputs at watershed scales through the addition of circulation-scale predictors in a postprocessing framework. The expanded focus includes additional CFSv2 circulation variables (see Table 1), and we test each predictor individually, performing cross-validated PLSR for each predictand, watershed, lead time, and forecast month.

4. Results

This section first presents a single watershed example to illustrate the approach, before reviewing outcomes

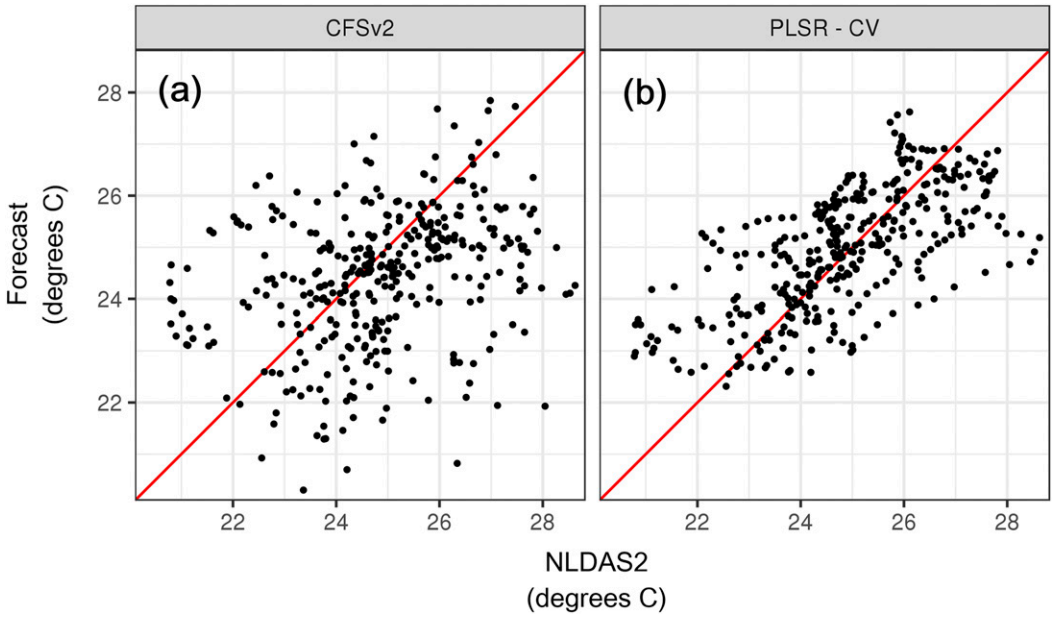

FIG. 2. June weeks 3–4 temperature forecasts are plotted vs NLDAS-2 observations for the Neosho and Verdigris watershed in southeastern Kansas. (a) The raw CFSv2 forecast and (b) the PLSR forecast.

with postprocessing augmented by SSTs for the entire CONUS domain. We then show findings for the exploration of additional predictability in other CFSv2 forecast fields.

a. Individual watershed example

The results of postprocessing varied across watersheds, with some watersheds performing well with predictor components based on SST and precipitation or temperature, while other watersheds showed either negligible benefits or even degradation of skill. Before turning to CONUS-wide results, we illustrate the approach for a single watershed using scatterplots of observations versus the CFSv2 raw and postprocessed forecasts, and maps of the PLSR loadings for each predictor and component. The PLSR loadings provide insight into the regions of the predictor fields that explain the highest covariance between the predictors and the predictand, which in turn is informative about the climate dynamics associated with the PLSR predictors.

The example shows a watershed in which postprocessing was successful in improving the skill of temperature forecasts. Figures 2a and 2b show the raw CFSv2 and postprocessed forecasts, respectively, for June weeks 3–4 temperature in the Neosho and Verdigris watershed in southeastern Kansas. The raw CFSv2 forecast does not differentiate between hot and cold temperature events with an ACC of 0.03 and MAE of 1.4°C over the biweekly period. The PLSR-based forecast reduces the forecast spread considerably, and the forecast distinguishes warm from cold outcomes much better than the raw forecasts. The ACC of the postprocessed forecasts improves to 0.54 and the MAE to 0.95°C. The loadings for the PLSR model are shown in Fig. 3 for SST and temperature. The first component SST loading patterns has strong positive loadings in the northern regions of the Pacific and Atlantic oceans, suggesting that warm temperature anomalies in Kansas are associated generally with warm midlatitude Pacific and Atlantic Ocean temperatures. The first predictor component temperature field also has positive loadings over most of the North American domain, which intuitively covaries positively with Kansas temperature. The second component for both predictors has lower loading magnitudes.

b. Seasonal CONUS domain analysis

The postprocessing approach using SSTs was applied to all CONUS HUC4 watersheds for biweekly periods of weeks 2–3 and 3–4. The PLSR model predictors are concurrent gridded SST and either precipitation or temperature, depending which is the predictand. Verification metrics were calculated for raw ensemble mean CFSv2 forecast and PLSR forecast on a seasonal basis for December–February (DJF), March–May (MAM), June–August (JJA), and September–November (SON).

In considering the following skill plots, which compare correlations of raw and postprocessed forecasts, we assessed whether changes in ACC were statistically significant. To calculate significance thresholds,

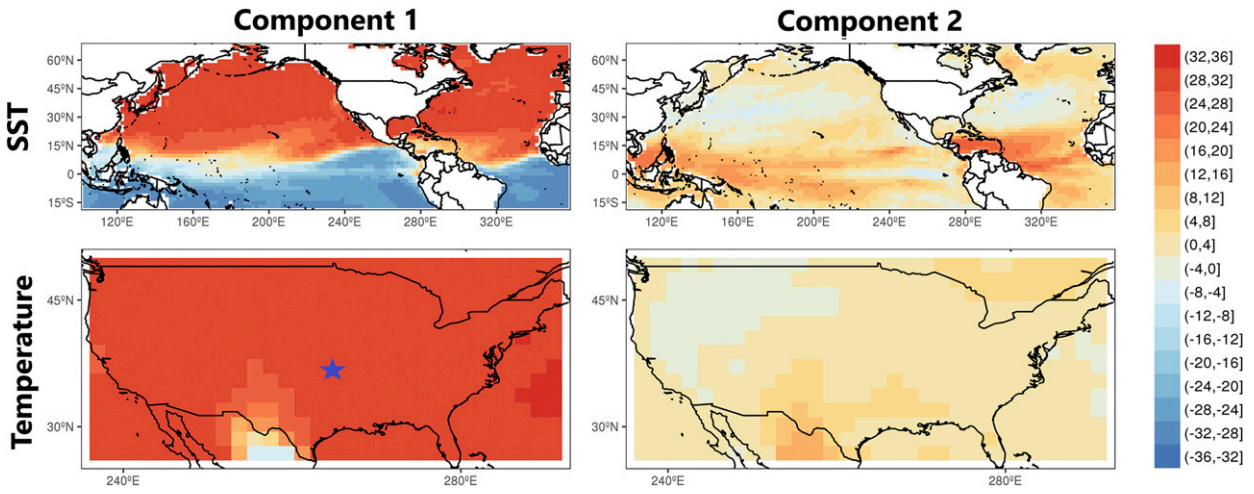

FIG. 3. Mean cross-validated loadings for PLSR model of June week 3–4 temperature forecast for the Neosho and Verdigris watershed. The predictors used in the PLSR model are SST and temperature, which are represented as rows. The two components are shown as columns. The star represents the location of the watershed in the domain.

we estimate an effective sample size that is significantly smaller than the number of forecasts evaluated, due to autocorrelation in the observed predictands (which overlap from forecast date to forecast date). There are ~1080 forecasts in each seasonal sample (12 years × 90 days). Given the noise of the forecasts (there is negligible autocorrelation from week to week), we treat each new week of forecasts as being effectively independent, reducing the sample size to 144, or one forecast per week of the seasonal period for 12 years. To estimate the significance of the difference, we apply the Fisher transformation to the correlation statistics and use $n - 3$ degrees of freedom to estimate standard error for a Student t test (due to sample size limitations). The effective sample size means increases in ACC must exceed 0.15 to be statistically significant at 90% confidence level (versus 0.055 if the sample size included all ~1080 forecasts) and must exceed 0.10 at an 80% confidence level. Recognizing recent arguments against significance testing (e.g., Amrhein et al. 2019), we offer these thresholds as guidance for the user to interpret the results, but nonetheless show all values in the figures rather than obscuring resulting values falling below an author-selected threshold. Aside from the absolute values of the skill increases or decreases, their spatial patterns and/or coherence can also provide support for accepting or rejecting the utility of the postprocessing shown here.

1) TEMPERATURE RESULTS

The raw CFSv2 ACC for weeks 3–4 temperature forecasts (Fig. 4, left column) varies seasonally and spatially over the CONUS domain. The highest raw skill occurs during DJF in the eastern half of the United States, in the northern plains in spring and in eastern Texas and Louisiana in summer, while the western United States does not generally exhibit much skill. The lowest raw CFSv2 skill for weeks 3–4 temperature forecast is during MAM in the Southwest and Rocky Mountains regions, and JJA and SON show mostly lower forecast skill over the entire domain. The center column in Fig. 4 shows the skill (ACC) from the best model (either raw CFSv2 or PLSR). Postprocessing here increases the number of watersheds showing forecast skill values of ACC above 0.3, a potential usability threshold used by forecast groups such the NCEP Climate Prediction Center (O'Lenic et al. 2008). In general, however, weeks 3–4 temperature forecast skill from CFSv2, with or without postprocessing, is not high for large regions of CONUS, at watershed scales.

Figure 4 (right column) shows the difference between the postprocessed and raw forecasts. The largest increases in skill from postprocessing occur during MAM and SON in watersheds of the Intermountain West, in the upper Mississippi during summer, and during SON along the Southeast coast near Florida. These instances, and their regional coherence, are encouraging, yet for most of the watersheds, in all seasons, the postprocessing either has no benefit or degrades raw skill. This outcome is often a result of overfitting predictors, which is a common problem when attempting to fit empirical predictive models to small datasets, and particularly those with noisy or weak predictors, as in the S2S context. The weeks 2–3 temperature forecast skill assessment shown in Fig. 5 revealed this outcome for the raw CFSv2 forecasts as well, performing equal to or better

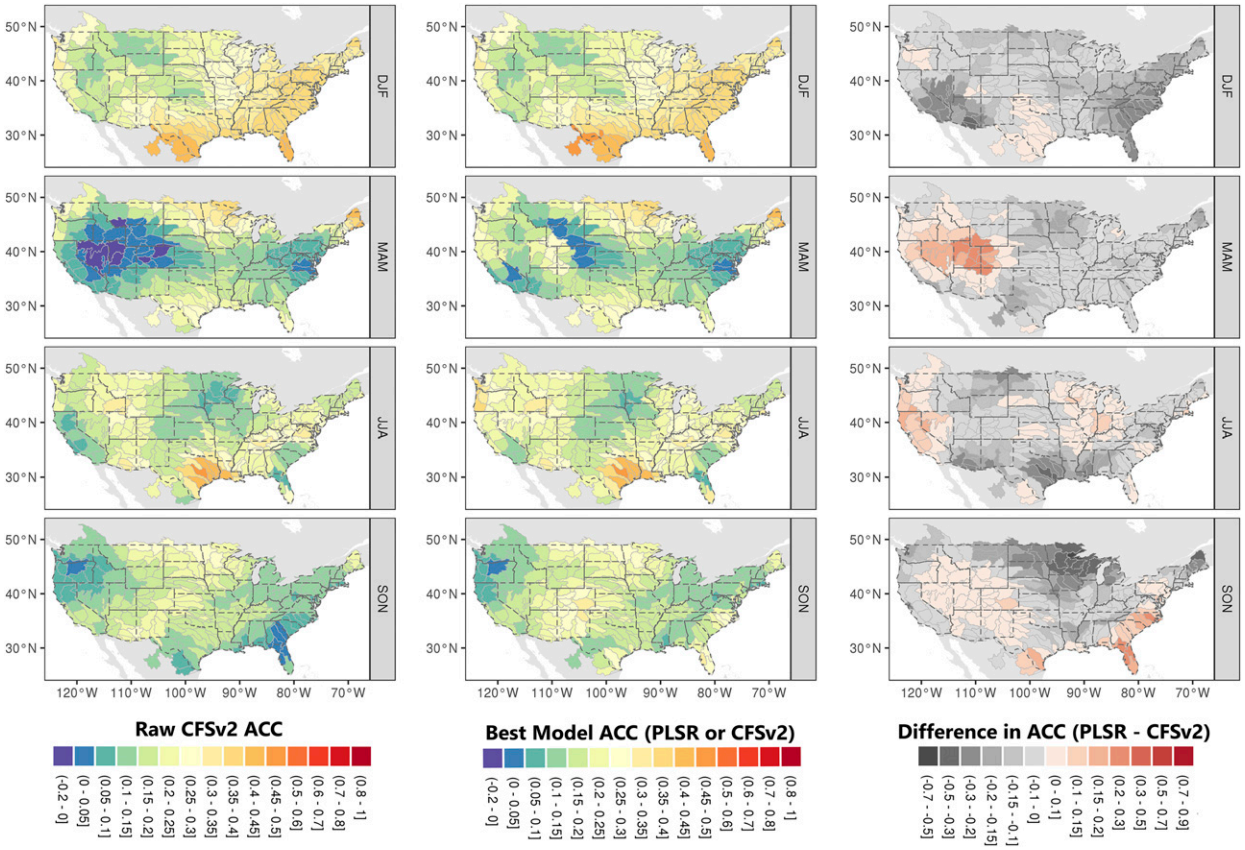

FIG. 4. Forecast results for 3–4 week temperature forecasts shown on a seasonal basis. (left) The raw CFSv2 forecast ACC; (center) the best forecast (either with PLSR-based postprocessing or the raw CFSv2) ACC; and (right) the difference in ACC between postprocessed and CFSv2 raw forecasts. Improvements in ACC (right column) that are greater than 0.10 and 0.15 are significant at 80% and 90% confidence levels, respectively (for details refer to the text).

than postprocessed PLSR results in most watersheds. The weeks 2–3 raw CFSv2 temperature skill is significantly higher (averaging to an ACC of 0.45) over the CONUS domain for all seasons, making it more challenging to improve upon. We note, however, that for all the predictands evaluated here, the statistical postprocessing substantially reduces climatological forecast bias, as it trains to the target observational dataset.

2) PRECIPITATION RESULTS

The seasonal forecast skill results for the weeks 2–3 precipitation forecasts (Fig. 6) show that the raw CFSv2 forecasts have nonnegligible skill in watersheds in the western United States and in the Great Lakes region during DJF. Lower skill values are shown in Texas, Louisiana, Alabama, and Kansas during MAM. During JJA, many watersheds show lower skill except a few watersheds in southern Idaho, northern Utah, and Nevada, which show areas with skill above 0.35. The increase in ACC with postprocessing (Fig. 6, right column) varies with season and watershed location. All seasons have a number of watersheds (but not a majority) that benefit from postprocessing. Watersheds in the north-central and northeastern United States show increases in skill during certain seasons. In some of these watersheds, the ACC increase is high enough to provide usable skill with the postprocessing PLSR model. There are also regions where the postprocessed precipitation forecasts lose skill relative to the raw CFSv2, for example, DJF in the western and southeastern United States. This could be especially true since the raw CFSv2 precipitation is not skillful over large portions of the United States. The spatial pattern of watersheds that show improved skill after postprocessing is notably less coherent for weeks 2–3 precipitation than for weeks 2–3 temperature (Fig. 4). This could indicate that the predictors in component form, such as precipitation, may be poor predictors for modeling precipitation, which is consistent with the fact that the raw watershed precipitation forecast performs poorly.

For weeks 3–4 precipitation forecasts (Fig. 7), the raw CFSv2 forecast ACC is very low with little to no skill for

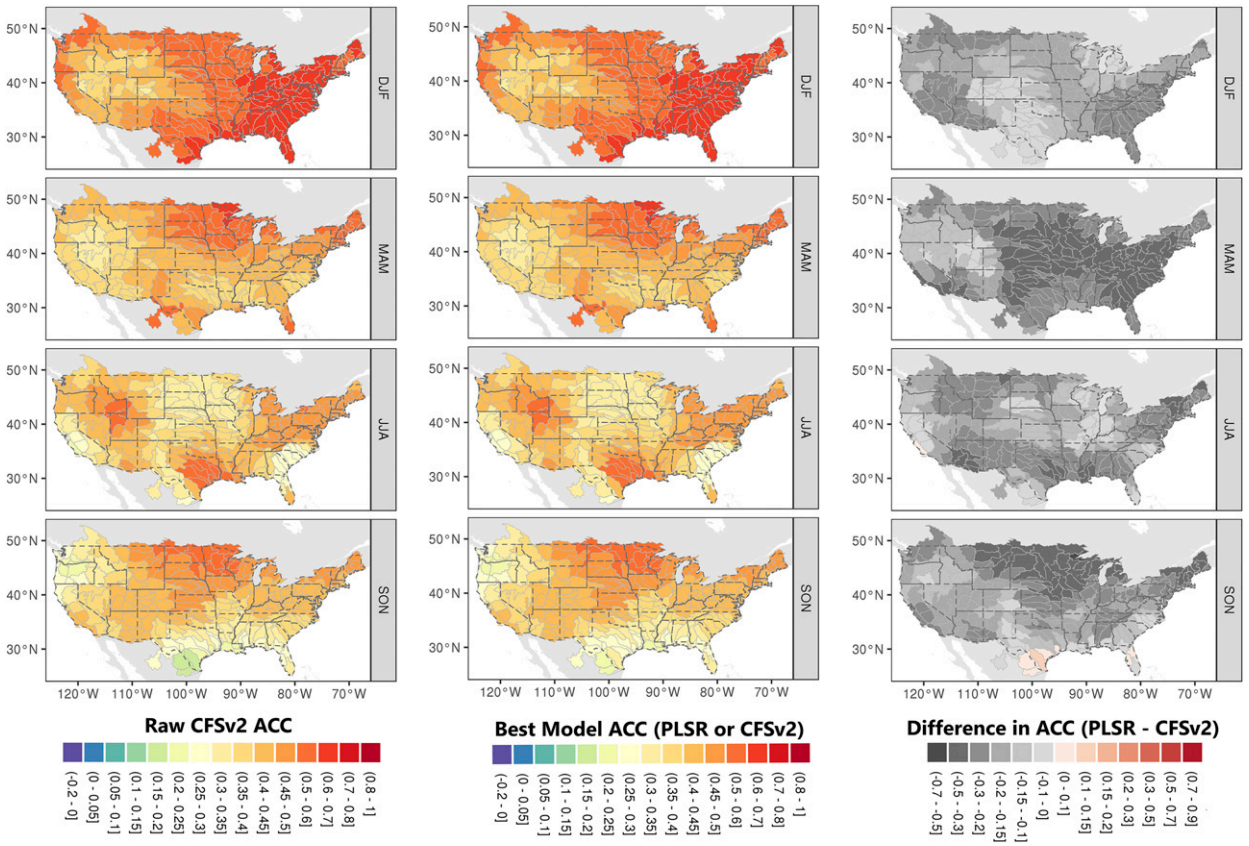

FIG. 5. Forecast results for weeks 2–3 temperature forecasts shown on a seasonal basis. The format is the same as in Fig. 4.

the weeks 3–4 precipitation forecasts for most watersheds and seasons. The raw CFSv2 forecasts are lowest during MAM and JJA over the CONUS domain. The postprocessed ACC values are higher than the raw CFSv2 forecast for many watersheds and seasons, especially during JJA, even exceeding 0.3 in some watersheds. Specifically, a few watersheds in Texas, New Mexico, and North Dakota during DJF show increases in ACC that propel the skill above the 0.3 threshold, as well as a few watersheds in the Northwest in spring. Yet for most watersheds, even moderate to large increases in skill, due to negligible raw skill, do not yield what might be considered skillful forecasts from a water management perspective. It is encouraging that the skill increases are more spatially contiguous, such as in the Pacific Northwest in spring or the upper Mississippi and Missouri River basins in summer, because it may suggest that that synoptic-scale conditioning of the SSTs contributes information to the forecasts.

c. *Additional CFSv2 predictor analysis*

Results thus far have shown that modest improvements in S2S forecast skill for certain regions and predictands may be possible through a relatively restrained multivariate postprocessing approach augmenting CFSv2 forecast components with SSTs, a primary driver of S2S North American climate variability. Yet this approach failed to provide extensive and consistent improvements throughout the forecast domain, thus we explore whether any of the additional predictors in Table 1 could potentially support further improvements. We train and cross validate a more varied set of postprocessed forecasts, using PLSR to predict precipitation or temperature with alternative CFSv2 predictor (Table 1), and we report forecasts for which any predictor could improve skill relative to the raw CFSv2.

Figure 8 illustrates this predictor investigation using the example of predicting weeks 2–3 precipitation for July. The raw CFSv2 skill (Fig. 8a) can be compared with the increases in the ACC skill metric using the highest performing PLSR-based variable (Fig. 8b), which includes watersheds for which postprocessing reduced skill relative to raw CFSv2 forecasts. We also show the maximum skill from either raw CFSv2 or PLSR-based forecasts with skill increases combined (Fig. 8c), and the individual predictor that resulted in the highest ACC

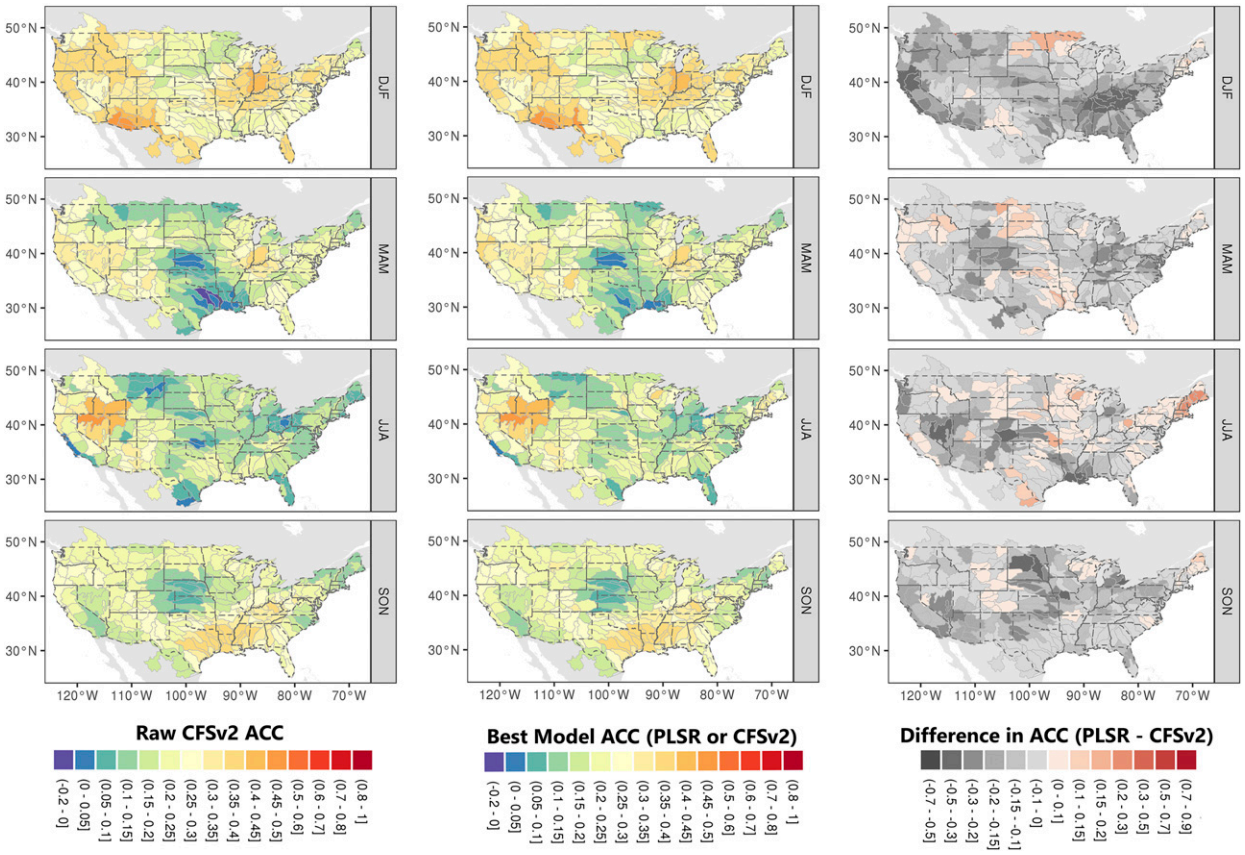

FIG. 6. Forecast results for weeks 2–3 precipitation forecasts shown on a seasonal basis. The format is the same as in Fig. 4.

(Fig. 8d), for those watersheds in which the postprocessing outperformed the raw watershed scale CFSv2 forecast.

In this July weeks 2–3 precipitation forecast example, the raw CFSv2 forecast performed poorly over most of the CONUS domain except in the Four Corners region and areas to the south. Since the raw skill is low, there is a potential for large improvements for many watersheds. The highest increases in skill are found over regions with the lowest raw CFSv2 forecast skill, for example in the Great Plains. Certain watersheds, especially those along the East Coast, do not have statistically significant (by common rejection thresholds) differences in skill between the raw CFSv2 and postprocessed forecasts, but the expanded use of predictors not surprisingly yields a map with more extensive positive results than in earlier figures. The predictors resulting in the highest skill improvements were SST, precipitation, precipitable water, and meridional and zonal winds (Fig. 8d).

Notably, the predictor with the highest impact on forecast skill varies from watershed to watershed, with limited regional consistency. Hypothesizing that multiple predictors may perform better than the raw CFSv2 forecast and that skill differences between predictors are driven to some extent by sample noise (despite the cross validation), we identify the top three predictors for each watershed (Fig. 9) for the July weeks 2–3 precipitation. This allows us to assess whether certain predictors may show more regional consistency in being a generally strong predictor (providing skill above the raw forecasts), than a top place ranking in a noisy field of predictors might otherwise indicate. Watersheds are displayed in color if the predictor ranks in the top three predictors for a watershed based on the ACC. The color is solid if the predictor provides higher ACC than the raw CFSv2 forecast, and transparent if not. For this example, the best predictors are SST, followed by wind speeds (both meridional and zonal), outgoing longwave radiation (OLR), and temperature.

Exploratory data analysis of the type described above was repeated for precipitation and temperature for all lead times in January and July, confirming both the lack of regional consistency for best-performing predictors, as well as underscoring that for almost any watershed of interest, there may be an optimal set of one or more atmospheric predictors that can be harnessed in

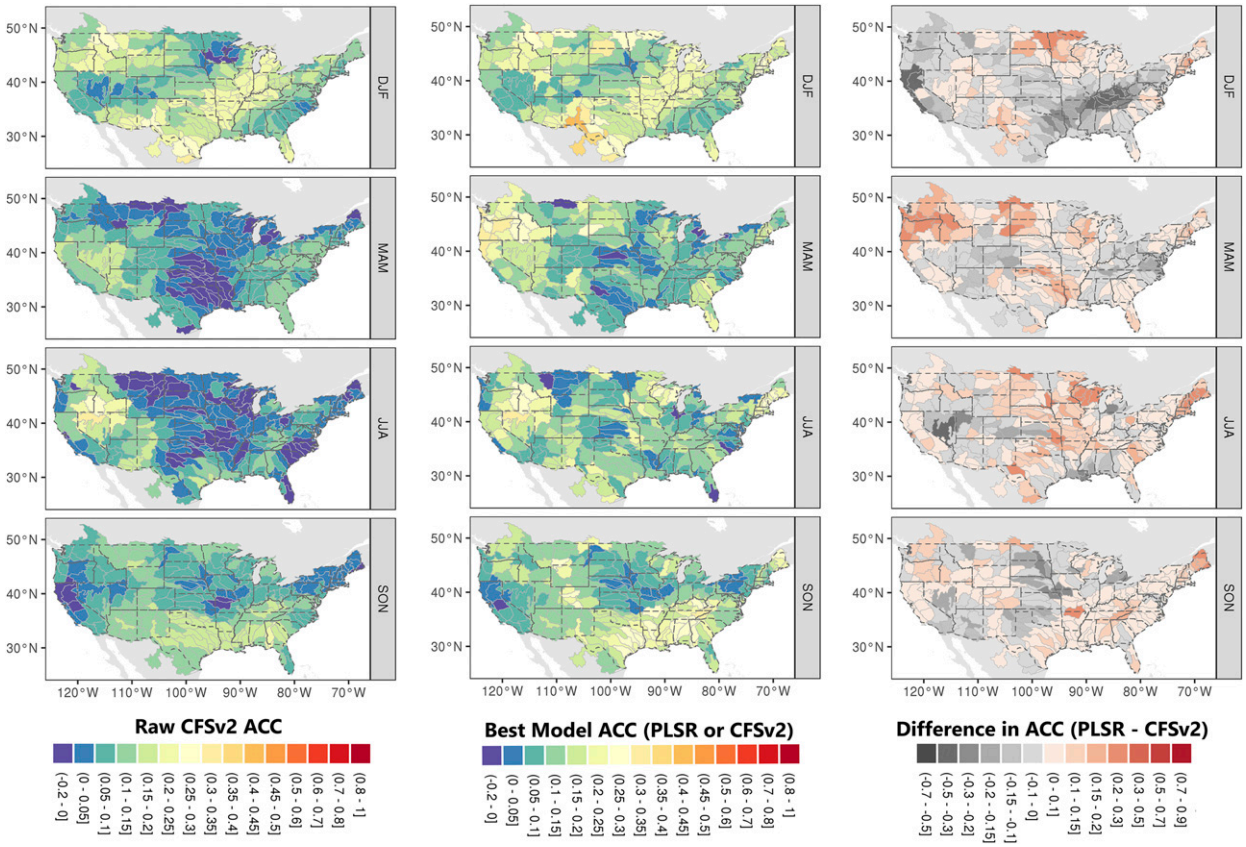

FIG. 7. Forecast results for weeks 3–4 precipitation forecasts shown on a seasonal basis. The format is the same as Fig. 4.

postprocessing to augment the skill of raw CFSv2 forecast output. In additional results obtained for this study, but not reported here, we also found that several other methodological choices (such as the number of components, the strategies used in training, and the use of multiple predictors), could provide additional benefits in optimizing localized postprocessing skill. For instance, we found that training PLSR models on only the extreme quantiles of precipitation events within a training sample could marginally increase forecast skill.

5. Discussion and conclusions

The subseasonal forecast time period has received increasing attention in both the climate forecast and applications communities (U.S. Bureau of Reclamation 2019). Both national projects such as the NOAA S2S Task Force (Mariotti et al. 2018) and international efforts such as the S2S prediction project (Vitart et al. 2017; Vitart and Robertson 2018) are working to improve forecast skill through enhancements of dynamical models and though techniques such as improved data assimilation, as well as through statistical postprocessing of dynamical model output. Some of these studies have used component based empirical regression methods to predict seasonal rainfall, but none have detailed an effort to enhance subseasonal biweekly climate forecasts from dynamical model forecasts by postprocessing or have focused on watershed scale outcomes.

This study's objective was to assess the potential of postprocessing in this context. Experimental results in postprocessing watershed-scale subseasonal climate forecasts via the PLSR method confirms there are opportunities to improve forecast skill via this avenue. Postprocessing of watershed scale biweekly climate forecasts showed that leveraging one additional known source of climate system predictability, SSTs, in postprocessing precipitation and temperature could lead to limited to moderate improvements in many watersheds. In some cases, postprocessing contributed large enough skill increases to produce usable forecasts where the raw forecasts fell below this threshold. Yet in the application shown here, postprocessing also did not perform well for many and even a majority of watersheds for some predictands. It is possible that alternative postprocessing

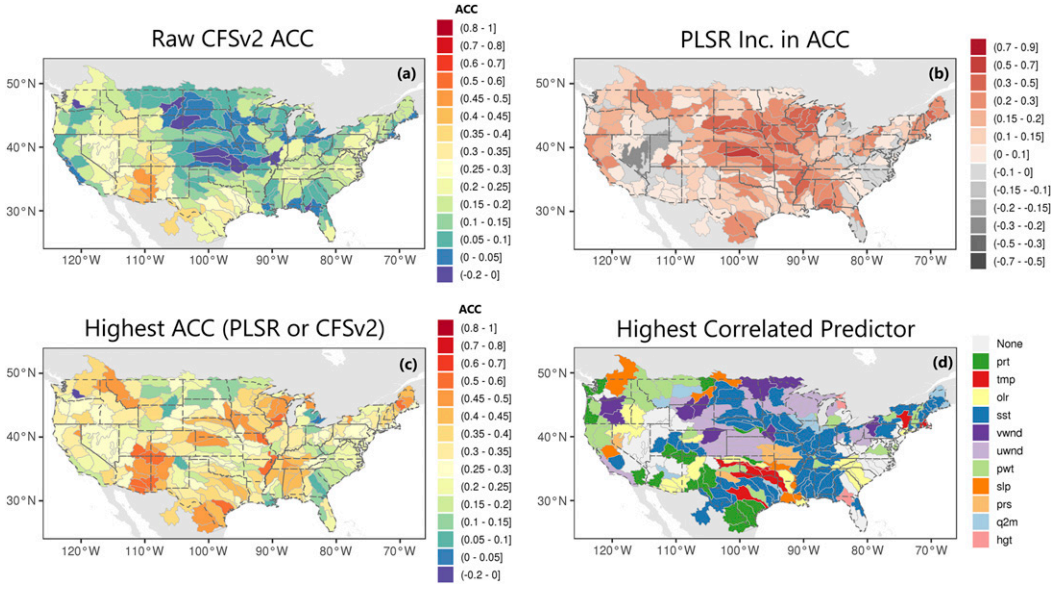

FIG. 8. Visual analysis of predictor performance for forecasts of July weeks 2–3 precipitation. Shown are (a) raw CFSv2 ACC, (b) increases in ACC from raw CFSv2 to PLSR with best predictor, (c) maximum ACC from either the raw CFSv2 or PLSR forecasts resulting in an increase in ACC, and (d) predictor corresponding to the increases in ACC from those shown in (b) and (c). The predictor variables are summarized in Table 1. The gray watersheds did not show improvements in ACC from any predictor using PLSR.

techniques or input climate datasets (such as reanalyses) would improve performance, but there may also be precipitation and temperature variability in certain watersheds that is not systematically forced by identifiable or predictable circulation patterns.

The shortcomings of using a prescribed, linear approach across CONUS watersheds revealed the potential value of investigating further a more data-driven postprocessing application to avoid forecast skill degradation where sampling uncertainty and variations in harnessable predictability lead to overfitting. A one-size-fits-all approach likely excludes additional, potentially usable predictability that may be harnessable through a consideration of dominant regional S2S climate dynamics and more expansive utilization of additional, relevant predictors (either from reanalyses, not considered here, or climate forecast models).

Another caveat to consider in interpreting the study results is that, through unintended oversight, we did not first normalize precipitation predictors before using them in PLSR, thus their distributions did not satisfy the inherent assumption that the predictors and predictands are Gaussian. Though temperature exhibits a Gaussian distribution for most watersheds, raw precipitation is not normally distributed. The time-averaging of precipitation (to 2-week periods) does reduce the distributional problems associated with intermittency, and improves normality, but not sufficiently. To render precipitation into Gaussian space, one can apply statistical transformations such as lognormal, power-law (e.g., square root), Box–Cox (Box and Cox 1964), or log-sinh (Wang et al. 2012), which are typically applied in precipitation and streamflow forecasting (e.g., Strazzo et al. 2019).

This study was scoped to offer a demonstration of concept rather than a comprehensive assessment of postprocessing techniques and dataset opportunities. Further research could home in on specific predictors for subseasonal climate forecasts, test different predictor domains for specific watersheds, use different lagged or pooled ensembles to reduce noise in raw forecasts, regionalize predictors within the CONUS domain, use newer reforecasting datasets than CFSv2, and longer training periods. The postprocessing method we selected, PLSR, may be limited relative to newer machine learning methods that can represent nonlinear and thresholded relationships between variables (Jones 2017), a speculation that can be confirmed if comparative or benchmarking analyses across a range of techniques are performed in the future.

Our focus on the skill of current operational subseasonal climate forecasts on a watershed scale is intended to familiarize potential stakeholders with their raw performance as well as provide an indication of the potential for postprocessing to enhance this performance.

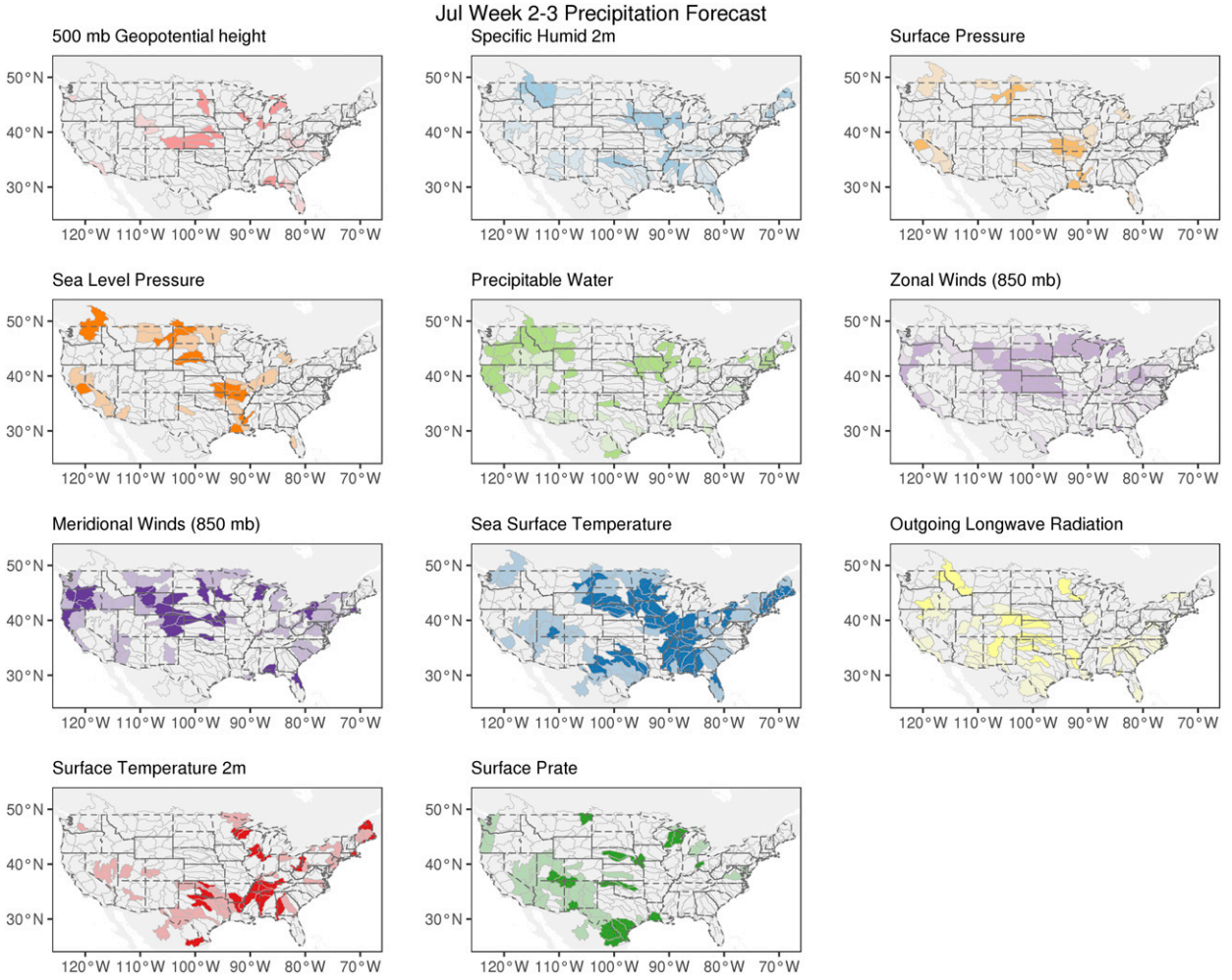

FIG. 9. Top three PLSR predictors for July week 2–3 precipitation. The watershed is colored if the mapped variable has the top three ACC for the watershed. If the PLSR model does not have an ACC that is significantly different than the raw CFSv2 ACC or if the raw CFSv2 forecast has a higher ACC than the PLSR model, the colored watershed is filled with a lighter shade.

To this end, Baker et al. (2019) earlier presented a real-time demonstration of climate forecasts related to those described here on an operational S2S Climate Outlooks for Watersheds web-based platform. Improvements to climate forecasts on such scales may help water managers improve decisions regarding reservoir operations, water allocation, flood control, hydropower generation, water treatment, and in-stream supported releases (Bolson et al. 2013). To wit, a number of studies have shown how S2S climate forecasts can be used to improve the skill of streamflow forecasts (e.g., Werner et al. 2004; Mendoza et al. 2014; Crochemore et al. 2017), a key input to water operations and management. Overall, we recommend an emphasis on postprocessing techniques as part of climate services based on operational climate forecasts because, notwithstanding the limitations of this study, it provided evidence of the potential benefit from the perspective of watershed scale subseasonal climate predictions.

Acknowledgments. This study was supported by NOAA's Climate Program Office's Modeling, Analysis, Predictions, and Projections (MAPP) Program under Awards NA16OAR4310138 and NA14OAR4310238, and the Bureau of Reclamation. The project also participated in the MAPP Sub-seasonal to Seasonal Climate Testbed. The project acknowledges high-performance computing support from Cheyenne (https://doi.org/10.5065/d6rx99hx) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. We also appreciate the scientific and pragmatic insights from Emily Becker and Dave DeWitt, and the helpful comments of two anonymous reviewers.

APPENDIX

Application of PLSR

This section describes further the application of the PLSR-based postprocessing in our study, which used the R statistical software package pls (Mevik and Wehrens 2019). The Mevik and Wehrens (2019) package vignette provides a detailed description of PLSR theory and also provides examples of performing PLSR in R. By way of illustration, we discuss postprocessing the weeks 3–4 precipitation forecasts for initializations in July.

As noted earlier in the main body of the paper, PLSR involves the formation of principal components (PCs), each of which are linear combinations of all elements of the predictor dataset, optimized such to maximize the predictability captured in the leading PC, followed by maximizing the remaining predictability in subsequent orthogonal PCs. The predictor dataset in this case is the daily time series of gridded SST and precipitation dataset, with the variable domain extents as shown in Table 1. Each day's predictor fields are an 8-member lagged ensemble forecast mean, for each grid cell, over the current and previous day's forecast updates (there are 4 per day), with values that are time averaged over the predictand period (e.g., weeks 3–4 ahead of the current day). The observational predictand is a time series of an individual watershed's climate values—in this example, the precipitation for the weeks 3–4 period from the current day.

For use in pls, the training dataset is organized in a large matrix in which the columns are individual grid cell values from the predictor fields and rows represent each record in the time series from the training period (e.g., 1 June 1990, 2 June 1990, . . . , 30 August 2009), when developing a model to predict during 1–31 July 2010. The predictand's training data are a matrix with a single column of observed weeks 3–4 precipitation for an individual watershed with rows matching the same dates as the predictor matrix. Both predictor and predictand datasets are standardized before calling the pls method. The pls method has an internal cross-validation mode that was not used in this study out of an interest in explicitly controlling the separation of training and test datasets, and allowing for broader assessment of training and test dataset characteristics than would be available through using the internal functionality.

We train PLSR forecasts independently for each month of the year, and to increase the sample size of the training, we pool forecasts for the 3-month season centered on the month of interest. In this example, we train the PLSR model for July precipitation on June, July, and August datasets. We cross validate the entire process by dropping a year from the training dataset, leaving 11 years, and verifying on the excluded year. The performance of the forecasts is averaged across all results from all of the excluded years.

REFERENCES

Abdi, H., 2010: Partial least squares regression and projection on latent structure regression (PLS Regression). *Wiley Interdiscip. Rev.: Comput. Stat.*, **2**, 97–106, https://doi.org/10.1002/wics.51.

Abudu, S., J. P. King, and T. C. Pagano, 2010: Application of partial least-squares regression in seasonal streamflow forecasting. *J. Hydrol. Eng.*, **15**, 612–623, https://doi.org/10.1061/(ASCE)HE.1943-5584.0000216.

Amrhein, V., S. Greenland, and B. McShane, 2019: Scientists rise up against statistical significance. *Nature*, **567**, 305–307, https://doi.org/10.1038/d41586-019-00857-9.

Baker, S. A., A. W. Wood, and B. Rajagopalan, 2019: Developing subseasonal to seasonal climate forecast products for hydrology and water management. *J. Amer. Water Resour. Assoc.*, **55**, 1024–1037, https://doi.org/10.1111/1752-1688.12746.

Barnston, A. G., A. Kumar, L. Goddard, and M. P. Hoerling, 2005: Improving seasonal prediction practices through attribution of climate variability. *Bull. Amer. Meteor. Soc.*, **86**, 59–72, https://doi.org/10.1175/BAMS-86-1-59.

Black, J., N. C. Johnson, S. Baxter, S. B. Feldstein, D. S. Harnos, and M. L. L'Heureux, 2017: The predictors and forecast skill of Northern Hemisphere teleconnection patterns for lead times of 3–4 Weeks. *Mon. Wea. Rev.*, **145**, 2855–2877, https://doi.org/10.1175/MWR-D-16-0394.1.

Bolson, J., C. Martinez, N. Breuer, P. Srivastava, and P. Knox, 2013: Climate information use among southeast US water managers: Beyond barriers and toward opportunities. *Reg. Environ. Change*, **13**, 141–151, https://doi.org/10.1007/s10113-013-0463-1.

Box, G. E. P., and D. R. Cox, 1964: An analysis of transformations. *J. Roy. Stat. Soc.*, **26B**, 211–252.

Callahan, B., E. Miles, and D. Fluharty, 1999: Policy implications of climate forecasts for water resources management in the Pacific Northwest. *Policy Sci.*, **32**, 269–293, https://doi.org/10.1023/A:1004604805647.

Crochemore, L., M.-H. Ramos, F. Pappenberger, and C. Perrin, 2017: Seasonal streamflow forecasting by conditioning climatology with precipitation indices. *Hydrol. Earth Syst. Sci.*, **21**, 1573–1591, https://doi.org/10.5194/hess-21-1573-2017.

de Jong, S., 1993: SIMPLS: An alternative approach to partial least squares regression. *Chemom. Intell. Lab. Syst.*, **18**, 251–263, https://doi.org/10.1016/0169-7439(93)85002-X.

DelSole, T., and A. Banerjee, 2017: Statistical seasonal prediction based on regularized regression. *J. Climate*, **30**, 1345–1361, https://doi.org/10.1175/JCLI-D-16-0249.1.

——, L. Trenary, M. K. Tippett, and K. Pegion, 2017: Predictability of week-3–4 average temperature and precipitation over the contiguous United States. *J. Climate*, **30**, 3499–3512, https://doi.org/10.1175/JCLI-D-16-0567.1.

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, https://doi.org/10.1002/wcc.217.

Glahn, H. R., and D. A. Lowry, 1972: The use of Model Output Statistics (MOS) in objective weather forecasting. *J. Appl. Meteor.*, **11**, 1203–1211, https://doi.org/10.1175/1520-0450(1972)011<1203:TUOMOS>2.0.CO;2.

Grantz, K., B. Rajagopalan, M. Clark, and E. Zagona, 2005: A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. *Water Resour. Res.*, **41**, W10410, https://doi.org/10.1029/2004WR003467.

Hamill, T. M., and J. S. Whitaker, 2006: Probabilistic quantitative precipitation forecasts based on reforecast analogs: Theory and application. *Mon. Wea. Rev.*, **134**, 3209–3229, https://doi.org/10.1175/MWR3237.1.

Hwang, J., P. Orenstein, K. Pfeiffer, J. Cohen, and L. Mackey, 2018: Improving subseasonal forecasting in the western U.S. with machine learning. Tech. Rep. 2018-05, Department of Statistics, Stanford University, Stanford, CA, 16 pp., https://statistics.stanford.edu/sites/g/files/sbiybj6031/f/2018-05.pdf.

Jones, N., 2017: How machine learning could help to improve climate forecasts. *Nature*, **548**, 379–380, https://doi.org/10.1038/548379a.

Kirchhoff, C. J., M. C. Lemos, and N. L. Engle, 2013: What influences climate information use in water management? The role of boundary organizations and governance regimes in Brazil and the U.S. *Environ. Sci. Policy*, **26**, 6–18, https://doi.org/10.1016/j.envsci.2012.07.001.

Koster, R. D., and Coauthors, 2017: Hydroclimatic variability and predictability: A survey of recent research. *Hydrol. Earth Syst. Sci.*, **21**, 3777–3798, https://doi.org/10.5194/hess-21-3777-2017.

Li, W., Q. Duan, C. Miao, A. Ye, W. Gong, and Z. Di, 2017: A review on statistical postprocessing methods for hydrometeorological ensemble forecasting. *Wiley Interdiscip. Rev.: Water*, **4**, e1246, https://doi.org/10.1002/wat2.1246.

Madadgar, S., A. AghaKouchak, S. Shukla, A. W. Wood, L. Cheng, K.-L. Hsu, and M. Svoboda, 2016: A hybrid statistical-dynamical framework for meteorological drought prediction: Application to the southwestern United States. *Water Resour. Res.*, **52**, 5095–5110, https://doi.org/10.1002/2015WR018547.

Mariotti, A., P. M. Ruti, and M. Rixen, 2018: Progress in subseasonal to seasonal prediction through a joint weather and climate community effort. *npj Climate Atmos. Sci.*, **1**, 4, https://doi.org/10.1038/s41612-018-0014-z.

McIntosh, P. C., A. J. Ash, and M. S. Smith, 2005: From oceans to farms: The value of a novel statistical climate forecast for agricultural management. *J. Climate*, **18**, 4287–4302, https://doi.org/10.1175/JCLI3515.1.

Mendoza, P. A., B. Rajagopalan, M. P. Clark, G. Cortés, and J. McPhee, 2014: A robust multimodel framework for ensemble seasonal hydroclimatic forecasts. *Water Resour. Res.*, **50**, 6030–6052, https://doi.org/10.1002/2014WR015426.

——, A. W. Wood, E. Clark, E. Rothwell, M. P. Clark, B. Nijssen, L. D. Brekke, and J. R. Arnold, 2017: An intercomparison of approaches for improving operational seasonal streamflow forecasts. *Hydrol. Earth Syst. Sci.*, **21**, 3915–3935, https://doi.org/10.5194/hess-21-3915-2017.

Mevik, B.-H., and R. Wehrens, 2019: Introduction to the pls Package. R package documentation, 24 pp., https://cran.r-project.org/web/packages/pls/vignettes/pls-manual.pdf.

O'Lenic, E. A., D. A. Unger, M. S. Halpert, and K. S. Pelman, 2008: Developments in operational long-range climate prediction at CPC. *Wea. Forecasting*, **23**, 496–515, https://doi.org/10.1175/2007WAF2007042.1.

Quan, X., M. Hoerling, J. Whitaker, G. Bates, and T. Xu, 2006: Diagnosing sources of U.S. seasonal forecast skill. *J. Climate*, **19**, 3279–3293, https://doi.org/10.1175/JCLI3789.1.

Raff, D. A., L. Brekke, K. Werner, A. Wood, and K. White, 2013: Short-term water management decisions: User needs for improved climate, weather, and hydrologic information. Climate Change and Water Working Group, http://www.ccawwg.us/docs/Short-Term_Water_Management_Decisions_Final_3_Jan_2013.pdf.

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, https://doi.org/10.1007/s10584-005-3148-z.

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

Sansom, P. G., C. A. Ferro, D. B. Stephenson, L. Goddard, and S. J. Mason, 2016: Best practices for post-processing ensemble climate forecasts. Part I: Selecting appropriate recalibration methods. *J. Climate*, **29**, 7247–7264, https://doi.org/10.1175/JCLI-D-15-0868.1.

Santosa, F., and W. Symes, 1986: Linear inversion of band-limited reflection seismograms. *SIAM J. Sci. Stat. Comput.*, **7**, 1307–1330, https://doi.org/10.1137/0907087.

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

Schepen, A., Q. J. Wang, and D. E. Robertson, 2014: Seasonal forecasts of Australian rainfall through calibration and bridging of coupled GCM outputs. *Mon. Wea. Rev.*, **142**, 1758–1770, https://doi.org/10.1175/MWR-D-13-00248.1.

Schwarz, G., 1978: Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464, https://doi.org/10.1214/aos/1176344136.

Smoliak, B. V., J. M. Wallace, M. T. Stoelinga, and T. P. Mitchell, 2010: Application of partial least squares regression to the diagnosis of year-to-year variations in Pacific Northwest snowpack and Atlantic hurricanes. *Geophys. Res. Lett.*, **37**, L03801, https://doi.org/10.1029/2009GL041478.

——, ——, P. Lin, and Q. Fu, 2015: Dynamical adjustment of the Northern Hemisphere surface air temperature field: Methodology and application to observations. *J. Climate*, **28**, 1613–1629, https://doi.org/10.1175/JCLI-D-14-00111.1.

Strazzo, S., D. C. Collins, A. Schepen, Q. J. Wang, E. Becker, and L. Jia, 2019: Application of a hybrid statistical–dynamical system to seasonal prediction of North American temperature and precipitation. *Mon. Wea. Rev.*, **147**, 607–625, https://doi.org/10.1175/MWR-D-18-0156.1.

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, https://doi.org/10.1175/JCLI-D-13-00481.1.

Tootle, G. A., A. K. Singh, T. C. Piechota, and I. Farnham, 2007: Long lead-time forecasting of U.S. Streamflow using partial least squares regression. *J. Hydrol. Eng.*, **12**, 442–451, https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(442).

U.S. Bureau of Reclamation, 2019: Teams complete Bureau of Reclamation's Sub-Seasonal Climate Forecast Rodeo—Outperforming the baseline forecasts. Accessed 7 May 2019, https://www.usbr.gov/newsroom/newsrelease/detail.cfm?RecordID=64969.

Vitart, F., and A. W. Robertson, 2018: The sub-seasonal to seasonal prediction project (S2S) and the prediction of extreme events. *npj Climate Atmos. Sci.*, **1**, 3, https://doi.org/10.1038/s41612-018-0013-0.

——, and Coauthors, 2017: The Subseasonal to Seasonal (S2S) prediction project database. *Bull. Amer. Meteor. Soc.*, **98**, 163–173, https://doi.org/10.1175/BAMS-D-16-0017.1.

Wang, F., Z. Liu, and M. Notaro, 2013: Extracting the dominant SST modes impacting North America's observed climate. *J. Climate*, **26**, 5434–5452, https://doi.org/10.1175/JCLI-D-12-00583.1.

Wang, Q. J., D. L. Shrestha, D. E. Robertson, and P. Pokhrel, 2012: A log-sinh transformation for data normalization and variance stabilization. *Water Resour. Res.*, **48**, W05514, https://doi.org/10.1029/2011WR010973.

Ward, M. N., and C. K. Folland, 1991: Prediction of seasonal rainfall in the north nordeste of Brazil using eigenvectors of sea-surface temperature. *Int. J. Climatol.*, **11**, 711–743, https://doi.org/10.1002/joc.3370110703.

Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2004: Climate index weighting schemes for NWS ESP-based seasonal volume forecasts. *J. Hydrometeor.*, **5**, 1076–1090, https://doi.org/10.1175/JHM-381.1.

White, C. J., and Coauthors, 2017: Potential applications of subseasonal-to-seasonal (S2S) predictions. *Meteor. Appl.*, **24**, 315–325, https://doi.org/10.1002/met.1654.

Wold, H., 1966: Estimation of principal components and related models by iterative least squares. *Multivariate Analysis*, P. R. Krishnajah, Ed., Academic Press, 391–420.

Xia, Y., and Coauthors, 2012: Continental-scale water and energy flux analysis and validation for the North American Land Data Assimilation System project phase 2 (NLDAS-2): 1. Intercomparison and application of model products. *J. Geophys. Res.*, **117**, D03109, https://doi.org/10.1029/2011JD016048.

Xing, W., B. Wang, and S.-Y. Yim, 2016: Long-lead seasonal prediction of China summer rainfall using an EOF–PLS regression-based methodology. *J. Climate*, **29**, 1783–1796, https://doi.org/10.1175/JCLI-D-15-0016.1.

Zhao, T., J. C. Bennett, Q. J. Wang, A. Schepen, A. W. Wood, D. E. Robertson, and M.-H. Ramos, 2017: How suitable is quantile mapping for postprocessing GCM precipitation forecasts? *J. Climate*, **30**, 3185–3196, https://doi.org/10.1175/JCLI-D-16-0652.1.