Seasonal Forecasting of Precipitation, Temperature, and Snow Mass over the Western United States by Combining Ensemble Postprocessing with Empirical Ocean–Atmosphere Teleconnections

WILLIAM D. SCHEFTIC,[a] XUBIN ZENG,[a] AND MICHAEL A. BRUNKE[a]

[a] *Department of Hydrology and Atmospheric Sciences, The University of Arizona, Tucson, Arizona*

(Manuscript received 28 May 2022, in final form 29 April 2023, accepted 17 May 2023)

ABSTRACT: Accurate and reliable seasonal forecasts are important for water and energy supply management. Recognizing the important role of snow water equivalent (SWE) for water management, here we include the seasonal forecast of SWE in addition to precipitation (*P*) and 2-m temperature (T2m) over hydrologically defined regions of the western United States. A two-stage process is applied to seasonal predictions from two models (NCEP CFSv2 and ECMWF SEAS5) through 1) postprocessing to remove biases in the mean, variance, and ensemble spread and 2) further reducing the residual errors by linear regression using climate indices. The adjusted forecasts from the two models are combined to form a superensemble using weights based on their prior skill. The adjusted forecasts are consistently improved over raw model forecasts probabilistically for all variables and deterministically for SWE forecasts. Overall skill of the superensemble usually improves upon the skill of forecasts from individual models; however, the percentage of seasons and regions with increased skill was approximately the same as those with decreased skill relative to the top performing postprocessed individual model. Seasonal SWE has the highest prediction skill, followed by T2m, with *P* showing lower prediction skill. Persistence contributes strongly to the skill of SWE and moderately to the skill of T2m. Furthermore, a distinct seasonality in the skill is seen in SWE, with a higher skill from late spring through early summer.

SIGNIFICANCE STATEMENT: Here we test the postprocessing of seasonal forecasts from two state-of-the-art seasonal prediction models for traditionally forecasted elements of precipitation and temperature as well as snowpack, which is important for water management. A two-stage procedure is utilized, including ocean and atmospheric teleconnection indices that have been shown to impact seasonal weather across the western United States. First, we adjust model output based on the average error in historic runs and then relate the remaining error to these teleconnection indices. A final step combines each adjusted model based on its historic performance. Forecasts are shown to improve upon the original models when assessed probabilistically. The snowpack forecasts perform better than temperature and precipitation forecasts with the best performance from late winter through early summer. Persistence is found to contribute strongly to the skill of snowpack and moderately to the skill of temperature.

KEYWORDS: North America; Snowpack; Superensembles; Probability forecasts/models/distribution; Seasonal forecasting; Postprocessing

1. Introduction

Knowledge of the climate for a given season is important to a variety of end-users including disaster preparedness and mitigation, agriculture, power supply generators, as well as water supply managers. Seasonal forecasts can be determined empirically using the state of long-term sources of memory in the ocean to predict temperature and precipitation via composite analysis (Ropelewski and Halpert 1986), analog methods (Livezey and Barnston 1988; Madadgar et al. 2016; Gao and Mathur 2021), linear regression (Barnett 1981), as well as more advanced methods that relate the dominant modes of variability of predictor fields such as sea surface temperature (SST) with the dominant modes of the predictand (e.g., precipitation) (Barnston and Smith 1996; Gershunov and Cayan 2003; Wang et al. 2021). Nonlinear empirical relationships have also recently been explored for forecasting, including using machine learning (Xu et al. 2020; Gibson et al. 2021; Stevens et al. 2021).

Dynamic seasonal forecasting using coupled climate models is able to simulate cyclical processes of the Earth system, such as El Niño–Southern Oscillation (ENSO; Barnston et al. 2019; Johnson et al. 2019), the Madden Julian oscillation (MJO; Kim et al. 2018), and the quasi-biennial oscillation (QBO; Stockdale et al. 2022), as well as propagating these persistent signals to the midlatitude atmosphere via Rossby waves (Vitart 2017). Still, such features are systematically biased in models, resulting in errors in the mean and variability. Therefore, skillful forecasts from these models must be obtained via statistical postprocessing. One approach to this is to simply adjust direct biases in the mean and variance of the predictands through simple mean bias correction or transformations that translate the forecast's distribution to the observed distribution [i.e., quantile matching (QM); Wood et al. 2004;

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/WAF-D-22-0099.s1.

Corresponding author: William D. Scheftic, scheftic@arizona.edu

DOI: 10.1175/WAF-D-22-0099.1

Thrasher et al. 2012]. More advanced methods that correct for interrelated skill and errors in spread through joint calibration between model and observed predictands have been developed (Zhao et al. 2017). There have also been attempts to correct model predictands based on relationships to antecedent patterns in observations as well as in the model (Robertson et al. 2004).

With the availability of multiple ensembles of dynamic model output from several centers, a superensemble (SENS) can be assembled. Efforts to postprocess superensembles have focused on calibration of each model separately (Vigaud et al. 2017) as well as determining the appropriate weights to combine models with different historic performance (Raftery et al. 2005; Gneiting et al. 2005). Furthermore, there have been additional methods developed to synthesize empirical information in combination with model output (Strazzo et al. 2019; Baker et al. 2020).

The focus of most seasonal forecasting has been to forecast temperature and/or precipitation. However, knowledge of the future state of snowpack is a necessity for water management in regions like the western United States where the amount and timing of reservoir replenishment is highly dependent on how much it snows and when it melts (Clow 2010). Pathak et al. (2018) found important relationships between in situ snow water equivalent (SWE) and SST over the various regions of the Pacific through principal component analysis (PCA). Thakur et al. (2020) looked at trends and abrupt shifts in in situ SWE associated with phases of ENSO and found significant negative trends and abrupt shifts in SWE for La Niña and El Niño that were absent in neutral years. Diro and Lin (2020) looked at three ensemble model forecasts of SWE and 2-m temperature (T2m) and found that the prediction of SWE is more skillful than prediction of T2m for forecasts of weeks 3–4. They also found that models had a weaker snow–temperature relationship than was observed, suggesting that some postprocessing techniques could improve forecasts of both SWE and T2m.

Here we evaluate seasonal forecasts of T2m, precipitation (P), and SWE derived from postprocessing two real-time state-of-the-art operational climate models: the fifth-generation seasonal forecasting system (SEAS5) at the European Centre for Medium-Range Weather Forecasts (ECMWF) and the Climate Forecast System version 2 (CFSv2) at the National Centers for Environmental Prediction (NCEP). Our two-stage approach first adjusts for biases in the mean, variance, and ensemble spread, and second adjusts for errors dependent on persistence and well-known modes of land–ocean–atmosphere climate variability. Finally, the prior skill of each adjusted ensemble forecast is used to create the superensemble forecast. Using the adjusted forecasts, we address three questions: How does each step in the postprocessing impact the skill of predicting each variable? Which variable has the most skill and why? How does seasonal prediction change with the annual cycle and are there seasons with consistently higher skill?

Section 2 provides the data used as well as our postprocessing and validation methodology. Section 3 presents the results of the postprocessed forecasts and their performance as a function of season and lead time. Section 4 provides further discussion of the results. Section 5 concludes with the implications of our results with respect to the above questions.

2. Data and methodology

a. *The region of study*

Forecasted and observed T2m, P, and SWE data are areally averaged over the hydrologic unit code (HUC) 4-digit subbasins in the USGS watershed boundary dataset for the western United States and is used to assess the spatial performance of forecasts. Since we are focusing here on forecasts of hydrological quantities of interest to stakeholders like water managers, power supply operators, and agricultural managers, the averages onto hydrological subbasins are more relevant than on a regular grid. The subbasins are chosen as they are relatively large, have hydrologic significance, and vary enough in climate response to be able to analyze the spatial variation of forecast performance.

b. *Data*

Forecasts are produced using model runs from the NCEP CFSv2 (Saha et al. 2014) and the ECMWF SEAS5 (Johnson et al. 2019) and calibrating them against observed monthly 4 km data from the Parameter-Elevation Regressions on Independent Slopes Model (PRISM) (Daly et al. 2008) for T2m and P and the University of Arizona (UA) daily 4-km SWE (Broxton et al. 2016; Zeng et al. 2018).

Here we use the pentad runs initialized four times per day of the CFSv2 from the reforecast (1982–2010) and the operational runs (2011–21), resulting in 24 or 28 ensemble members per month from separate initializations. The ECMWF SEAS5 system consists of 25 members in the reforecast period (1982–2016) and 51 members for operation runs (2017–21).

We also use a pool of climate indices as additional predictors to help improve the final forecast. These predictor indices consist of SST and atmosphere-based teleconnections, as well as the observed standardized anomalies for the target variable of the prior season for a given subbasin (persistence). The number of indices in the pool was reduced manually by looking at a correlation matrix and removing indices that were highly correlated to each other to create more independence between indices and more consistent regression selection. A total of nine indices were used in this research (listed in Table S1 in the online supplemental material). Four [the 109-month average global mean sea surface temperature (GMSST), Niño-3.4 regional mean sea surface temperatures (Niño 3.4), the Atlantic multidecadal oscillation (AMO), and the Western Hemisphere warm pool (WHWP)] of the nine indices showed some residual relationship to other indices. All the indices were preprocessed using a Theil–Sen regression to remove the effect of the preceding four indices to create independent indices.

c. *Postprocessing methodology*

Two stages of postprocessing of each model are tested. The first stage removes biases in the mean, variance, and spread.

The second stage attempts to correct for systematic errors in each model related to teleconnection indices with known effects on the climate. Calibrations are performed separately for each season and lead time out to 3 months.

Bias correction (Stage 1) is undertaken through quantile matching (Wood et al. 2004; Thrasher et al. 2012). This step removes the bias of the model in the mean and variance by matching the forecast distribution to the observed distribution. To adjust a forecast, we find the percentile of the forecast, and then find what value that percentile equals in the observed distribution. We have chosen the simplest distributions with the least number of parameters to fit to each variable that reasonably represents its true distribution. A two-parameter Gaussian distribution is chosen for T2m data. While the temperature trend tends to stretch out the distribution to a certain degree from a true normal distribution, implementing the trend into the adjustment adds an additional uncertainty parameter. Therefore, we do not consider this in our Stage 1 adjustment. The variables P and SWE have highly skewed distributions that are bounded at zero. The mixed gamma distribution is a good approximation for such variables, since it both evaluates the probability of having nonzero P and SWE as well as the distribution of P and SWE conditioned on occurrence. This is also a common distribution used for the standardized precipitation index (SPI).

In practice, forecasts are not automatically converted from percentile to actual P values but rather both the forecast and the observations are transformed from their original distributions to standard normal distributions as standardized anomalies. These standardized anomalies are used in the Stage 1 correction here, as they are directly comparable across variables and different seasons where absolute values of the variables can vary a lot. Furthermore, Stage 2 as well as the ensemble combination used to create the superensemble is simplified greatly by using standardized anomalies. Adjusted forecasts can easily be converted to absolute precipitation and temperature values using the observed distribution parameters.

The benefit of quantile matching is its simplicity and preservation of ensemble spatial patterns. However, it can only correct local biases and cannot adjust for differences between the forecast error and ensemble spread (defined as the standard deviation of ensemble members for a given forecast). Therefore, Stage 1 also incorporates a simple spread correction for each season and lead time such that

$$F_{\text{new},i} = \overline{F} + R(F_i - \overline{F}),$$

where R is the ratio of the ensemble mean forecast error over the forecast spread calculated during the training period, where $\overline{F}$ is the mean forecast and F_i is the forecast of ensemble i. By doing this, the ensemble is either expanded or compressed about its mean.

In Stage 2, we adjust the residual error (i.e., ensemble mean minus observation) from Stage 1 based on the relationship to teleconnection indices from the season prior to forecast initialization. Out of a pool of possible indices (Table S1) we only use two for a given calibration. An initial step is to transform the variable to a standardized anomaly since linear regression is best performed on normal variables.

First, we find the best fit linear regression of this residual error with a predictor index. To avoid overfitting, the correlation squared (R^2) must be greater than a critical value (taken as 0.125). Then a 95% confidence interval of the slope is calculated and the slope with the lowest absolute value is chosen as the fit line to adjust the residual error. If the confidence interval includes a slope of zero or the R^2 is less than or equal to 0.125, the predictor is not selected.

Once we have the regression line, we can subtract this from each ensemble member in the forecast and adjust the new residual error by repeating the above regression to select the second-best index for a given lead time and month. If at any iteration no predictor meets the R^2 and slope interval thresholds, then the procedure is stopped. Effectively, this means there can be 0, 1, or 2 indices selected as predictors for Stage 2, and when it is 0, no correction is made in Stage 2 (i.e., the Stage 2 forecast is the same as the Stage 1). In this way, only relatively strong relationships are selected, and the forecast is only adjusted for the amount that we have some confidence in.

Figure S1 in the online supplemental material shows the progression of ensemble SWE forecasts from SEAS5 at 0-month lead time [e.g., early January prediction for January–March (JFM)] from the original forecast to Stage 1 and then to Stage 2. Stage 1 forecasts adjust the scale and shape of the forecast distribution to match the climatological distribution more closely, whereas the Stage 2 forecast tries to adjust situation dependent errors. The final forecast more closely follows the observed SWE. In the first stage, the root-mean-square error (RMSE) decreases by 14% and the correlation increases by 16%. After Stage 2 we see an additional but small 6% reduction in RMSE and a 3% increase in correlation.

Once each model ensemble forecast is adjusted through the above two-stage process, we combine each ensemble into a SENS with each ensemble member's weight based on the cross-validated R^2 of the associated forecast during the training period. We found a weak but positive relationship between the prior estimated correlation difference between the two models (CFSv2 and SEAS5) and the actual correlation difference in our validation for all three variables. These weightings are used to create a distribution (using a weighted kernel density estimation) from which ensemble members can be chosen. This leverages the skill of each model and still maintains the well-calibrated structure of the bias-corrected model forecast. The weights are applied to individual ensemble members, so that the multimodal nature of the original ensemble forecasts that occasionally happens when ensembles are keying on two or more potential outcomes can be preserved.

d. Validation metrics

The forecasts are validated using three metrics both deterministically and probabilistically. The deterministic forecast represented by the ensemble mean (ENSM) is evaluated using Pearson product-moment standardized anomaly correlations (SAC) as well as the mean square error (MSE) based skill score (SS); i.e., $(1 - \text{MSE}_{\text{ENSM}}/\text{MSE}_{\text{ref}})$ with MSE_{ref}

TABLE 1. RPSS and SAC for seasonal forecasts at 1-month lead; median over all seasons then median across all subbasins. SENS combines CFSv2 and SEAS5. For stage 1 and 2, values that are bolded show significant improvement at the 95% level. For SENS, metrics that are significant above at least one model are italicized, and those significant above both models are also bolded. The results are split into three different validation strategies: last 11 years using prior data for calibration (L11), first 11 years using future data for calibration (F11), and leave 3-out validation (L3O) for all years (1982–2021).

			L11			F11			L3O		
Metric	Variable	Stage	SENS	CFSv2	SEAS5	SENS	CFSv2	SEAS5	SENS	CFSv2	SEAS5
RPSS	T2m	Raw		−0.10	−0.81		−0.13	−0.96		−0.17	−1.01
		1	*0.16*	**0.10**	**0.15**	0.14	**0.12**	**0.11**	***0.12***	**0.08**	**0.11**
		2	*0.16*	0.10	0.15	0.13	0.11	0.11	***0.13***	0.07	0.11
	P	Raw		−0.60	−0.07		−0.47	−0.07		−0.56	−0.08
		1	*0.00*	**−0.04**	**0.00**	0.03	**0.01**	**0.02**	*0.03*	**−0.01**	**0.01**
		2	−0.01	−0.05	−0.01	0.03	0.01	0.01	*0.01*	−0.02	0.00
	SWE	Raw		−3.07	−0.10		−3.71	−0.09		−3.60	−0.07
		1	*0.11*	**0.05**	**0.05**	*0.16*	**0.09**	**0.17**	*0.12*	**0.02**	**0.10**
		2	*0.16*	**0.08**	0.11	*0.20*	0.12	0.16	***0.20***	**0.09**	**0.16**
SAC	T2m	Raw		0.27	0.30		0.36	0.35		0.45	0.51
		1	0.36	0.26	0.31	0.35	0.34	0.31	*0.51*	0.45	0.49
		2	0.34	0.26	0.29	0.33	0.34	0.30	*0.51*	0.44	0.49
	P	Raw		0.23	0.17		0.31	0.23		0.21	0.25
		1	0.18	0.23	0.17	0.29	0.31	0.22	*0.26*	0.20	0.24
		2	0.17	0.17	0.16	0.23	0.26	0.20	*0.22*	0.22	0.22
	SWE	Raw		0.15	0.34		0.42	0.60		0.30	0.55
		1	*0.45*	0.18	**0.45**	*0.57*	0.38	0.58	*0.56*	0.31	0.52
		2	*0.52*	**0.31**	0.47	*0.60*	**0.50**	0.57	*0.61*	**0.45**	**0.59**

being the MSEs of the reference forecast (using the climatological mean over the training period). A score of 1 for SS is a perfect forecast, a score of zero has the same skill as the reference forecast, and a negative score represents forecasts that are worse than the reference.

To assess the skill of the full ensemble forecast, we use the ranked probability skill (RPS) score (RPSS). RPSS is derived from the forecast ensemble RPS ($\mathrm{RPS_{Fcst}}$) and the reference ensemble forecast ($\mathrm{RPS_{ref}}$) such that

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS_{Fcst}}}{\mathrm{RPS_{ref}}},$$

$$\mathrm{RPS} = \int_{x=-\infty}^{x=\infty} (\mathrm{CDF_{Fcst(or\_ref)}} - \mathrm{CDF_{Obs}})^2 dx,$$

where CDF is the cumulative distribution function for the standardized anomalies (denoted by x) from negative infinity. The reference forecast for a given season is an ensemble containing all observations for that season from the training period in which each member consists of the observations from one of the training years. RPSS can be considered a probabilistic analog to the MSE-based skill score where similarly a score of 1 is a perfect score, a score of 0 has the same skill as the reference forecast and a negative score represents forecasts that are worse than the reference forecast. It is also considered an estimate of the uncertainty, reliability, and resolution of a probabilistic forecast (Murphy 1973).

e. Validation strategies

All forecasts are validated using a prior training with update approach. Validation is done on an 11-yr period from 2011 to 2021 (L11), while the training period includes all years from 1982 up to the year prior to the forecast, with each new year of forecasting updating the training period to include the previous year's data to simulate an operational forecast mode. This is considered a more accurate assessment of error than typical leave-*n*-year-out cross-validation approaches (Risbey et al. 2021). Cross validation can be especially misleading for regression-based forecasts including the assessment of trends and predictor selector biases such as could be the case with our two-stage approach (Barnston and van den Dool 1993; DelSole and Shukla 2009).

Other seasonal forecasting research has expressed the ideal of using a prior training with update validation approach but cited the limited number of years as making it infeasible for estimating performance metrics (Shao et al. 2022). Therefore, to assess the robustness of our results we have also implemented two additional validation strategies, and these results will be discussed in section 4. The first strategy validates forecasts from 1982 to 1992 (F11) where each validation year uses all future years as the training period. The second additional strategy is a leave-3-years-out cross validation (L3O) for forecasts from 1982 to 2021 where each validation year uses all remaining years for training except the year prior, year of, and year after the validation year. The 40 years in this validation strategy provides more stability in the performance metrics but it can overestimate forecast performance (Risbey et al. 2021).

For T2m and *P*, all months are used for all subbasins in our analysis. However, since snow is ephemeral, the lack of snow or limited sample of years with sufficient snow in certain subbasins will cause overall averaged skill scores to be

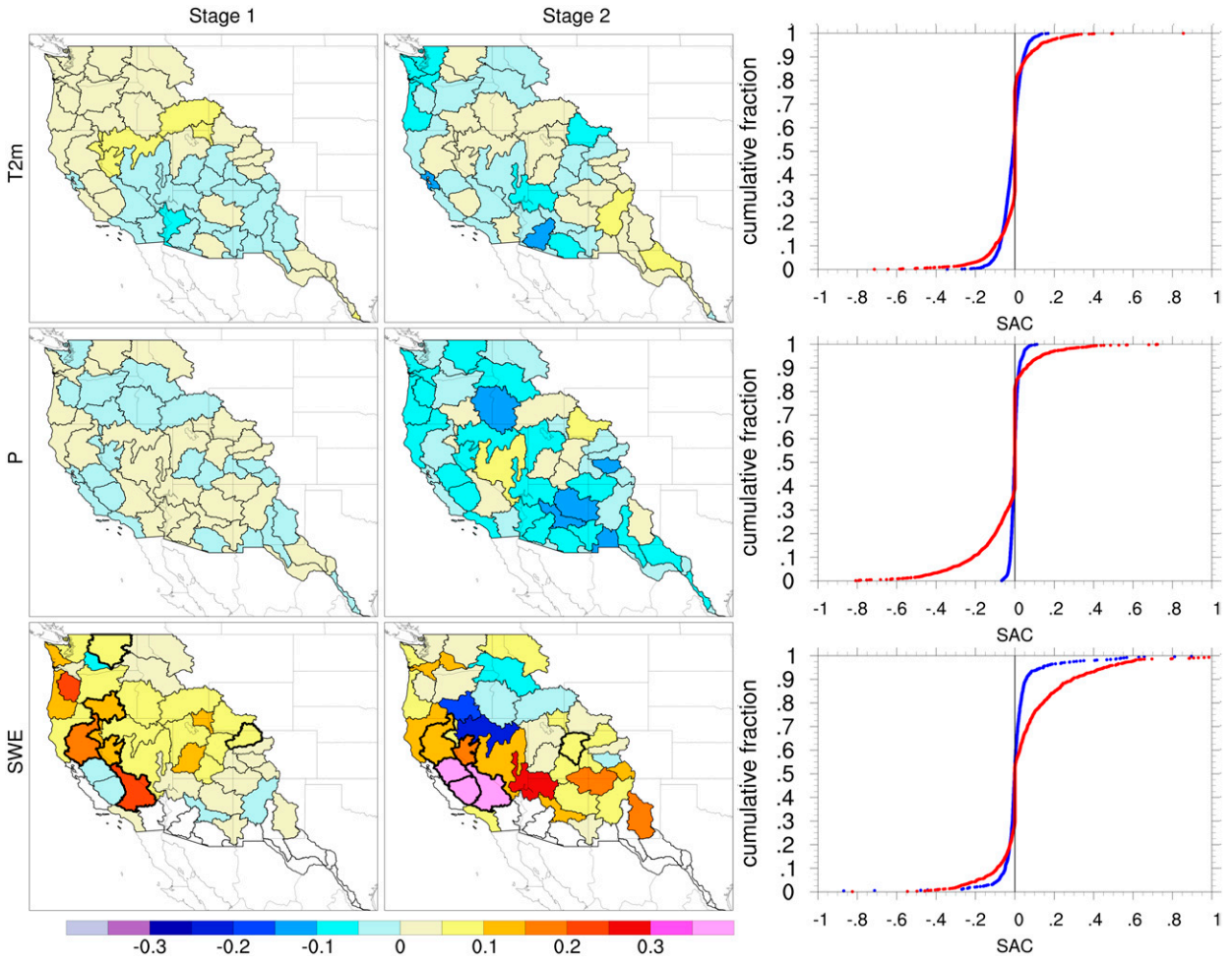

FIG. 1. (left) The differences between original and Stage 1 forecasts and (center) the differences between Stage 1 and 2 forecasts of the average of CFSv2 and SEAS5 median SAC across all seasons for all HUC 4-digit subbasins in study for (top) T2m, (middle) *P*, and (bottom) SWE at 1-month lead over all seasons. (right) The cumulative distribution of SAC differences for all seasons, subbasins and models for Stage 1 (blue) and Stage 2 (red). Subbasins outlined in bold black are 95% significant according to a bootstrap analysis.

unrepresentative of the skill of SWE forecasts over most months. Therefore, we perform a masking of SWE forecasts in two steps adapted from a similar gridded methodology in Zeng et al. (2018). First, we define a snowy season as the full season in which 3-month periods of the climatology have greater than 1.0 mm SWE averaged across a subbasin. Then we only analyze subbasins with at least three consecutive overlapping 3-month periods in the snowy season.

f. Significance testing

Here we are testing which steps in the postprocessing improve the performance of the forecast. Thus, the null hypothesis is a one-sided test (significance level $\alpha = 0.05$) that there is no improvement in the forecast going from the raw forecast to Stage 1, from Stage 1 to Stage 2, and from the individual postprocessed forecasts to the merged SENS. We also test whether the SENS forecast outperforms persistence and trend in the further discussion in section 4. Significance results for the superensemble merging are presented relative to significance over one or both models; however, only those results where SENS significantly outperforms both models are considered a significant improvement for the merging step.

Significance is assessed by bootstrapping (resampling with replacement) the validation years for all forecasts and observations 500 times. Forecasts are paired and the corresponding verification metrics of one stage are subtracted from the next stage in the forecast for each resample. If the percentile value of zero difference in performance metric is less than 0.05 for a given stage, then that stage in postprocessing rejects the null hypothesis and can be considered significant since over 95% of resamples show improvement. Bootstrapping has been chosen as it can be applied consistently across all metrics used in the study and does not assume a distribution. This method is analogous to the resampling methodology used in Hamill (1999) where here only the years are reshuffled. The significance testing is performed for all three validation strategies as

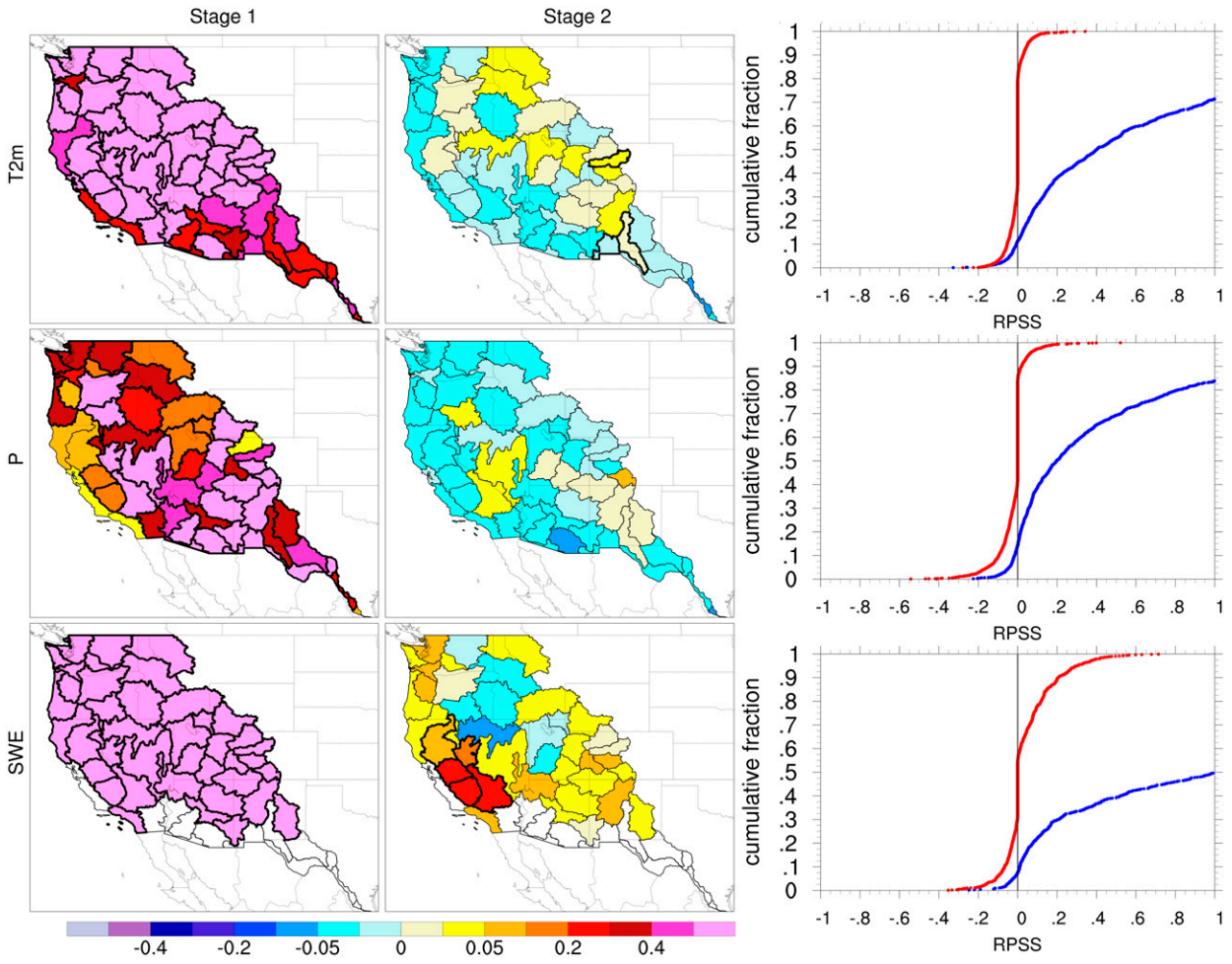

FIG. 2. As in Fig. 1, but for RPSS.

the small 11-yr sample of the L11 validation will have a lot of overlap in the evaluated samples.

3. Results

a. *The impact of the two-stage bias correction and ensemble merging method*

The forecast skill as measured by median SAC and RPSS over all seasons and subbasins at 1-month lead time for the original forecast (raw) and Stages 1 and 2 in each model forecast (CFSv2, SEAS5) and the combined SENS forecast is summarized in Table 1, focusing here only on the L11 results. RPSS steadily increases from raw to Stage 2 for SWE forecasts. T2m and *P* forecasts improve from raw to Stage 1, with a slight improvement (for T2m) or degradation (for *P*) from Stage 1 to Stage 2. The Stage 1 and Stage 2 SWE forecasts in SENS are improved over those from either of the individual models, while the SENS results for T2m and *P* are comparable to the better performing model (SEAS5). Results are similar for median MSE SS in Table S2.

Table 1 also shows that SAC steadily increases from 0.15 and 0.34 in the raw CFSv2 and SEAS5 SWE forecasts, respectively, to 0.31 and 0.47 in Stage 2. However, Stage 1 may increase or decrease the SAC for T2m and *P*, with a slight degradation from Stage 1 to Stage 2. While SENS after the Stage 1 correction shows the best skill for T2m, CFSv2 after Stage 1 correction has the best skill for *P*.

With regard to significance, all variables show significant improvement in RPSS for Stage 1 adjustments, and some of the improvements in RPSS seen for SWE for Stage 2 are also significant. For RPSS, the superensemble merging is generally only significantly better than CFSv2 but not SEAS5 for the L11 validation. SAC metrics only show significant improvement for SWE during Stage 1 or 2, with most significance due to the Stage 2 adjustment. The superensemble merging shows significant improvements in SAC over CFSv2 for SWE, but not SEAS5.

To understand why the teleconnection adjustment (Stage 2) fails to improve forecasts of most notably *P* and to a certain extent T2m, Fig. S2 shows the percentage of time when a Stage 2 adjustment was actually made. More often than not, zero

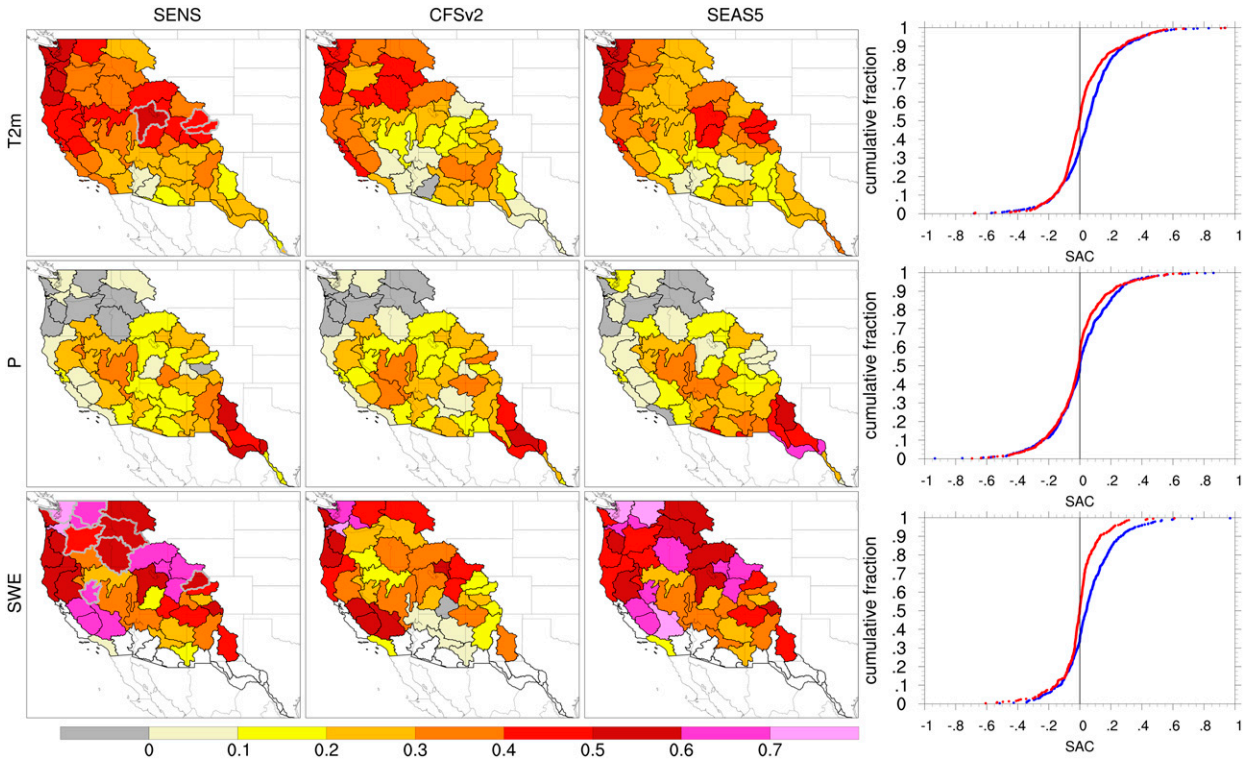

FIG. 3. Median SAC across all seasons for all HUC 4-digit subbasins for (top) T2m, (middle) *P*, and (bottom) SWE at 1-month lead for Stage 2 forecasts of (first column) SENS, (second column) CFSv2, and (third column) SEAS5. (fourth column) The cumulative distribution of SAC differences for all seasons, subbasins and models between SENS and CFSv2 (blue) and SEAS5 (red). Subbasins in SENS outlined in thick gray show significant improvement (at 95% level) over one of the two models and those in bold black show significant improvement over both models.

predictors are selected for T2m and *P*: 80% of the time for all *P* forecasts and 80% of the time for each of the two forecast models and 55% for SENS for T2m (Fig. S2). GMSST_9yr and NINO34 do get used a little more often for SENS T2m, while persistence is most often used (other than no indices getting picked up) for SWE. This suggests that forecast errors are most often not related to global teleconnections in the models.

The spatial variability of the change in SAC going from the raw forecasts through the two stages, where median 1-month lead SAC across all seasons from CFSv2 and SEAS5 is averaged together, is shown in Fig. 1. For T2m and *P* forecasts, little change is made with the Stage 1 correction with mostly positive changes in SAC in the northern part of the domain and mostly negative changes in the south. Small increases or decreases in SAC occur in the Stage 2 correction of the T2m forecasts, whereas most subbasins experience a decrease in SAC in Stage 2 for the *P* forecasts. SWE forecasts are more influenced by the Stage 1 correction, especially with strong increases along the West Coast and Great Basin. A stronger effect of the Stage 2 correction of SWE is garnered in California and the Great Basin. The cumulative distribution of SAC differences after Stages 1 and 2 are also plotted. They show that for Stage 1, the percentage of validation points showing improved forecasts is similar to the percentage of points decreasing. Stage 2 shows fewer points improving versus a degradation of skill for both T2m (19%–27%) and *P* (15%–36%). On the other hand, SWE shows 44% of points improving and 25% degrading.

The Stage 1 correction has a much stronger impact (with most subbasins showing significant improvement) on RPSS for all variables across the whole region (Fig. 2). While RPSS increases overall for SWE prediction from Stage 1 to Stage 2, there is a mixture of positive and negative changes to RPSS in the Stage 2 correction of T2m and *P* forecasts. The Stage 1 improvements are further highlighted in the cumulative distribution where 80%–90% of points show improvement in RPSS after Stage 1. For Stage 2, only SWE has more points showing improved forecasts (41%–25%). Less increases than decreases in RPSS after Stage 2 are shown for T2m (14%–27%) and *P* (12%–36%) forecasts.

b. Which variable has the most skill?

The higher SACs in SENS SWE in Table 1 suggests that SWE might be more predictable than T2m and *P* across the western United States. Here, we explore the spatial variations of median skill across all seasons at a lead of 1 month for the three forecasted variables. Figure 3 shows how the temporal SAC varies spatially for each variable in SENS and the two model ensembles after the two-stage adjustment. The number

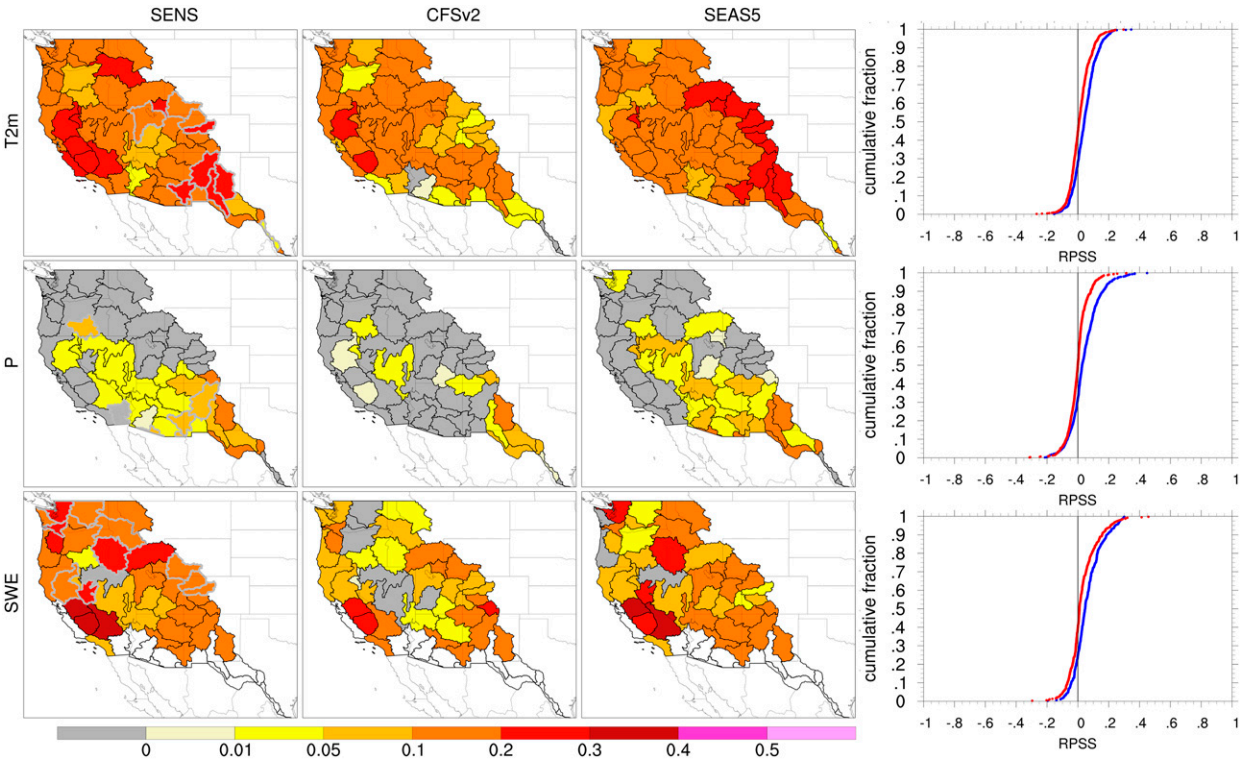

FIG. 4. As in Fig. 3, but for RPSS.

of subbasins with a SAC of at least 0.2 is higher in SENS forecasts of T2m than in either CFSv2 or SEAS5. Over the southern portion of the domain, *P* has higher SACs, whereas T2m and SWE have more skill over the Northwest, West Coast, and Rocky Mountains. For T2m there are some differences in the spatial variation in skill between the two models, while *P* and SWE SACs show similar spatial patterns across both models with higher SACs overall in SEAS5. Both models' skill patterns are combined in SENS. Thus, SENS is typically close to or surpasses the peak skill of the individual model forecasts. However, no individual subbasin has significant skill improvement over both models. In general, SWE forecasts have the highest SACs (0.05–0.78), whereas *P* SACs are the lowest (from −0.17 to 0.58) with T2m SACs in between (0.06–0.59). The cumulative distributions show that SENS more frequently increases SAC over CFSv2 for each variable. However, compared to SEAS5, SENS has lower SAC for *P* 50% of the time versus higher SAC only 37% of the time. For T2m and SWE, the number of points showing increased versus decreased SAC is similar. Thus, the ensemble merging has a SAC that is typically near that of the better performing of the two individual models.

Figure 4 shows the spatial variability of RPSS at 1-month lead. There is broad probabilistic skill for T2m and SWE across most subbasins with only a slight drop off for SWE in areas with limited snowfall in the southern part of the domain. The spatial distribution in RPSS for T2m is much more uniform compared to the SACs in Fig. 3, especially in SEAS5. Precipitation *P* has skill over the Rio Grande basin and the Great Basin but only marginal or no skill elsewhere. As in Table 1, RPSS benefits more from the ensemble merging for all three variables (Fig. 4) than SAC which only sees broad increases for T2m (Fig. 3). As with SAC, the highest RPSS is produced for SWE forecasts (from −0.02 to 0.39), followed by T2m (0.02–0.30) and *P* (from −0.09 to 0.16). This indicates that, of the quantities looked at here, SWE is the most predictable, followed by T2m, with *P* being the least predictable. The cumulative distribution in Fig. 4 shows how RPSS of SENS changed relative to each model for all forecast locations and seasons. With the exception of SEAS5 *P* forecasts (42% decrease, 38% increase), SENS improves RPSS more often than it decreases after merging the two ensemble forecasts. Note that RPSS (in Fig. 4) improves more frequently than SAC (in Fig. 3). One possible reason is that SENS combines the distribution of each model forecast, leading to enhanced RPSS, while the computation of SAC is based on the ensemble mean (with minimal impacts of the distributions).

c. *How does forecast skill change during the annual cycle?*

The results above discussed median skill at 1-month lead time over all seasons. Here, we explore the dependence of skill on season and lead month. Figure 5 shows the anomaly SACs averaged over all subbasins in SENS after the two-stage adjustment as a function of season and lead time. Seasonal forecasts with 0-month lead time have consistently higher

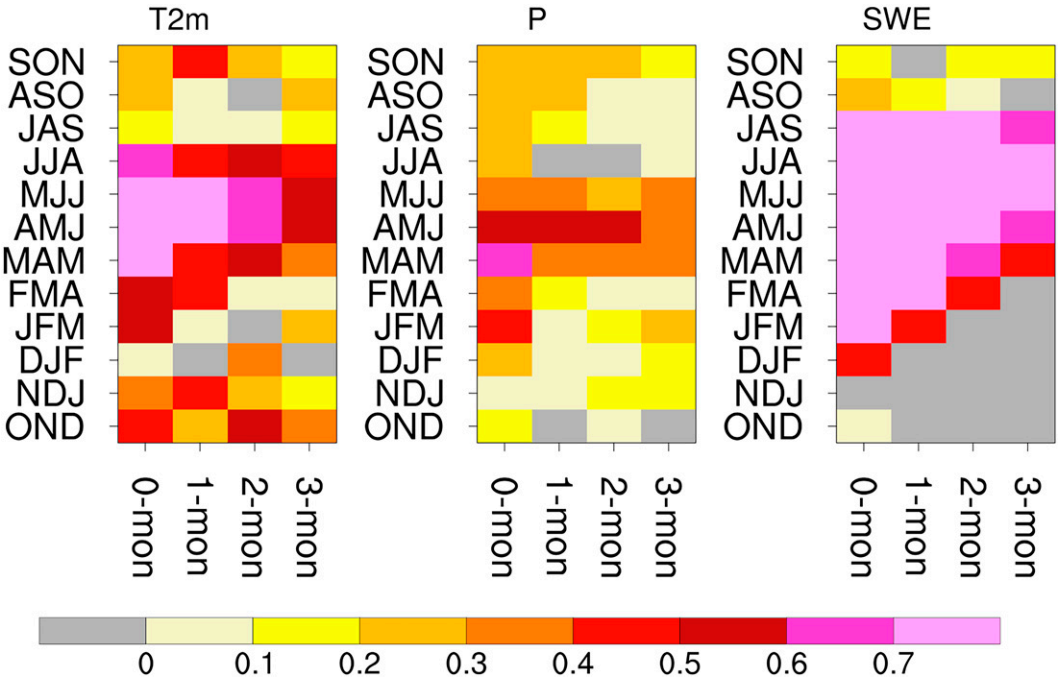

FIG. 5. Median SAC for SENS across all HUC 4-digit subbasins for each season and lead time for (left) T2m, (center) *P*, and (right) SWE. For instance, the value corresponding to "0-mon" and "OND" refers to the median SAC for early October forecasts of OND, while the value corresponding to "1-mon" and "NDJ" refers to the median SAC for early October forecasts of NDJ.

SAC than longer lead times. SAC drops off quickest for *P* and slowest for SWE, suggesting that SWE has longer memory than *P*. Both SWE and T2m have a distinct seasonality in SAC with the highest values occurring from late winter through early summer.

Figure 6 shows the seasonality and lead-time dependence of probabilistic skill represented by RPSS of SENS. There is broad skill for T2m with some drop off in late fall and early winter. A similar springtime peak in skill is seen for T2m and SWE during the spring snowmelt period (Fig. 6) as is seen for SAC (Fig. 5). T2m seems to hang on to some greater skill in late summer through fall compared to SAC which sees a sharper drop off. Precipitation *P* has the most skill at a 0-month lead, with only marginal skill at longer leads except in early spring.

The spatial variations in the timing of peak SAC at 1-month lead time are shown in Fig. 7. For T2m, the peak SAC is fairly uniform with a peak usually in late spring and early summer (AMJ and MJJ). The SWE SAC peak tends to be later over the Pacific Northwest and earlier further south, roughly corresponding to the timing of peak SWE and subsequent snowmelt. These results are consistent with those in Fig. 5. Precipitation *P* exhibits more spatial variability in the timing of peak SAC. Parts of the interior western United States and the Rocky Mountains have a late spring peak. An interesting divide occurs in the Pacific Northwest due to the Cascade Range, with peak skill in early winter on the windward side of the range and an autumn peak in subbasins leeward of the Cascades. This pattern is more pronounced in SEAS5 than CFSv2 (not shown) and the overall weaker seasonality and skill for precipitation may make this pattern highly dependent on the small sample of years in the validation.

4. Further discussion

a. What role does persistence and trend play in forecast skill?

The above results show that there is quite a bit of variability in skill both spatially and temporally. To understand the reason for this, we compare the SENS forecasts to persistence forecasts in which the observed standardized anomaly of the season prior to forecast initialization to the forecasted season at each lead time is applied. SENS is also compared to the trend forecast which calculates a trend line for each season using the Theil–Sen trend estimation from the training period and applies it to the forecast year. Table 2 shows a summary of the results of SENS compared to persistence and trend for both the MSE-based skill score (SS) and SAC. For SS, SENS is significantly more skilled than both persistence and trend for all three validation strategies. SENS also has higher SAC to observation than both persistence and trend, but only shows significant improvements in the longer L3O validation. Figure 8 shows the spatial pattern of the difference in median SAC between SENS and both the persistence and trend forecasts across all seasons at 1-month lead time. These can be compared to that of the SENS forecasts in Fig. 3. Many of the subbasins with the highest SAC in SENS also have high SAC in the persistence forecast as indicated by the similarity in

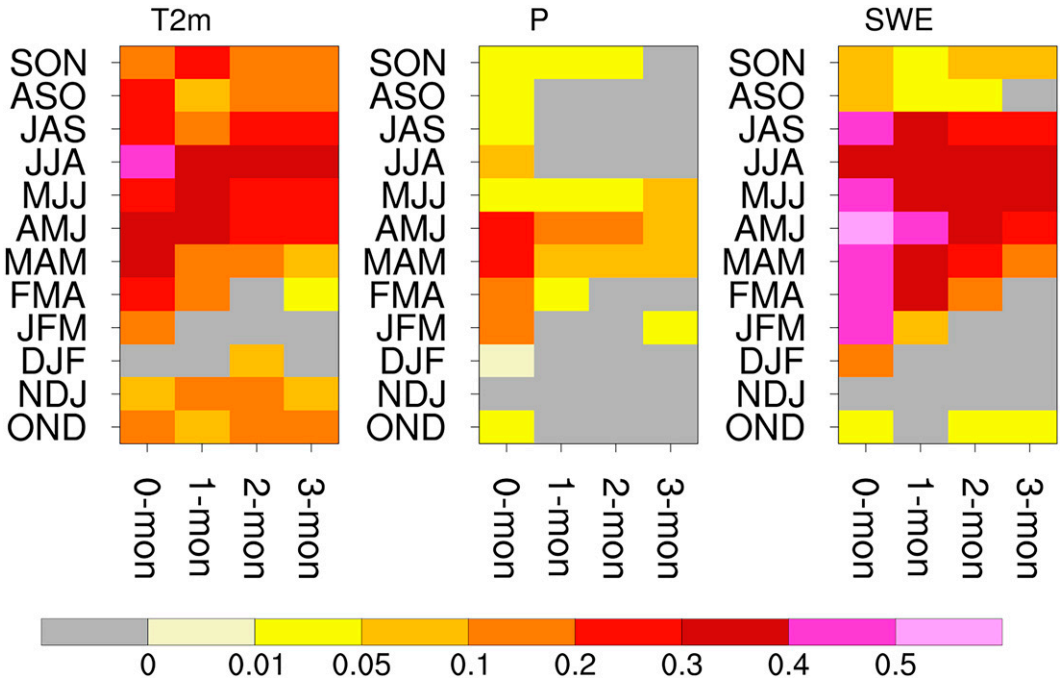

FIG. 6. Median RPSS for SENS across all HUC 4-digit subbasins for each season and lead time for (left) T2m, (center) *P*, and (right) SWE.

probability density functions in the panels in the right column of Fig. 8 and the small number of subbasins in which SENS is statistically significant over persistence. This is most apparent for SWE, but there are also regions where this is the case for T2m and *P* as well. For T2m, there is high persistence skill over the coastal regions consistent with the high skill in SENS. Skill in the interior is less dependent on persistence. The high *P* SACs in the Rio Grande region also correspond to high SACs in the persistence forecasts. However, the higher skill of *P* over California, the Desert Southwest, and the Great Basin are not tied to persistence. Trend forecasts are shown to have poor skill in predicting any variable (Table 2) with SENS being statistically significantly different from Trend in all subbasins (Fig. 8). However, the 11-yr validation period may not be sufficiently long to evaluate the skill of the trend using anomaly SAC. This is discussed further in the next subsection.

The seasonal and lead-time dependence of higher and lower skill in the persistence and trend forecasts are also explored (Fig. S3). Persistence is very similar to the SENS forecasts in Fig. 5 with high skill during the spring snowmelt period and low skill prior to snowpack initiation for SWE,

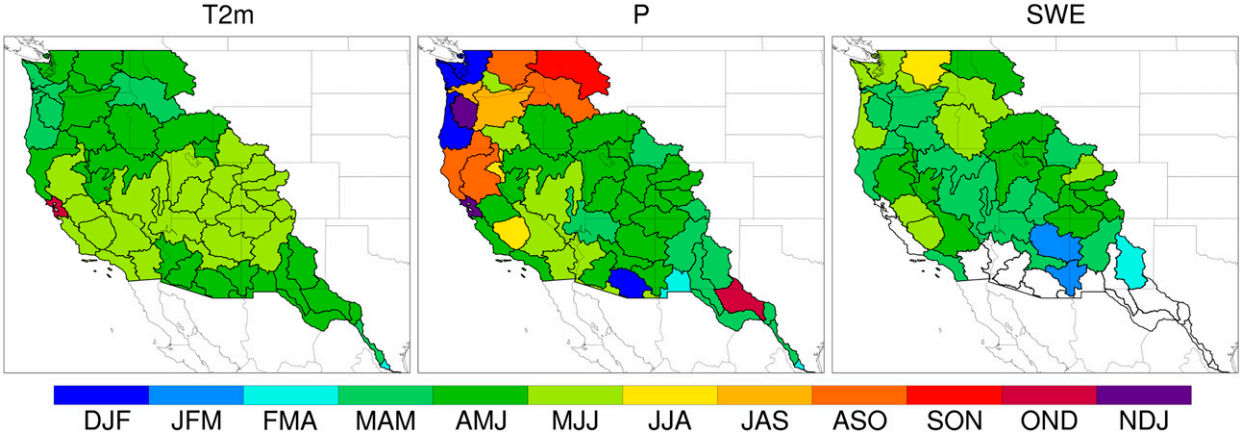

FIG. 7. The 3-month season of maximum SAC for all HUC 4-digit subbasins for (left) T2m, (center) *P*, and (right) SWE at 1-month lead for Stage 2 forecasts of SENS.

TABLE 2. MSE-based skill score (SS) and SAC for seasonal forecasts [SENS, persistence (Pers.), and trend] at 1-month lead; median over all seasons then median across all subbasins. SENS combines CFSv2 and SEAS5. For SENS, metrics that are significant above either Pers. or trend are italicized, and those significant above both are also bolded. The results are split into three different validation strategies: last 11 years using prior data for calibration (L11), first 11 years using future data for calibration (F11) and leave-3-out validation (L3O) for all years (1982–2021).

		L11			F11			L3O		
Metric	Variable	SENS	Pers.	Trend	SENS	Pers.	Trend	SENS	Pers.	Trend
SS	T2m	***0.30***	−0.33	0.14	***0.25***	−0.73	0.11	***0.24***	−0.63	0.08
	P	***0.00***	−1.05	−0.09	***0.02***	−0.81	−0.09	***0.03***	−0.91	−0.05
	SWE	***0.32***	−0.14	0.02	***0.35***	0.11	−0.05	***0.36***	−0.20	−0.03
SAC	T2m	*0.34*	0.21	−0.22	***0.33***	0.04	−0.28	***0.51***	0.15	0.33
	P	*0.17*	0.09	−0.33	*0.23*	0.12	−0.27	***0.22***	0.08	−0.03
	SWE	*0.52*	0.41	−0.25	*0.60*	0.56	−0.21	***0.61***	0.43	0.10

suggesting that much of the skill as well as its seasonal variation is largely due to persistence in the target variables. Trend shows little to no skill when looking at SAC. Only slightly above zero skill for trend is seen in late spring and early summer T2m forecasts. However, trend has higher skill when looking at the MSE-based skill score shown in Fig. S4.

Table 3 presents the pattern correlation between the skill in SENS and that of persistence and trend. This explores how much persistence and trend contribute to patterns of variability in the skill of SENS. Persistence explains more of the variability in SENS's SAC than trend for each variable. Persistence explains 58% of SENS's SWE variability, 21% of its T2m variability, and around 6% of its *P* variability. Trend explains 6%, 3%, and 0% of SENS's variability for T2m, SWE, and *P*, respectively. Around 5% of persistence's SAC is also explained by trend for T2m and SWE. Results show that a higher percentage of spatiotemporal variation of SS is explained by trend than was the case for SAC, especially for T2m where 29% of variance in SS is explained by trend versus 6% for persistence. Precipitation shows largely equal contributions and SWE has similar results for SS than SAC where most skill is explained by persistence. These results suggest that persistence's influence on SENS skill is the dominant contributor for SWE. Trend does have a strong influence on SS for T2m but not as much for SAC which appears to be driven more by persistence.

b. *How robust are the results?*

The alternative validations (F11 and L3O) described in section 2e are analyzed here to discuss the robustness of the F11 validation results presented in section 3. From Table 1, all three validation strategies show that the SENS forecasts generally show an increase in RPSS for SWE and T2m above the raw model forecasts. Similar results are also seen for each step in the postprocessing regardless of strategy, where Stage 1 removes biases but does not improve SAC much. The improved SAC for Stage 2 SWE forecasts is consistent across the validation strategies. All three validation strategies also have small overall decreases in SAC for *P* from Stage 1 to Stage 2. T2m consistently shows minimal difference between Stages 1 and 2. Overall RPSS increases for SWE after Stage 2 for each validation strategy. RPSS only changes by very small amounts for T2m and *P* from Stage 1 to Stage 2. Table 1 also shows that the merging results in small increases in RPSS and SAC for all three validation strategies. Whereas SENS T2m and SWE are only significantly better than one model in the L11 validation, the L3O validation SENS is significantly improved in RPSS over both models for two quantities.

The alternative validation strategies were also looked at to see if we saw the same seasonal progression of skill for each variable that was shown in Figs. 5–7. The seasonal cycle of SWE skill for both SAC and RPSS is consistent for all three validation strategies. However, we find that the seasonal cycle of skill for *P* and T2m seen in the L11 analysis are not consistent across the different validation strategies (e.g., Fig. S5 for SAC).

Furthermore, we explored the skill of persistence and trend across the three different validation strategies. Persistence contributes a high fraction of skill to SENS for all validation strategies, being a bit weaker contribution for F11 and L3O for T2m predictions (not shown). Trend shows mostly negative SAC skill for either variable for both L11 and F11. However, positive skill is present across most seasons in L3O (see Fig. S6). Two reasons may contribute to this. First, the short 11-yr validation periods in L11 and F11 may not represent the long-term trend well. Second, small errors in the trend estimation can introduce larger errors when extrapolating the trend as a forecast, whereas the trend is interpolated across each validation year in the L3O forecasts.

5. Conclusions

In this study we tested postprocessing of two state-of-the-art coupled seasonal prediction models (NCEP CFSv2 and ECMWF SEAS5), empirical ocean–atmosphere climate information and quality observations to create ensemble forecasts of 3-month seasonal T2m, *P*, and SWE out to a lead time of 3 months (e.g., early January prediction for AMJ). Each model was postprocessed in a two-stage approach where biases in the mean, variance and spread are removed first, and then systematic errors related to well-established empirical climate relationships are calculated to remove additional errors. The adjusted models are then combined using prior cross-validated skill as weighting to create the final superensemble (SENS) forecasts. Training for these forecasts used hindcasts beginning in

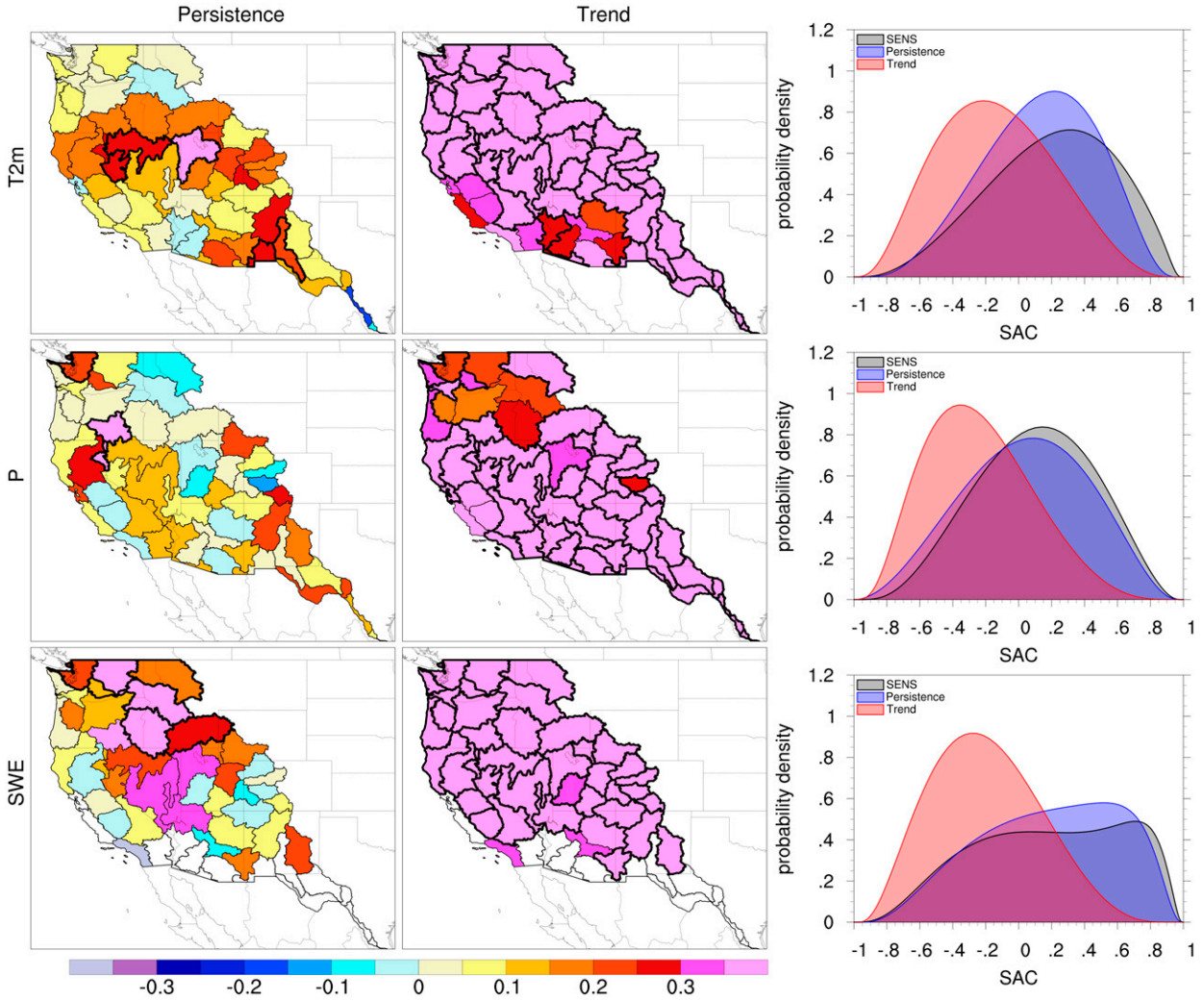

FIG. 8. Median SAC of Stage 2, 1-month lead forecasts of SENS across all seasons for all HUC 4-digit subbasins for (top) T2m, (middle) *P*, and (bottom) SWE minus the SACs of (left) persistence and (center) trend. (right) The probability density function of SAC for each forecast [SENS (gray), persistence (blue), trend (red)] across all seasons and subbasins. Subbasins outlined in bold black show significant skill (at 95% level) of SENS over either persistence or trend.

1982 and were validated in a simulated operational forecast mode from 2011 to 2021.

In assessing the skill of the forecasts of the individual models and SENS, we find the following:

TABLE 3. Pattern correlation between seasonal forecasts [SENS, persistence (Pers.), and trend] across all basins and seasons for skill score (SS) and SAC at 1-month lead for the last 11 years using prior data for calibration (L11).

Metric	Variable	SENS–Pers.	SENS–trend	Pers.–trend
SS	T2m	0.25	0.54	0.12
	P	0.37	0.32	0.17
	SWE	0.71	0.04	0.06
SAC	T2m	0.46	0.24	0.22
	P	0.24	−0.02	0.04
	SWE	0.76	0.18	0.21

1) After applying the two-stage adjustment and combining the adjusted forecasts, the final forecasts (SENS) showed a consistent increase in probabilistic skill for all variables. However, deterministic skill through SAC only shows consistent improvements for SWE.
2) Forecast skill also changes with each step of the postprocessing depending on the variable being predicted. The effect of the bias removal (Stage 1) is more pronounced probabilistically (on RPSS) with significant improvement for all variables. Varying regional changes in skill occur in the Stage 2 correction incorporating the teleconnection indices with the strongest increases occurring for SWE in California and the Great Basin.
3) When strong consistent relationships exist between predictor indices and the target variable as in the case for SWE, we find that Stage 2 has great utility. For SWE, poor initialization in the models (Broxton et al. 2017) has

been effectively adjusted through using anomaly persistence as a predictor index. When these strong relationships do not exist such as for T2m and *P*, then Stage 2 does not benefit the forecast. For these quantities, Stage 1 only is needed prior to merging the models.

4) Overall, prediction of seasonal SWE in SENS was found to be most skillful relative to climatology, followed by T2m, with *P* being the least predictable. Spatial variations in predictability show that SWE is most skillful in snow dominant regions, T2m is broadly skillful, and *P* is most skillful across the southern tier, especially the Rio Grande region.
5) A strong seasonal cycle in prediction skill was found for SWE. In general, peak skill is seen from late winter through early summer with weak skill in the fall. Persistence was shown to be a big contributor to skill and this seasonal skill variation. However, other signals of the seasonal cycle of skill for T2m and *P* were not found to be robust across three separate validation strategies.

The current study is limited by the 40-yr length of the hindcast and 11-yr length of the primary validation used (L11). Such a limit on the length of hindcasts and of the validation period would prohibit small changes in the forecasts from having any statistical significance (Barnston and Tippett 2017). To test the robustness of our validation strategy on future forecasts (L11), we also perform validation on previous forecasts instead (F11) and by leaving three years out (L3O). Similar skill is produced using these two additional validation strategies. Also, trend has little effect on SAC in any of the validation strategies, emphasizing the importance of persistence in the SACs, especially for SWE. However, trend is still related to overall MSE-based skill score.

Furthermore, longer-term relationships between ocean–atmosphere processes and model errors may be missed by the number of years looked at here. The teleconnection relationships to model errors here were assumed to be linear. However, it is possible that some of these teleconnection indices may yield stronger nonlinear relationships and interactions. Future work should explore the possibility of using these nonlinear relationships and further investigate the length of hindcast, and ensemble size needed for optimal adjustments.

Here we only considered ocean and atmospheric predictors from the most recent season for our Stage 2 adjustments, but recent work has also shown that considering the evolution of antecedent conditions can better predict seasonal climate than only using the most recent season (Nigam and Sengupta 2021; Switanek et al. 2020; Rieger et al. 2021). Such work would involve adjustments being made on multiple lags for predictors. Finally, we only considered overall a priori performance for combining multiple models, but some models may perform better than others given specific climate conditions. Therefore, adjusting the weights based upon specific forecasts of opportunity (Mariotti et al. 2020) may improve forecasts.

Acknowledgments. The authors thank Duane E. Waliser at the Jet Propulsion Laboratory (JPL) and Nick Dawson at Idaho Power Corporation for their helpful feedback. One anonymous reviewer is thanked for insightful comments and suggestions that helped us clarify and improve several aspects of our study. This work was supported by the California Department of Water Resources [through subcontracts from JPL and the Center for Western Weather and Water Extremes (CW3E)]. It was initiated under prior support from Idaho Power Corporation.

Data availability statement. The NCEP CFSv2 data are available from the National Centers for Environmental Information at https://www.ncei.noaa.gov/products/weather-climate-models/climate-forecast-system. The ECMWF SEAS5 data are available from the Climate Data Store at https://cds.climate.copernicus.eu/cdsapp#!/dataset/seasonal-monthly-single-levels. PRISM monthly precipitation and temperature data are available from the PRISM Climate Group at https://prism.oregonstate.edu/. The UA SWE product is available from the National Snow and Ice Data Center at https://nsidc.org/data/nsidc-0719. Most teleconnections indices were downloaded from the NOAA Physical Sciences Laboratory at https://psl.noaa.gov/data/climateindices/; please see Table S1 for more information. The USGS Watershed Boundary Dataset is available from the USGS at https://www.usgs.gov/national-hydrography/watershed-boundary-dataset.

REFERENCES

Baker, S. A., A. W. Wood, and B. Rajagopalan, 2020: Application of postprocessing to watershed-scale subseasonal climate forecasts over the contiguous United States. *J. Hydrometeor.*, **21**, 971–987, https://doi.org/10.1175/JHM-D-19-0155.1.

Barnett, T. P., 1981: Statistical prediction of North American air temperatures from Pacific predictors. *Mon. Wea. Rev.*, **109**, 1021–1041, https://doi.org/10.1175/1520-0493(1981)109<1021:SPONAA>2.0.CO;2.

Barnston, A. G., and H. M. van den Dool, 1993: A degeneracy in cross-validated skill in regression-based forecasts. *J. Climate*, **6**, 963–977, https://doi.org/10.1175/1520-0442(1993)006<0963:ADICVS>2.0.CO;2.

——, and T. M. Smith, 1996: Specification and prediction of global surface temperature and precipitation from global SST using CCA. *J. Climate*, **9**, 2660–2697, https://doi.org/10.1175/1520-0442(1996)009<2660:SAPOGS>2.0.CO;2.

——, and M. K. Tippett, 2017: Do statistical pattern corrections improve seasonal climate predictions in the North American multimodel ensemble models? *J. Climate*, **30**, 8335–8355, https://doi.org/10.1175/JCLI-D-17-0054.1.

——, ——, M. Ranganathan, and M. L. L'Heureux, 2019: Deterministic skill of ENSO predictions from the North American multimodel ensemble. *Climate Dyn.*, **53**, 7215–7234, https://doi.org/10.1007/s00382-017-3603-3.

Broxton, P. D., N. Dawson, and X. Zeng, 2016: Linking snowfall and snow accumulation to generate spatial maps of SWE and snow depth. *Earth Space Sci.*, **3**, 246–256, https://doi.org/10.1002/2016EA000174.

——, X. Zeng, and N. Dawson, 2017: The impact of a low bias in snow water equivalent initialization on CFS seasonal forecasts. *J. Climate*, **30**, 8657–8671, https://doi.org/10.1175/JCLI-D-17-0072.1.

Clow, D. W., 2010: Changes in the timing of snowmelt and streamflow in Colorado: A response to recent warming. *J. Climate*, **23**, 2293–2306, https://doi.org/10.1175/2009JCLI2951.1.

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

DelSole, T., and J. Shukla, 2009: Artificial skill due to predictor screening. *J. Climate*, **22**, 331–345, https://doi.org/10.1175/2008JCLI2414.1.

Diro, G. T., and H. Lin, 2020: Subseasonal forecast skill of snow water equivalent and its link with temperature in selected SubX models. *Wea. Forecasting*, **35**, 273–284, https://doi.org/10.1175/WAF-D-19-0074.1.

Gao, X., and S. Mathur, 2021: Predictability of U.S. regional extreme precipitation occurrence based on large-scale meteorological patterns (LSMPs). *J. Climate*, **34**, 7181–7198, https://doi.org/10.1175/JCLI-D-21-0137.1.

Gershunov, A., and D. R. Cayan, 2003: Heavy daily precipitation frequency over the contiguous United States: Sources of climatic variability and seasonal predictability. *J. Climate*, **16**, 2752–2765, https://doi.org/10.1175/1520-0442(2003)016<2752:HDPFOT>2.0.CO;2.

Gibson, P. B., W. E. Chapman, A. Altinok, L. Delle Monache, M. J. DeFlorio, and D. E. Waliser, 2021: Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. *Commun. Earth Environ.*, **2**, 159, https://doi.org/10.1038/s43247-021-00225-4.

Gneiting, T., A. E. Raftery, A. H. Westveld III, and T. Goldman, 2005: Calibrated probabilistic forecasting using ensemble model output statistics and minimum CRPS estimation. *Mon. Wea. Rev.*, **133**, 1098–1118, https://doi.org/10.1175/MWR2904.1.

Hamill, T. M., 1999: Hypothesis tests for evaluating numerical precipitation forecasts. *Wea. Forecasting*, **14**, 155–167, https://doi.org/10.1175/1520-0434(1999)014<0155:HTFENP>2.0.CO;2.

Johnson, S. J., and Coauthors, 2019: SEAS5: The new ECMWF seasonal forecast system. *Geosci. Model Dev.*, **12**, 1087–1117, https://doi.org/10.5194/gmd-12-1087-2019.

Kim, H., F. Vitart, and D. E. Waliser, 2018: Prediction of the Madden–Julian oscillation: A review. *J. Climate*, **31**, 9425–9443, https://doi.org/10.1175/JCLI-D-18-0210.1.

Livezey, R. E., and A. G. Barnston, 1988: An operational multifield analog/antianalog prediction system for United States seasonal temperatures 1. System design and winter experiments. *J. Geophys. Res.*, **93**, 10 953–10 974, https://doi.org/10.1029/JD093iD09p10953.

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., and Coauthors, 2020: Windows of opportunity for skillful forecasts subseasonal to seasonal and beyond. *Bull. Amer. Meteor. Soc.*, **101**, E608–E625, https://doi.org/10.1175/BAMS-D-18-0326.1.

Murphy, A. H., 1973: A new vector partition of the probability score. *J. Appl. Meteor.*, **12**, 595–600, https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2.

Nigam, S., and A. Sengupta, 2021: The full extent of El Niño's precipitation influence on the United States and the Americas: The suboptimality of the Niño 3.4 SST index. *Geophys. Res. Lett.*, **48**, e2020GL091447, https://doi.org/10.1029/2020GL091447.

Pathak, P., A. Kalra, K. W. Lamb, W. P. Miller, S. Ahmad, R. Amerineni, and D. P. Ponugoti, 2018: Climatic variability of the Pacific and Atlantic oceans and western US snowpack. *Int. J. Climatol.*, **38**, 1257–1269, https://doi.org/10.1002/joc.5241.

Raftery, A. E., T. Gneiting, F. Balabdaoui, and M. Polakowski, 2005: Using bayesian model averaging to calibrate forecast ensembles. *Mon. Wea. Rev.*, **133**, 1155–1174, https://doi.org/10.1175/MWR2906.1.

Rieger, N., Á. Corral, E. Olmedo, and A. Turiel, 2021: Lagged teleconnections of climate variables identified via complex rotated maximum covariance analysis. *J. Climate*, **34**, 9861–9878, https://doi.org/10.1175/JCLI-D-21-0244.1.

Risbey, J. S., and Coauthors, 2021: Standard assessments of climate forecast skill can be misleading. *Nat. Commun.*, **12**, 4346, https://doi.org/10.1038/s41467-021-23771-z.

Robertson, A. W., U. Lall, S. E. Zebiak, and L. Goddard, 2004: Improved combination of multiple atmospheric GCM ensembles for seasonal prediction. *Mon. Wea. Rev.*, **132**, 2732–2744, https://doi.org/10.1175/MWR2818.1.

Ropelewski, C. F., and M. S. Halpert, 1986: North American precipitation and temperature patterns associated with the El Niño/Southern Oscillation (ENSO). *Mon. Wea. Rev.*, **114**, 2352–2362, https://doi.org/10.1175/1520-0493(1986)114<2352:NAPATP>2.0.CO;2.

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.

Shao, Y., Q. J. Wang, A. Schepen, D. Ryu, and F. Pappenberger, 2022: Improved trend-aware postprocessing of GCM seasonal precipitation forecasts. *J. Hydrometeor.*, **23**, 25–37, https://doi.org/10.1175/JHM-D-21-0099.1.

Stevens, A., R. Willett, A. Mamalakis, E. Foufoula-Georgiou, A. Tejedor, J. T. Randerson, P. Smyth, and S. Wright, 2021: Graph-guided regularized regression of Pacific Ocean climate variables to increase predictive skill of southwestern U.S. winter precipitation. *J. Climate*, **34**, 737–754, https://doi.org/10.1175/JCLI-D-20-0079.1.

Stockdale, T. N., and Coauthors, 2022: Prediction of the quasi-biennial oscillation with a multi-model ensemble of QBO-resolving models. *Quart. J. Roy. Meteor. Soc.*, **148**, 1519–1540, https://doi.org/10.1002/qj.3919.

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.

Switanek, M. B., J. J. Barsugli, M. Scheuerer, and T. M. Hamill, 2020: Present and past sea surface temperatures: A recipe for better seasonal climate forecasts. *Wea. Forecasting*, **35**, 1221–1234, https://doi.org/10.1175/WAF-D-19-0241.1.

Thakur, B., A. Kalra, V. Lakshmi, K. W. Lamb, W. P. Miller, and G. Tootle, 2020: Linkage between ENSO phases and western US snow water equivalent. *Atmos. Res.*, **236**, 104827, https://doi.org/10.1016/j.atmosres.2019.104827.

Thrasher, B., E. P. Maurer, C. McKellar, and P. B. Duffy, 2012: Technical note: Bias correcting climate model simulated daily temperature extremes with quantile mapping. *Hydrol. Earth Syst. Sci.*, **16**, 3309–3314, https://doi.org/10.5194/hess-16-3309-2012.

Vigaud, N., A. W. Robertson, and M. K. Tippett, 2017: Multimodel ensembling of subseasonal precipitation forecasts over North America. *Mon. Wea. Rev.*, **145**, 3913–3928, https://doi.org/10.1175/MWR-D-17-0092.1.

Vitart, F., 2017: Madden–Julian Oscillation prediction and teleconnections in the S2S database. *Quart. J. Roy. Meteor. Soc.*, **143**, 2210–2220, https://doi.org/10.1002/qj.3079.

Wang, G., Y. Zhuang, R. Fu, S. Zhao, and H. Wang, 2021: Improving seasonal prediction of California winter precipitation using canonical correlation analysis. *J. Geophys. Res. Atmos.*, **126**, e2021JD034848, https://doi.org/10.1029/2021JD034848.

Wood, A. W., L. R. Leung, V. Sridhar, and D. P. Lettenmaier, 2004: Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. *Climatic Change*, **62**, 189–216, https://doi.org/10.1023/B:CLIM.0000013685.99609.9e.

Xu, L., N. Chen, X. Zhang, and Z. Chen, 2020: A data-driven multi-model ensemble for deterministic and probabilistic precipitation forecasting at seasonal scale. *Climate Dyn.*, **54**, 3355–3374, https://doi.org/10.1007/s00382-020-05173-x.

Zeng, X., P. Broxton, and N. Dawson, 2018: Snowpack change from 1982 to 2016 over conterminous United States. *Geophys. Res. Lett.*, **45**, 12 940–12 947, https://doi.org/10.1029/2018GL079621.

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.