Evaluation of the Subseasonal Forecast Skill of Floods Associated with Atmospheric Rivers in Coastal Western U.S. Watersheds

QIAN CAO,[a] SHRADDHANAND SHUKLA,[b] MICHAEL J. DEFLORIO,[c] F. MARTIN RALPH,[c] AND DENNIS P. LETTENMAIER[a]

[a] *Department of Geography, University of California, Los Angeles, Los Angeles, California*
[b] *University of California, Santa Barbara, Santa Barbara, California*
[c] *Center for Western Weather and Water Extremes, Scripps Institution of Oceanography, University of California, San Diego, La Jolla, California*

(Manuscript received 11 September 2020, in final form 16 January 2021)

ABSTRACT: Atmospheric rivers (ARs) are responsible for up to 90% of major flood events along the U.S. West Coast. The time scale of subseasonal forecasting (from 2 weeks to 1 month) is a critical lead time for proactive mitigation of flood disasters. The NOAA Climate Testbed Subseasonal Experiment (SubX) is a research-to-operations project with almost immediate availability of forecasts. It has produced a reforecast database that facilitates evaluation of flood forecasts at these subseasonal lead times. Here, we examine the SubX-driven forecast skill of AR-related flooding out to 4-week lead using the Distributed Hydrology Soil Vegetation Model (DHSVM), with particular attention to the role of antecedent soil moisture (ASM), which modulates the relationship between meteorological and hydrological forecast skill. We study three watersheds along a transect of the U.S. West Coast: the Chehalis River basin in Washington, the Russian River basin in Northern California, and the Santa Margarita River basin in Southern California. We find that the SubX-driven flood forecast skill drops quickly after week 1, during which there is relatively high deterministic forecast skill. We find some probabilistic forecast skill relative to climatology as well as ensemble streamflow prediction (ESP) in week 2, but minimal skill in weeks 3–4, especially for annual maximum floods, notwithstanding some probabilistic skill for smaller floods in week 3. Using ESP and reverse-ESP experiments to consider the relative influence of ASM and SubX reforecast skill, we find that ASM dominates probabilistic forecast skill only for small flood events at week 1, while SubX reforecast skill dominates for large flood events at all lead times.

KEYWORDS: Soil moisture; Forecast verification/skill; Flood events; Atmospheric river

1. Introduction

The subseasonal forecasting time scale (from 2 weeks to 1 month lead) is a critical lead time window for proactive disaster mitigation efforts such as reservoir operations for flood control. However, past research on dynamical forecast skill of flooding events has not focused on subseasonal lead times until recently (e.g., Vitart et al. 2017). During the past several years, a joint effort between the weather and climate communities has been made to bridge the weather–climate prediction gap at the Subseasonal-to-Seasonal (S2S) range (Mariotti et al. 2018; Merryfield et al. 2020), which typically is defined by lead times ranging from 2 weeks to 2 (or 3) months, but in this study we only focus on lead times from 1 to 4 weeks. Several forecast databases have been developed, such as the World Weather Research Programme (WWRP)/World Climate Research Program (WCRP) S2S Prediction Project (Vitart et al. 2017) and the NOAA/Climate Testbed Subseasonal Experiment (SubX) project (Pegion et al. 2019). The former consists of 11 models, while the latter consists of 7 models and focuses on operational subseasonal forecasts. The two databases have two models in common: the National Centers for Environmental Prediction (NCEP) model and the Environment and Canada Climate Change (ECCC) model. Both databases have been increasingly used by the S2S research community for a variety of applications (e.g., DeFlorio et al. 2019b; Gibson et al. 2020).

Atmospheric rivers (ARs) are responsible for most of the storm events leading to extreme precipitation and runoff along the coastal western United States (e.g., Ralph et al. 2006, 2019; Dettinger et al. 2011; Neiman et al. 2011; Barth et al. 2017; Konrad and Dettinger 2017). Forecasts of opportunity, where predictions of midlatitude extremes at S2S lead times over North America are more skillful than normal, can be identified by examining subsets of predictions made during certain active phases of large-scale climate variability, such as the Madden–Julian oscillation (MJO) and El Niño–Southern Oscillation (ENSO) (Mariotti et al. 2020; Merryfield et al. 2020). The effects of these mechanisms on AR forecast skill have been demonstrated along the U.S. West Coast (e.g., Baggett et al. 2017; DeFlorio et al. 2018, 2019a,b; Mundhenk et al. 2018; Nardi et al. 2018). For example, DeFlorio et al. (2019b) evaluated the S2S hindcast skill of ARs out to 4-week lead over the western United States using three models from the WWRP/WCRP S2S database, including the European Centre for Medium-Range Weather Forecasts (ECMWF)

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-20-0219.s1.

Corresponding author: Dennis P. Lettenmaier, dlettenm@ucla.edu

DOI: 10.1175/JHM-D-20-0219.1

model, and the NCEP and ECCC models. They found that the ECMWF and ECCC models showed larger magnitude and spatial extent of forecast skill at week 3 in comparison with NCEP, but there was generally very little skill at week 4 in all three models. In addition, the active MJO phases showed stronger modulation of hindcast skill than ENSO phases (DeFlorio et al. 2019b). To date, however, an evaluation of the hindcast skill of subseasonal AR-related flooding has not been made.

Several recent studies have evaluated hindcasts of precipitation and their potential application in streamflow forecasts from the WWRP/WCRP S2S database. For example, Lin et al. (2018) examined the forecast skill of 11 S2S models for extreme precipitation at a lead time of about 2 weeks for a 2017 flood event in eastern Canada. They found that most of the models predicted above-normal precipitation during the flood event but most underestimated precipitation amounts in comparison with observations, possibly due to the underestimation of the amplitudes of the MJO teleconnections in boreal winter and due to inherent biases in model climatology. Pan et al. (2019) evaluated S2S precipitation prediction skill over the western United States. They found that the best-performing models had useful deterministic skill at week 2 but beyond that only their probabilistic skill was useful. Li et al. (2019) evaluated the deterministic skill of precipitation reforecasts of eight S2S models and the S2S-driven streamflow forecasts using a hydrologic model in four basins in China, with drainage areas of 3300–52 000 km^2. They found that the S2S models had precipitation prediction skill up to a lead time of 11 days. They also found that bias-correcting streamflow as a postprocessing step could substantially improve the deterministic skill for peak flow prediction. Schick et al. (2019) used a regression method to examine the application of the ECMWF output in predicting monthly average streamflow in 16 European catchments. They found that the prediction skill varied greatly among the predictor combinations, catchments, and dates of prediction. The skill was frequently lower than climatology at a lead time of 20 days.

All the above studies used the WWRP/WCRP S2S database to evaluate the hindcast skill of several relevant variables for water resource management applications. NOAA's SubX differs from the WWRP/WCRP reforecasts by having a research-to-operations focus, and hence includes operational as well as research models and produces forecasts in near real-time (Pegion et al. 2019). Here, we use NOAA's SubX precipitation and temperature subseasonal reforecasts given their almost immediate availability, with specific attention to AR-related storms and flooding along the coastal western United States. Pegion et al. (2019) evaluated the skill of the week-3 averages of the seven SubX models globally. They found greater skill in temperature compared to precipitation forecasts at 3-week lead time. Baker et al. (2019) found similar results in an evaluation of the skill of the NCEP Climate Forecast System version 2 (CFSv2), one of the SubX models, over the contiguous United States (CONUS) domain. Despite the fact that precipitation skill dropped quickly by weeks 2–3, the West Coast showed the highest skill over the CONUS during the winter months (Baker et al. 2019).

The forecast skill of meteorological variables (particularly precipitation) is an important determinant of flood prediction skill; however, antecedent hydrological conditions play an important role as well (e.g., Mahanama et al. 2008). For instance, low antecedent soil moisture (ASM) (as is often the case along the Pacific Coast early in the winter season) has been shown to be an offsetting factor for several extreme historical AR events in California's Russian River basin that otherwise would have led to major flooding (Cao et al. 2019).

Ensemble streamflow prediction (ESP) (Day 1985) and reverse-ESP (revESP) (Wood and Lettenmaier 2008) experiments have been used in previous studies to examine the relative importance of initial hydrological conditions (denoted as "IHCs"; primarily ASM in lowland coastal watersheds) and climate forecast error as sources of streamflow forecast uncertainty at seasonal time scales (e.g., Wood and Lettenmaier 2008; Li et al. 2009; Shukla and Lettenmaier 2011). In ESP, a hydrologic model with assumed perfect IHCs is forced by an ensemble of meteorological forcings resampled from past observations. In contrast in revESP, the model is forced with assumed perfect meteorological forecasts with an ensemble of resampled IHCs. Here we used the ESP/revESP method(s) to partition the relative contributions of ASM and meteorological forecast skill to errors in flood forecasts at subseasonal time scales.

In addition, ESP can be used as a baseline (and arguably is more hydrologically relevant than climatology, often used for evaluation of weather and climate forecasts) to assess the performance of meteorological forecast-driven streamflow forecasts (Wood et al. 2005; Li et al. 2009). For streamflow, the ESP ensembles, unlike climatology from historical observations (which is equivalent to unconditional ESP), benefit from knowledge of IHCs (Wood et al. 2005). For example, Monhart et al. (2019) evaluated subseasonal forecasts from ECMWF (after statistical downscaling) against the ESP approach in three alpine catchments with areas of 80–1700 km^2. They found that the ECMWF forecasts provided added value in streamflow predictions relative to ESP especially at shorter lead times.

Given this background, our motivating questions in this study are:

1) What is the subseasonal forecast skill (at 1–4-week lead times) of AR-related flooding driven by downscaled SubX reforecasts in coastal western U.S. watersheds? Are SubX-based flood forecasts more skillful than traditional ESP?
2) What are the relative influences of ASM and SubX reforecast skill on subseasonal flood forecast skill for coastal western U.S. watersheds?

To answer these questions, we first downscaled the SubX reforecasts to a finer spatial resolution, given their coarse native resolution of $1° \times 1°$ with respect to our study domain, and the high spatial resolution of our hydrological model. We then implemented the Distributed Hydrology Soil Vegetation Model (DHSVM) (Wigmosta et al. 1994) in the three basins and ran the model with the downscaled SubX reforecast forcings. We evaluated both the deterministic and probabilistic skill of SubX-based flood forecasts (all of the "forecasts" in this paper technically are reforecasts; i.e., not real time). We also examined how the relative contribution of

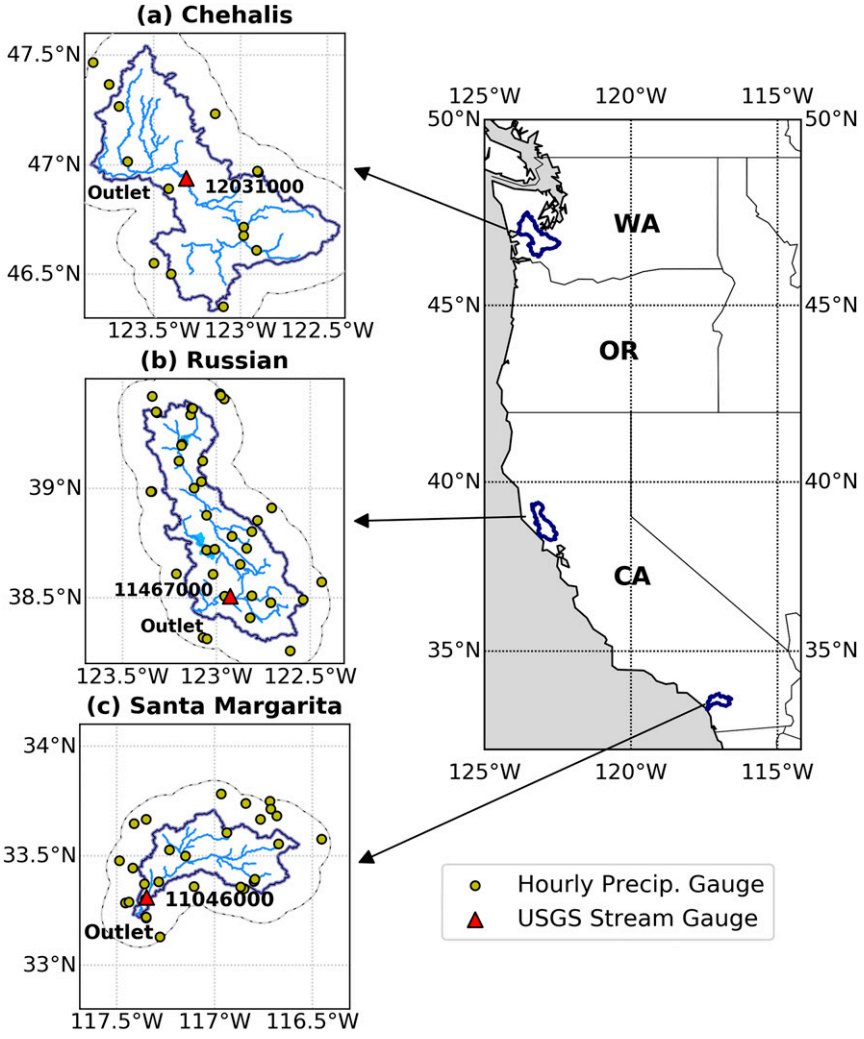

FIG. 1. Map of study region including (a) the Chehalis River basin in Washington State, (b) the Russian River basin in Northern California, and (c) the Santa Margarita River basin in Southern California.

uncertainties in meteorological forcings and in ASM to errors in streamflow reforecasts evolved with lead time by conducting ESP/revESP experiments.

2. Study region

We focused on three watersheds that form a transect along the U.S. Pacific Coast: the Chehalis River basin in Washington State, the Russian River basin in Northern California, and the Santa Margarita River basin in Southern California, with drainage areas of 5400, 3850, and 1870 km^2, respectively (see Fig. 1). These are the same watersheds used in our previous study (Cao et al. 2020), where we examined future climate impacts on the role of ASM in AR-related floods. There are systematic geographic variations in the hydroclimatic conditions in coastal watersheds along the transect within which these three watersheds lie (Cao et al. 2020). Moreover, their geographical locations reflect different AR landfalling signatures. AR landfalls show a marked seasonal progression from the Pacific Northwest in the mid-to-late fall (when they are most frequent) to northern California in early winter (Gershunov et al. 2017). All three basins have moderate topographic variations with elevation ranges of 0–1429, 0–1324, and 143–1736 m in the Chehalis, Russian, and Santa Margarita River basins, respectively. They are all rain-dominated basins, which avoids the added complexity of the influence of snowmelt on streamflow that more commonly occurs in mountainous basins; only the Chehalis has modest contributions of snowmelt to flood runoff. The precipitation in all three basins is strongly winter dominant and varies interannually, with ranges of 1560–2700 mm, 320–1580 mm, and 160–750 mm in the Chehalis, Russian, and Santa Margarita River basins, respectively, during 1999–2016. The precipitation regime is more variable in the Santa Margarita River than the other two basins (Gershunov et al. 2019).

TABLE 1. List of SubX models used in this study. Community column indicates SEAS for seasonal prediction community and NWP for numerical weather prediction community.

Model	Ensemble members	Initialization interval (days)	Forecast period (days)	Community	Reference(s)
ECCC-GEPS5	4	7	32	NWP	Lin et al. (2016)
EMC-GEFS	11	7	35	NWP	Zhou et al. (2016, 2017) and Zhu et al. (2018)
ESRL-FIMr1p1	4	7	32	NWP	Sun et al. (2018a,b)
GMAO-GEOS_V2p1	4	5	45	SEAS	Koster et al. (2000), Molod et al. (2012), Reichle and Liu (2014), and Rienecker et al. (2008)
RSMAS-CCSM4	3	7	45	SEAS	Infanti and Kirtman (2016)
NCEP-CFSv2	4	1	45	SEAS	Saha et al. (2014)

During these years, 79%, 87%, and 83% of precipitation fell between October and March, respectively.

3. Data and methods

a. *SubX reforecasts*

There are seven models in the SubX database, but one of them (NRL-NESM) only has one member, and we therefore omitted it from this study. We used six models in total, with five of them from the SubX database (available at the IRI data library; http://iridl.ldeo.columbia.edu/SOURCES/.Models/.SubX/), plus a sixth (NCEP-CFSv2) from another source (https://www.ncdc.noaa.gov/data-access/model-data/model-datasets/climate-forecast-system-version2-cfsv2) because some of the variables we needed had not been uploaded to the IRI data library. Forecasts of all models are available in real-time. There are 30 ensemble members in total for the six models we used (see Table 1). The reforecasts (retrospective forecasts) cover the period 1999–2016. We only used data during winter months (October–March) when most precipitation and AR-related flood events occur. The initialization interval of the models is at least once a week and the lead time for each model is at least 32 days. The spatial resolution of the SubX output is $1° \times 1°$ and its temporal resolution is daily. We downscaled the SubX output to $1/16° \times 1/16°$ using the gridded observation dataset of Livneh et al. (2013) [extended to 2018 as described in Su et al. (2021)] as the training dataset. The wind speed data in the Livneh dataset is from interpolated NCEP–NCAR Reanalysis (Kalnay et al. 1996).

1) DOWNSCALING METHOD

We used a statistical downscaling method, the bias correction and spatial downscaling (BCSD; Wood et al. 2004). We implemented BCSD at a daily time scale [termed as "daily BCSD" following Abatzoglou and Brown (2012)]. Daily BCSD has been shown to be an effective approach for removing bias (e.g., Monhart et al. 2018; Baker et al. 2019) in atmospheric model output. We applied daily BCSD to precipitation, maximum daily temperature (Tmax), minimum daily temperature (Tmin), and wind speed. Some analyses have indicated that constructed analog-based techniques may outperform daily BCSD in capturing the magnitude of extremes (e.g., Abatzoglou and Brown 2012). Therefore we also implemented a constructed analog-based method, localized constructed analogs (LOCA; Pierce et al. 2014), for the downscaling of precipitation (arguably the most important hydrologic forcing) of one SubX model in order to compare these two methods. The procedures of this method can be found in the online supplemental material (see Text S1). We will only describe BCSD in detail here.

There are two primary steps in our implementation of daily BCSD: 1) spatial (bilinear) interpolation of the $1° \times 1°$ daily SubX model output to $1/16° \times 1/16°$; 2) bias correction of the interpolated forecasts at each 1/16° grid cell using the quantile mapping (QM) method (Wood et al. 2002). The training period we used was 1999–2016. We applied both steps to each of the six SubX models.

When applying the QM, we pooled the reforecast days and observation days with similar climatology (see Fig. S1 for illustration). In so doing, we wanted to preserve the forecast skill with respect to lead times. Hence, we performed the QM in a lead-time-dependent manner similar to Monhart et al. (2018) in their bias correction of ECMWF model output. The general steps in our processing included: 1) for a given forecast initialization date, we first found the initialization falling within a 15-day window centered on the given forecast initialization date over the 18-yr reforecast period; 2) for each lead on the given initialization date, we selected the reforecast days with the same lead time from all ensemble members of the same model; 3) we pooled the climatology (observations) based on a 15-day window centered on reforecast calendar days over 18 years; and 4) we replaced the reforecast value with a value from the observation climatology with the same quantile based on their empirical distributions (based on Weibull plotting positions). Additionally, we examined the effect of populating the sample distribution in the second step by pooling the following 7 days (denoted as "BCSD_7d") versus using 1 day only (denoted as "BCSD_1d") (see Fig. S1).

When the percentile of a reforecast value was outside the range of the empirical percentile of observations, we fit theoretical probability distributions to the data. Following Wood et al. (2002), we used the Gumbel distribution for upper tails and the Weibull distribution for lower tails for precipitation. Similar to precipitation, we applied daily BCSD to Tmax, Tmin, and wind speed. For temperature (Tmax and Tmin), we used a normal distribution for the tails. For wind speed, we used only the empirical distribution, limiting the tails to the range of the empirical distribution.

2) EVALUATION OF SUBX PRECIPITATION AND TEMPERATURE

We evaluated the precipitation and temperature before and after bias correction. Following Pegion et al. (2019), we used the anomaly correlation coefficient (ACC; Wilks 2006), which was computed over time, to assess the SubX precipitation and temperature skill. To obtain anomalies, the climatology was calculated for each calendar day using three steps: 1) calculate daily ensemble means for each model and each forecast run; 2) calculate a multiyear average for each day of the year (1–366); 3) apply a smoothing window of ±15 days (Pegion et al. 2019). The ACC was calculated between each model and corresponding observations and averaged over weeks 1–4. We estimated a critical correlation value of 0.23 for the ACC in a single month and 0.19 for the ACC across all months (October–March), exceeding which the ACC values are statistically different from zero at the 5% significance level using a two-sided t test based on the number of reforecasts we examined (we selected reforecasts with initialization dates closest to certain fixed dates in each month; see section 3b). To evaluate the performance of downscaling methods, we calculated relative biases for precipitation and biases for temperature.

b. Model implementation

We implemented DHSVM in the three basins with essentially the same model setup as in Cao et al. (2020). DHSVM requires meteorological inputs including precipitation, wind speed, air temperature, relative humidity, and downward solar and longwave radiation. To run the model at an hourly time step, we disaggregated the daily data to hourly using the Mountain Microclimate Simulation Model (MTCLIM) algorithms as described and implemented by Bohn et al. (2013) for the last four variables. We took wind speed to be constant throughout each day.

Following Cao et al. (2020), we disaggregated the gridded daily precipitation in each basin to hourly using a regionalized method of fragments (denoted as "MoF") algorithm (Westra et al. 2012), which samples the hourly to daily precipitation ratio from storms with similar magnitudes. We collected hourly precipitation gauge data with at least 5-yr records during 1999–2016 from multiple networks (see Cao et al. 2019) in our study watersheds. There are 15, 42, and 28 hourly precipitation gauges that met our criteria in the Chehalis, Russian, and Santa Margarita basins, respectively (see gauge locations in Fig. 1). We applied the same disaggregation approach to both Livneh et al. (2013) data and, bias-corrected and downscaled SubX reforecasts.

We ran DHSVM using the Livneh et al. (2013) forcings for the period 1999–2016 as a control run. The initialization interval for most SubX models is 7 days (see Fig. S1 for example), but different models are initialized on different days. One primary purpose of the control run is to provide IHCs for forecast simulations using SubX forcings. A multimodel ensemble is usually generated by averaging all forecasts from the same start date, and has been termed a lagged average ensemble (Vitart et al. 2017; Pegion et al. 2019). Following this method, we output model states for fixed dates including the 7th, 14th, 21st, and 28th of each month. This results in 432 IHC dates in total over 18 years. For each IHC date, we identified the latest SubX model initialization date within the previous week. For each SubX ensemble member and each identified initialization, we ran DHSVM for 28 days (4-week forecast). Given that the initialization interval of NCEP-CFSv2 is one day, we examined how its forecast skill evolved with lead days.

c. Assessment of flood forecast skill

1) IDENTIFICATION OF AR-RELATED EXTREME EVENTS

We used the peaks-over-threshold (POT) method (e.g., Lang et al. 1999) to identify extreme discharge events as in Cao et al. (2019, 2020). We first applied the event independence criteria from USWRC (1982) to daily streamflow data in order to identify independent discharge events. We set thresholds at each stream gauge that resulted in 1, 2, and 3 extreme events per year on average, which we denote as POT_{N1}, POT_{N2} and POT_{N3}.

We examined AR contributions to extreme events by identifying the flood events that were coincident with AR events. The AR date catalog we used is based on the ECMWF interim reanalysis (ERA-Interim) dataset, from Guan and Waliser (2015). We extracted the grid cells from the catalog that intersected each basin and identified them as potential AR-related floods.

2) EVALUATION METRICS

(i) Deterministic skill

We evaluated the deterministic skill of NCEP-CFSv2-based flood forecasts because they are initialized every day as opposed to other SubX models that are initialized weekly and on different days of the week. NCEP-CFSv2 based flood forecasts hence facilitate examination of the variation in the flood forecast skill with lead-time more precisely than would be possible with the other SubX models. We used the Kling–Gupta efficiency (KGE) (Gupta et al. 2009) as the evaluation metric for POT_{N1}, POT_{N2}, and POT_{N3} extreme discharge forecasts in each of the three basins. We calculated the KGE between observed peak flows and forecasted peak flows for reforecasts with initialization dates 1, 2, . . . , and 28 days prior to the peak of each individual event among POT events (see Fig. S2 for an illustration). Given that KGE does not include a built-in benchmark (Gupta et al. 2009; Knoben et al. 2019), we chose a benchmark KGE that was calculated based on mean flow [or median, depending on whichever results in a higher benchmark KGE following Knoben et al. (2020)] of the same calendar day for a given POT date over 18 years excluding the target year.

(ii) Probabilistic skill

We evaluated the probabilistic flood forecast skill of all six SubX models with 30 ensemble members in total. We used the evaluation metrics from DeFlorio et al. (2019b) where they assessed the subseasonal forecast skill of ARs, including 1) debiased Brier skill score (BSS) (Weigel et al. 2007) and 2) relative operating characteristic (ROC)-like (Hanley and McNeil 1982) diagrams (DeFlorio et al. 2019b), which account

for both hit rate and false alarm rate in a lead-dependent manner. We discuss these two skill measures and our application of them briefly below.

1) BSS

BSS is a measure of probabilistic skill relative to climatology that is sensitive to small ensemble sizes. Following DeFlorio et al. (2019b), we used debiased BSS, which adds a correction term in the denominator of BSS to overcome the small ensemble size issue. The debiased BSS is calculated as follows:

$$\mathrm{BSS} = 1 - \frac{\mathrm{BS}}{\mathrm{BS}_{\mathrm{ref}} + D}, \tag{1}$$

$$\mathrm{BS} = \frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2, \tag{2}$$

$$\mathrm{BS}_{\mathrm{ref}} = \frac{1}{N}\sum_{i=1}^{N}(P_{\mathrm{clim}} - O_i)^2, \tag{3}$$

$$D = \frac{1}{M}P_{\mathrm{clim}}(1 - P_{\mathrm{clim}}), \tag{4}$$

where N is the number of reforecast samples during October–March; P_i represents the forecast ability of a particular level of discharge events, which is the fraction of ensemble members that predict maximum discharge falling into a particular category during a week-long lead period for a single reforecast event; O_i is the binary representation of whether the observed discharge fell into that category (1 if yes, 0 if no); M is the ensemble size; and P_{clim} is the probability of the reference climatology.

This metric can be sensitive to the choice of the reference climatology (Bartholmes et al. 2009). Here, we focused on POT extreme discharge events. For $\mathrm{POT}_{\mathrm{N1}}$ events (with the threshold set to one event per year on average), we set P_{clim} to 1/26, corresponding to one extreme event occurring over 6 months (i.e., 26 weeks). For $\mathrm{POT}_{\mathrm{N2}}$ events (with the threshold set to two events per year on average), we set P_{clim} to 2/26 (the interval of two independent peaks is greater than 1 week based on the criteria), and so forth. BSS ranges from $-\infty$ to 1. Values above 0 indicate that the reforecast skill is higher than skill of reference climatological forecast.

Moreover, as mentioned previously, ESP is more suitable than climatology as the baseline for the evaluation of forecast skill because ESP leverages the knowledge of IHCs. Therefore, we also report the difference in the SubX-based and ESP-based BSS values (denoted as "$\Delta\mathrm{BSS}_{\mathrm{SubX\text{-}ESP}}$"; see section 3d for details of ESP). Besides, in order to examine whether all of the hydrologic skill comes from the first 1–2 weeks, we included additional runs driven by the ensemble mean of NCEP-CFSv2 forecasts for the first 2 weeks followed by ESP for weeks 3–4 [denoted as "NCEP_ESP"; similar to Shukla et al. (2012)]. We considered NCEP-CFSv2 to be a standard medium range weather forecast since the NCEP forecast system in the SubX dataset is the same as the Global Ensemble Forecast System (GEFS) v11 with some upgrades (Guan et al. 2019). The difference between the SubX-based and NCEP_ESP-based BSS values is denoted as "$\Delta\mathrm{BSS}_{\mathrm{SubX\text{-}NCEP\_ESP}}$." Moreover, previous studies evaluated the influence of limited number of extreme events on BSS using bootstrapping (e.g., Addor et al. 2011; Liechti et al. 2013). Similarly, we estimated 90% confidence intervals for $\Delta\mathrm{BSS}_{\mathrm{SubX\text{-}ESP}}$ and $\Delta\mathrm{BSS}_{\mathrm{SubX\text{-}NCEP\_ESP}}$ based on a 1000-sample bootstrapping procedure with replacement.

2) ROC-LIKE DIAGRAMS

A ROC-like diagram shows the ensemble mean, hit rates, and false alarm rates for the prediction of a quantity of interest. The hit rate is calculated as the number of hits divided by the total number of hits + misses. The false alarm rate is calculated as the number of false alarms divided by the total number of false alarms + correct rejections. For each POT category and each week of lead time, the terms comprising these rates are defined as below:

Hit = a POT extreme discharge event is observed, and it is forecasted.
Miss = a POT extreme discharge event is observed, but it is not forecasted.
False alarm = a POT extreme discharge event is forecasted, but it is not observed.
Correct rejection = a POT extreme discharge event is not forecasted, and it is not observed.

d. ESP and revESP implementation

In our implementation of ESP, we used IHCs from section 3b. ESP considers the IHCs to be "true" and the model is forced with resampled gridded observations. For each IHC date, we extracted the following 28 days of observed forcings for the same calendar day over the 18 years of the reforecast period with the target year excluded. We then forced the model with 17 ensemble members starting from the IHC date for a period of 28 days. We compared the SubX-based forecasts with ESP to see which was more skillful.

Similarly, revESP samples IHCs from climatology to initialize the model, which is forced with true observations. For a given forecast date, the revESP experiments sampled 17 IHCs for the same calendar day over 18 years of the reforecast period with the target year excluded to initialize the model. The model was then forced with the true (observed) forcings for a period of 28 days.

4. Results

a. Evaluation of SubX reforecasts

1) PRECIPITATION AND TEMPERATURE SKILL

We examined the precipitation skill of the individual SubX models, as well as the multimodel ensemble mean (denoted as "MME"), at lead times of 1–4 weeks in each river basin (see Fig. 2 and Table 2; week 4 refers to days 22–28). We also examined each month during the October–March period separately. Figure 2 shows that precipitation skill (as measured by ACC) drops quickly after week 1. In week 2, almost all models have positive ACC in all months, but by week 3, some models show negative ACC in certain months. Over all months (October–March; see bottom panel in Fig. 2 and also Table 2),

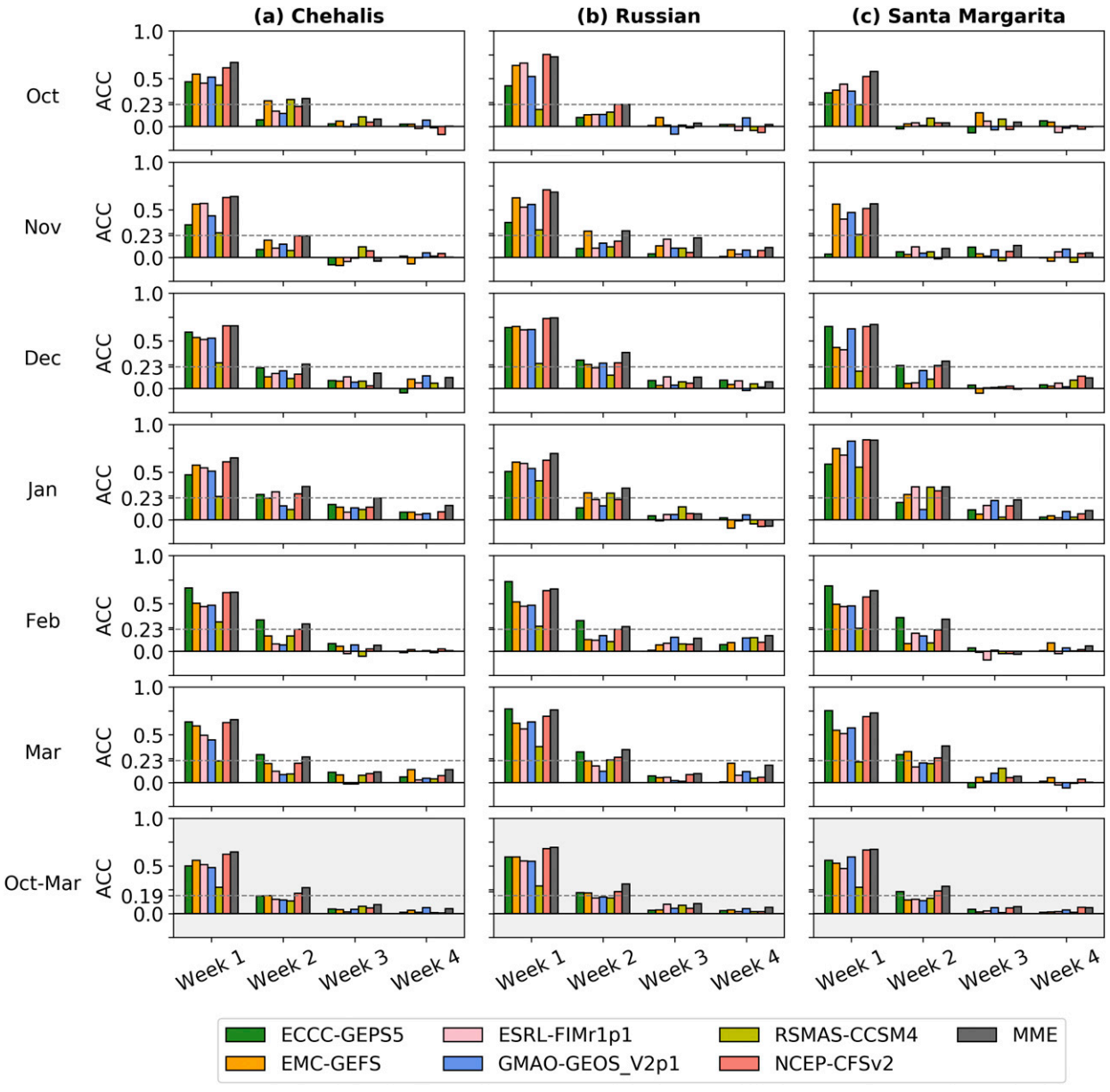

FIG. 2. Precipitation prediction skill [as measured by the anomaly correlation coefficient (ACC)] of SubX models averaged over each basin and each week before bias correction. Different rows are for different months and the bottom row shows ACC values over all months (October–March). The critical correlation shown by dashed lines (beyond which the ACC is statistically different from zero at the 5% significance level).

SubX models as well as MME only show statistically significant skill in weeks 1–2. Among individual models, NCEP-CFSv2 performs best in weeks 1–2 in the three basins, with similar skills to MME. However, the model performance at longer lead times varies by lead time and basin. In weeks 3–4, GMAO-GEOS_V2p1 generally shows better performance than other models (although none of them are statistically significant) across three basins.

We also examined temperature forecast skill. Figure 3 shows the results for Tmax (since the pattern for Tmin is similar, it is shown in Fig. S3). Similar to precipitation, Tmax skill drops quickly after week 1. However, there are fewer negative ACC values for Tmax in weeks 3–4 in comparison with precipitation. In general, Tmax generally has higher skill than precipitation at all weeks. Over all months (October–March; see bottom panel in Fig. 3 and also Table 2), SubX models as well as MME generally show statistically significant skill in all weeks. Across models, NCEP-CFSv2 shows better performance in weeks 1–2 than the other models across the three basins, while in weeks 3–4,

TABLE 2. Precipitation and temperature skill (as measured by the ACC) of SubX models and multimodel ensemble mean (MME) during winter months (October–March). The ACC is averaged over each basin before bias correction. The maximum ACC value across SubX models in each row is marked in bold font.

Variable	Basin	Week	ECCC-GEPS5	EMC-GEFS	ESRL-FIMr1p1	GMAO-GEOS_V2p1	RSMAS-CCSM4	NCEP-CFSv2	Min	Max	MME
Precipitation	Chehalis	1	0.50	0.56	0.51	0.48	0.28	**0.62**	0.28	0.62	0.64
		2	0.18	0.19	0.15	0.14	0.13	**0.21**	0.13	0.21	0.27
		3	0.05	0.04	0.02	0.05	**0.08**	0.06	0.02	0.08	0.10
		4	0.02	0.03	0.01	**0.06**	0.01	0.01	0.01	0.06	0.05
	Russian	1	0.60	0.59	0.55	0.55	0.29	**0.68**	0.29	0.68	0.69
		2	0.22	0.22	0.16	0.18	0.16	**0.23**	0.16	0.23	0.31
		3	0.04	0.04	**0.10**	0.06	0.09	0.06	0.04	0.10	0.10
		4	0.03	0.04	0.02	**0.05**	0.02	0.02	0.02	0.05	0.06
	Santa Margarita	1	0.56	0.53	0.47	0.59	0.28	**0.67**	0.28	0.67	0.67
		2	0.23	0.14	0.15	0.14	0.16	**0.24**	0.14	0.24	0.29
		3	0.05	0.02	0.03	**0.06**	0.01	**0.06**	0.01	0.06	0.07
		4	0.01	0.02	0.02	0.04	0.01	**0.07**	0.01	0.07	0.06
Tmax	Chehalis	1	0.64	0.59	0.57	0.50	0.51	**0.78**	0.50	0.78	0.78
		2	0.50	**0.56**	0.50	0.39	0.41	0.52	0.39	0.56	0.59
		3	0.36	**0.46**	0.37	0.22	0.33	0.39	0.22	0.46	0.46
		4	0.30	**0.39**	0.31	0.19	0.24	0.34	0.19	0.39	0.42
	Russian	1	0.67	0.62	0.58	0.52	0.65	**0.81**	0.52	0.81	0.81
		2	**0.54**	0.52	0.44	0.35	0.44	0.53	0.35	0.54	0.58
		3	0.34	**0.38**	0.32	0.18	0.30	0.33	0.18	0.38	0.43
		4	0.26	**0.33**	0.32	0.16	0.19	0.32	0.16	0.33	0.37
	Santa Margarita	1	0.63	0.57	0.56	0.51	0.64	**0.81**	0.51	0.81	0.80
		2	**0.48**	0.45	0.42	0.32	0.40	0.47	0.32	0.48	0.51
		3	0.26	0.27	0.24	0.10	0.24	**0.29**	0.10	0.29	0.31
		4	0.18	**0.23**	0.19	0.12	0.13	**0.23**	0.12	0.23	0.25
Tmin	Chehalis	1	0.40	0.45	0.43	0.53	0.38	**0.73**	0.38	0.73	0.71
		2	0.19	0.34	0.30	0.37	0.26	**0.43**	0.19	0.43	0.43
		3	0.12	0.22	0.18	0.20	0.17	**0.24**	0.12	0.24	0.26
		4	0.10	**0.15**	0.11	0.12	0.09	0.12	0.09	0.15	0.19
	Russian	1	0.35	0.36	0.44	0.40	0.41	**0.68**	0.35	0.68	0.62
		2	0.18	0.32	0.31	0.31	0.27	**0.38**	0.18	0.38	0.40
		3	0.13	**0.23**	0.16	0.18	0.21	**0.23**	0.13	0.23	0.27
		4	0.16	**0.20**	0.14	0.17	0.11	0.18	0.11	0.20	0.23
	Santa Margarita	1	0.36	0.46	0.50	0.55	0.56	**0.73**	0.36	0.73	0.71
		2	0.25	0.46	0.42	0.43	0.39	**0.51**	0.25	0.51	0.53
		3	0.18	**0.34**	0.28	0.29	0.26	**0.34**	0.18	0.34	0.36
		4	0.16	**0.27**	0.23	0.22	0.14	**0.27**	0.14	0.27	0.30

EMC-GEFS generally shows better performance. Same was found for Tmin.

2) PERFORMANCE OF DAILY BCSD

The difference in precipitation skill (as measured by ACC) before and after applying the daily BCSD is very small. This meets our expectation since the QM is performed in a lead-time-dependent manner. Figures 4a and 4b shows the basin-average relative bias (the bias is defined as model minus observation) of precipitation forecasts before and after applying daily BCSD_7d in three basins, averaged over October–March. Before applying daily BCSD, the absolute relative biases were up to 25%, 37%, and 60% across models and over weeks 1–4 in the three basins, respectively. They were reduced to below 4%, 10%, and 15% (across models and over weeks 1–4) after applying BCSD_7d, and 5%, 11%, and 23% after applying BCSD_1d. Also, the precipitation biases are more spatially coherent after bias correction (see Figs. S4–S6). The biases in temperature were also reduced from up to 2.7°C to below 0.5°C (see Figs. 4c,d).

b. *Hydrologic model evaluation*

We examined model performance of the control run, forced by the Livneh et al. (2013) data after hourly disaggregation. We evaluated streamflow time series at gauges across each basin using KGE, normalized root-mean-square error (RMSE) and relative bias (see Table S1, Text S2, and Fig. S7). The KGE values are greater than 0.65 at most gauge locations but are lower (0.45–0.61) at certain gauges in the Russian and Santa Margarita basins (see Table S1), although this mostly reflects errors in prediction

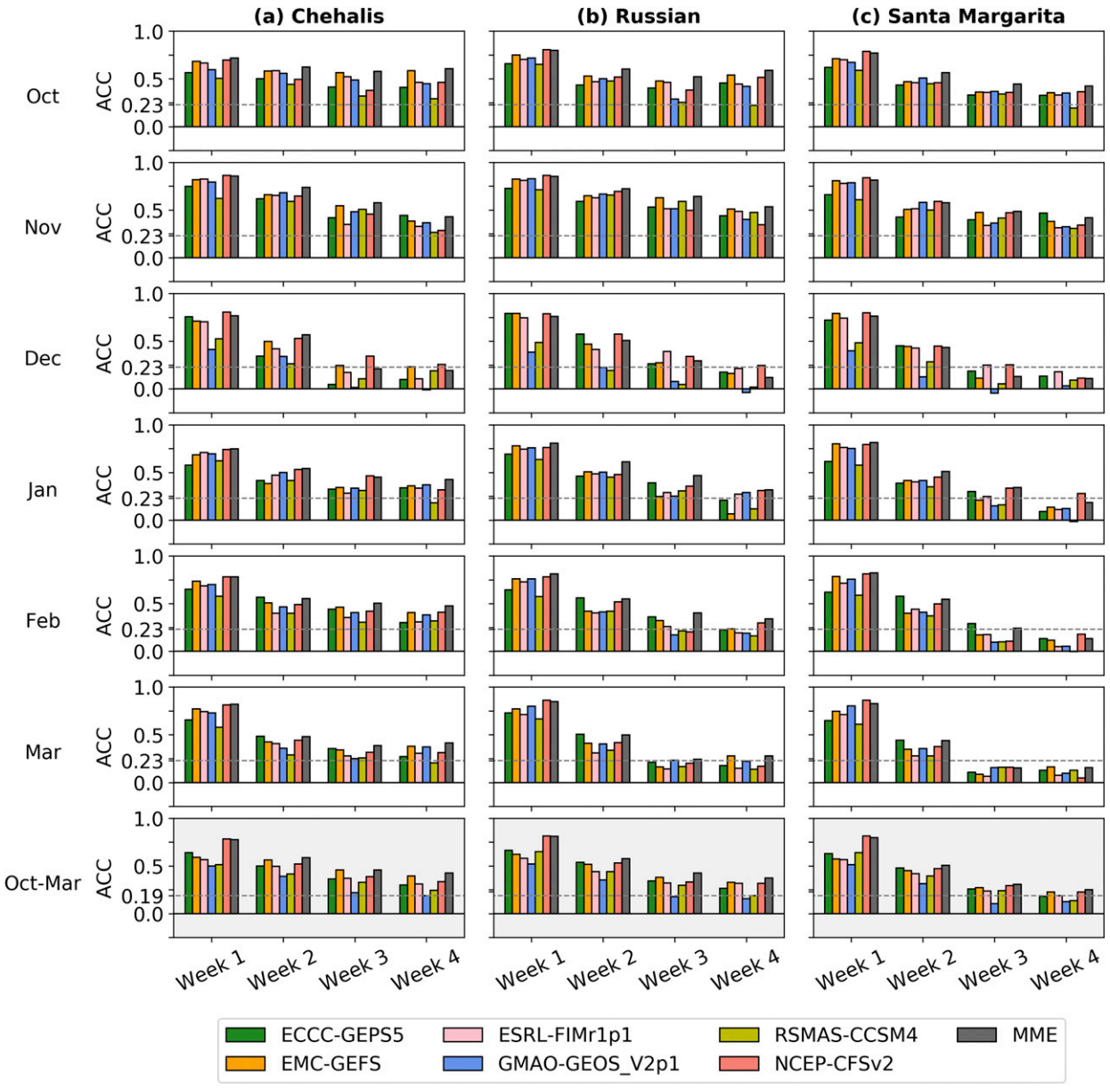

FIG. 3. Maximum daily temperature (Tmax) prediction skill [as measured by the anomaly correlation coefficient (ACC)] of SubX models averaged over each basin and each week before bias correction. Different rows are for different months and the bottom row shows ACC values over all months (October–March). The critical correlation shown by dashed lines (beyond which the ACC is statistically different from zero at the 5% significance level).

of low flows. Given that this study focuses on peak flows, we further examined model performance during extreme discharge events given our interest in flood forecasting at the downstream-most USGS stream gauge in each basin (see Fig. 1 for gauge locations). Figure 5 compares simulated and observed peak flows for POT_{N3} extreme discharge events during the SubX period. The simulated peaks show reasonable matches with observed peaks, with KGE values of 0.68, 0.89, and 0.86 at downstream-most gauges in the Chehalis, Russian, and Santa Margarita River basins, respectively. In Fig. 5, we also show events associated with ARs. The percentages of POT_{N3} extreme discharge events that were coincident with ARs during 1999–2016 are 52%, 74%, and 41% in each of the three respective basins. We can see that most of the largest extreme discharge events were AR-related, especially in the Russian River basin. The AR-related percentage increases as POT threshold increases.

c. *Assessment of flood forecast skill*

1) DETERMINISTIC SKILL

Figure 6 shows the deterministic forecast skill (KGE) of the NCEP-CFSv2 model (based on the average of its four

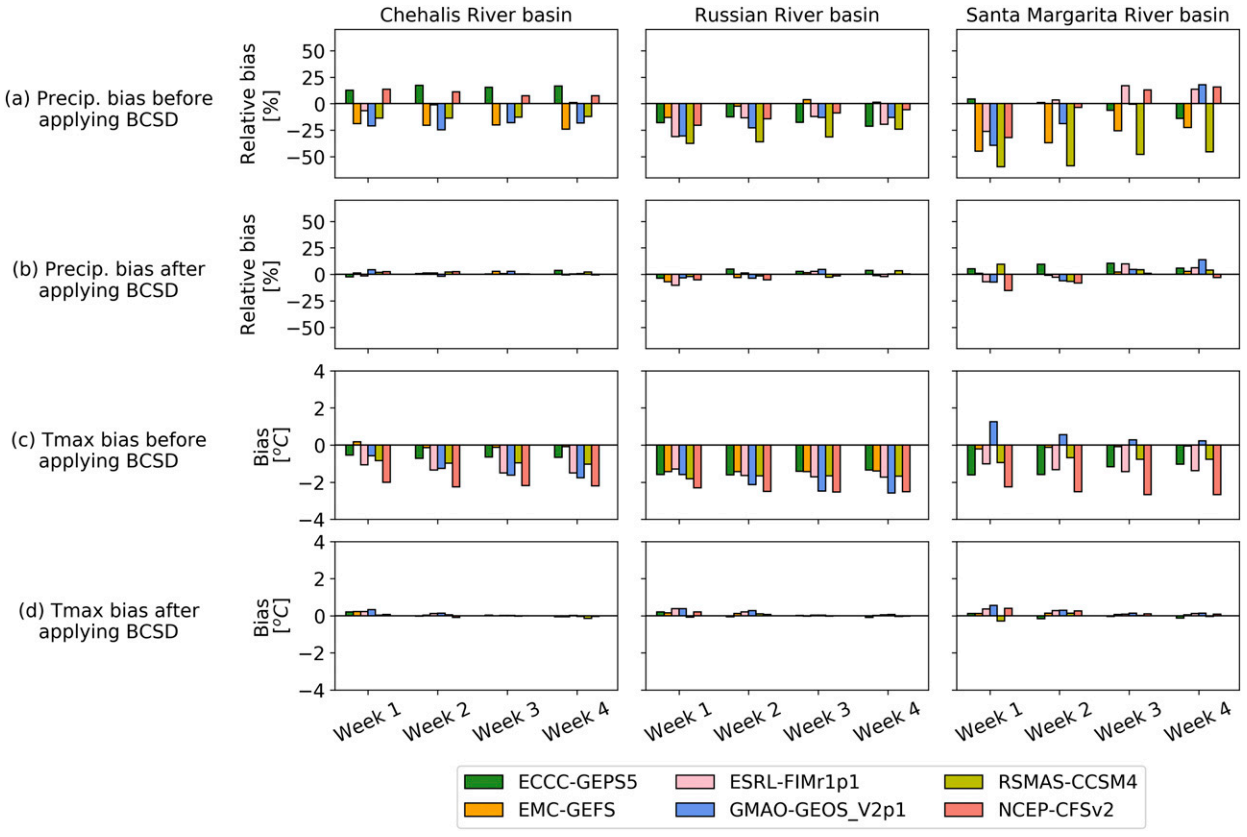

FIG. 4. Precipitation bias of SubX models averaged over each basin and over October–March (a) before and (b) after bias correction. (c),(d) As in (a) and (b), but for maximum daily temperature (Tmax). The bias is defined as model minus observation.

ensemble members) as a function of lead time for POT_{N1}, POT_{N2} and POT_{N3} extreme discharge events (with thresholds set to 1, 2, and 3 events per year on average). The KGE drops quickly with lead time. For POT_{N1} events, KGE drops below benchmarks (i.e., when the skill is no better than climatology) after lead times of 14, 17, and 13 days in the Chehalis, Russian, and Santa Margarita River basins, respectively. For POT_{N2} (POT_{N3}) events, KGE drops below benchmarks after lead times of 19, 18, and 8 days (24, 22, and 8 days) in the three respective basins, several days longer

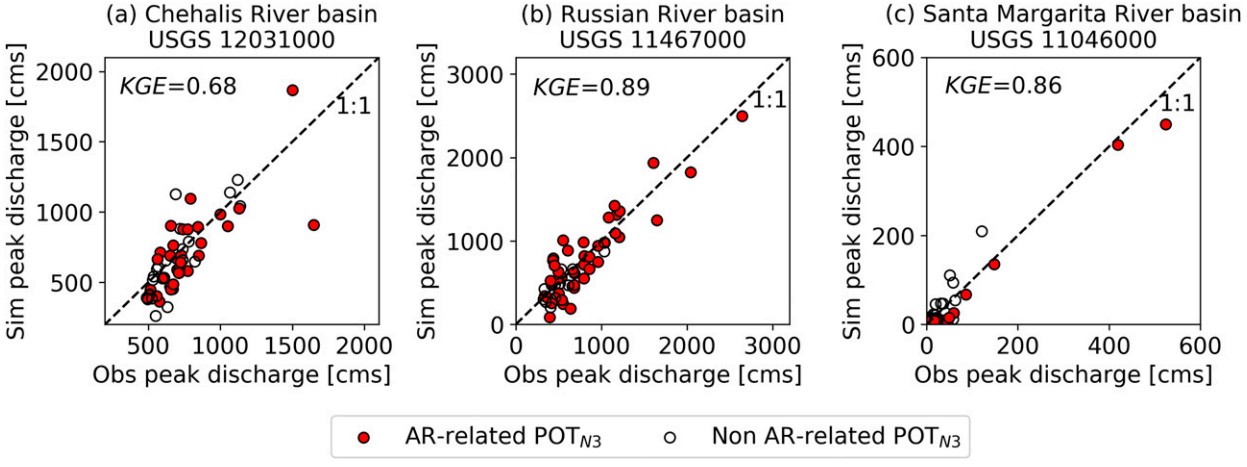

FIG. 5. Evaluation of hydrologic model performance: simulated vs observed daily peak flow of POT_{N3} extreme discharge events (with threshold set to 3 events per year on average) during the period of 1999–2016. The events associated with ARs are marked by circles filled with red color.

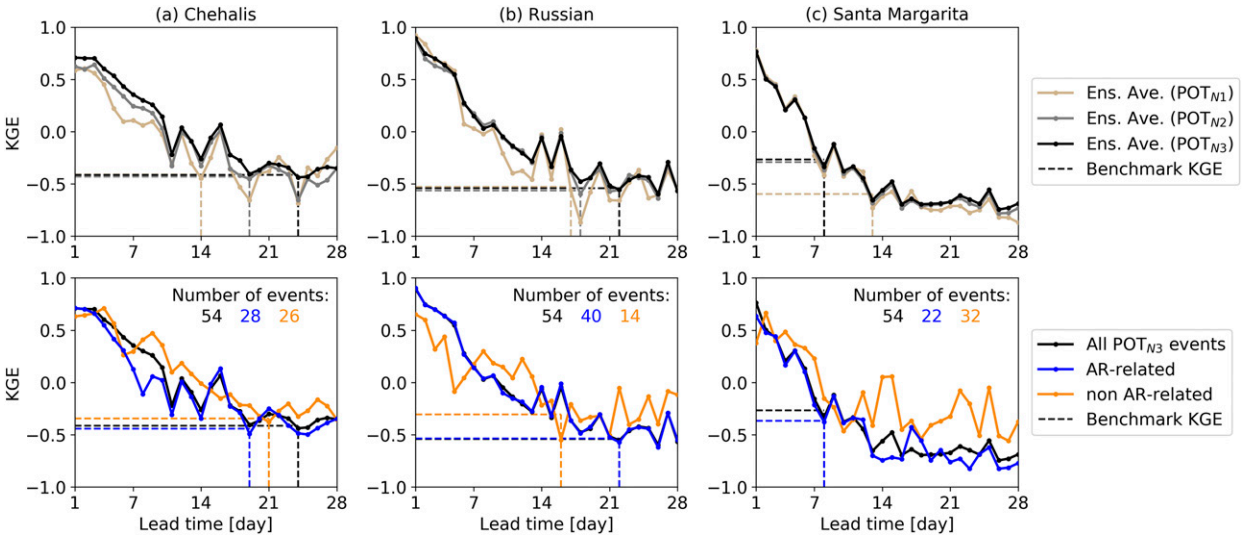

FIG. 6. (a)–(c) The deterministic flood forecast skill [in Kling–Gupta efficiency (KGE)] of the NCEP-CFSv2 model vs lead time. (top) The skill for POT_{N1}, POT_{N2}, and POT_{N3} extreme discharge events (with threshold set to 1, 2, and 3 events per year on average); (bottom) the skill for AR-related and non-AR-related POT_{N3} events. Horizontal dashed lines show benchmark KGE values and vertical dashed lines indicate lead times when the forecast skill drops below them. In the lower panel of (c), the benchmark KGE of non-AR-related events is below −1.0 and hence no dashed orange line is shown.

than POT_{N1} events in the Chehalis and Russian River basins. Figure 6 also shows the difference in the skill of AR-related and non-AR-related POT_{N3} events. The lead time when there is skill for the AR-related events is generally shorter than the non-AR-related ones, except in the Russian River basin. This is possibly due to the relatively small sample size of non-AR-related POT_{N3} events (14 out of 54 events) in this basin.

2) PROBABILISTIC SKILL

Figure 7a shows the SubX-based BSS values (denoted as "BSS_{SubX}") for week 1–4 lead time for POT_{N1} and POT_{N3} extreme discharge events (denoted as "BSS_{POTN1}" and "BSS_{POTN3}," respectively) in the three basins. The BSS_{POTN1} drops quickly after week 1 and is close to 0 after week 2 in the three basins. The BSS values are 0.40, 0.07, 0.03, and 0.04 for weeks 1–4 in the Chehalis River basin, 0.39, 0.10, 0.04, and 0.01 in the Russian River basin, and 0.40, 0.17, 0.06, and 0.06 in the Santa Margarita River basin. When we lower the threshold to POT_{N3}, BSS_{POTN3} generally increases in all weeks, which may be partly due to the increasing influence of ASM as opposed to SubX model skill, as the role of ASM increases for smaller events (as discussed in the following sections). Figure 7a also shows the 90% confidence interval of differences between BSS_{POTN1} and BSS_{POTN3} (denoted as "$\Delta BSS_{POTN3\text{-}POTN1}$") derived by bootstrapping. BSS is generally higher for small events than for large events.

Figure 7b shows the difference between the BSS_{SubX} and the ESP-based BSS values ($\Delta BSS_{SubX\text{-}ESP}$) as well as the NCEP_ESP-based values ($\Delta BSS_{SubX\text{-}NCEP\_ESP}$) for POT_{N1} events. The median of $\Delta BSS_{SubX\text{-}ESP}$ generally decreases with lead time. The $\Delta BSS_{SubX\text{-}ESP}$ is generally statistically different from zero in weeks 1–2 across the three basins, which indicates that SubX is more skillful than ESP, while in weeks 3–4, both $\Delta BSS_{SubX\text{-}ESP}$ and $\Delta BSS_{SubX\text{-}NCEP\_ESP}$ are not statistically different from zero. When we lower the threshold to POT_{N3}, the $\Delta BSS_{SubX\text{-}ESP}$ is statistically different from zero in weeks 1–3 across the three basins. So is the $\Delta BSS_{SubX\text{-}NCEP\_ESP}$ in week 3, indicating that the skill (for POT_{N3} events) does not all come from the improvement in the first 2 weeks.

We examined the hit rate and false alarm rate in ROC-Like diagrams (see Fig. 8a). For POT_{N1} events, the hit rate drops quickly from week 1 to week 2 in all three basins and the false alarm rate slightly increases. After week 2, the positive skill is low. As the threshold is lowered to POT_{N2} and to POT_{N3}, the hit rate increases, but the false alarm rate also increases. Across basins, the hit rate in the Santa Margarita River basin is higher after week 1 in comparison with the other two basins, but its false alarm rate is also higher, reflecting the larger natural variability (and hence forecast uncertainty) in this basin.

3) INFLUENCING FACTORS IN FLOOD FORECAST SKILL

We examined factors, including ASM, ARs, and storm precipitation, that might affect flood forecast skill. We separated all POT events shown in Fig. 8a based on each factor. The three pairs for comparison are 1) events with wet ASM versus events with dry ASM, separated by median values; 2) AR-related events versus non AR-related events; 3) events with large storm precipitation versus events with small storm precipitation, separated by median values (see Figs. 8b–d; see also Table S2 for the number of events in each category).

Figure 8b shows that forecast skill is generally higher when ASM is wet than dry for all weeks and across all basins, except for weeks 3–4 in the Russian River basin. However, the difference between the two groups decreases as the POT threshold increases and lead time increases. Also, forecast skill

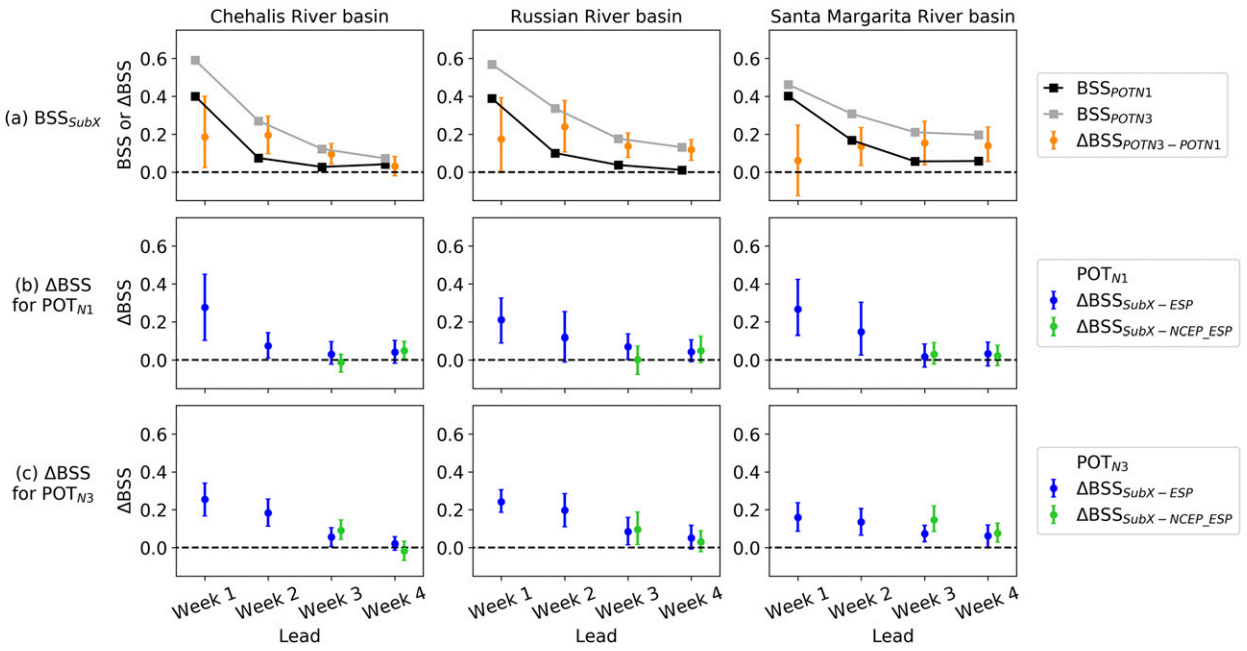

FIG. 7. (a) SubX-based Brier skill score (BSS; denoted as "BSS$_{SubX}$" and shown as square symbols) over weeks 1–4 lead time for POT$_{N1}$ (denoted as "BSS$_{POTN1}$") and POT$_{N3}$ (denoted as "BSS$_{POTN3}$") extreme discharge events (with threshold set to 1 and 3 events per year on average). The boxplot shows a 90% confidence interval of their differences (denoted as "ΔBSS$_{POTN3\text{-}POTN1}$") derived by bootstrapping. The case when there is no overlapping with zero indicates that the difference is significant. (b) Difference between the BSS$_{SubX}$ and the ESP-based BSS (denoted as "ΔBSS$_{SubX\text{-}ESP}$"), and difference between the BSS$_{SubX}$ and the NCEP_ESP (i.e., the NCEP is used for weeks 1–2 and ESP for weeks 3–4) based BSS (denoted as "ΔBSS$_{SubX\text{-}NCEP_ESP}$") for POT$_{N1}$ events. (c) As in (b), but for POT$_{N3}$ events.

is generally higher for the non-AR-related events than for AR-related ones (see Fig. 8c). If we consider the POT$_{N3}$ events, which have a larger sample size (and hence smaller magnitudes) than the other two POT categories, this tendency (higher skill for non-AR events) is stronger in weeks 1–2 in the Chehalis and Russian River basins, and for all lead times in the Santa Margarita River basin. Events associated with ARs generally have higher storm precipitation than non-AR events, so this result may well be reflective of the fact that forecast skill is higher for less extreme events (because the role of ASM is greater). The pattern for the skill of events separated by storm precipitation is similar to ARs versus non-ARs (see Fig. 8d). The forecast skill of extreme discharge events is higher when storm precipitation is small. Despite the fact that AR versus non-AR, and groups separated storm precipitation amounts show similar differences (higher forecast skill for smaller events), they are not as distinct as the effects of ASM, especially at shorter lead times and for lower POT thresholds, as shown in section 4d (and section 5) below.

d. Relative influence of SubX reforecast skill and ASM

To further examine the relative influence of SubX reforecast skill and ASM on streamflow forecast skill and how it evolves with lead times, we conducted ESP/revESP experiments. Streamflow forecast errors and lead times for ESP and revESP experiments are shown in Fig. 9. The RMSE for ESP generally increases with lead time, while the RMSE for revESP generally decreases with lead times in all three basins, but both with large variations in the Santa Margarita River basin. The RMSE for revESP is larger than that for ESP in the first few days (with longer time when using median values instead of mean values; see Fig. S8), indicating that ASM dominates the streamflow forecast skill at shorter lead times. The periodic cycles in Fig. 9 are an overlaying effect of all reforecasts.

We also compared ESP with SubX-based forecasts. The RMSE for SubX is lower than ESP over all lead times in all three basins, with a few exceptions in the Santa Margarita River basin. This indicates that SubX forecasts are more skillful than ESP. In addition, we compared the effect of populating the sample distribution for QM in daily BCSD by pooling the following 7 days of a model day versus using 1 day only. We found that the RMSE of BCSD_7d was slightly lower than that of BCSD_1d, with greater difference at longer lead times.

We further examined the RMSE ratios to evaluate the relative contributions of ASM and SubX reforecast skill to streamflow forecast skill. Figure 10 shows the ratios by month. The ESP-based ratio is generally higher than the SubX-based ratio, but the difference becomes smaller with lead time. If the RMSE ratio is less than one, we infer that ASM dominates the streamflow forecast skill and vice versa. The lead time when the SubX-based RMSE ratio exceeds one generally decreases from October to March in the Chehalis and Russian River basins, suggesting that ASM dominates the streamflow forecast skill in fall months, but that the effect is reduced through the winter. This is consistent with the fact that early fall soil moisture is dominated by the prolonged preceding summer dry period. There is no clear pattern in the Santa Margarita River basin for the lead time when RMSE ratios exceed one across months, possibly due to its larger relative variability of streamflow in comparison with the other two basins and hence larger estimation difficulties. Overall, the SubX forecast

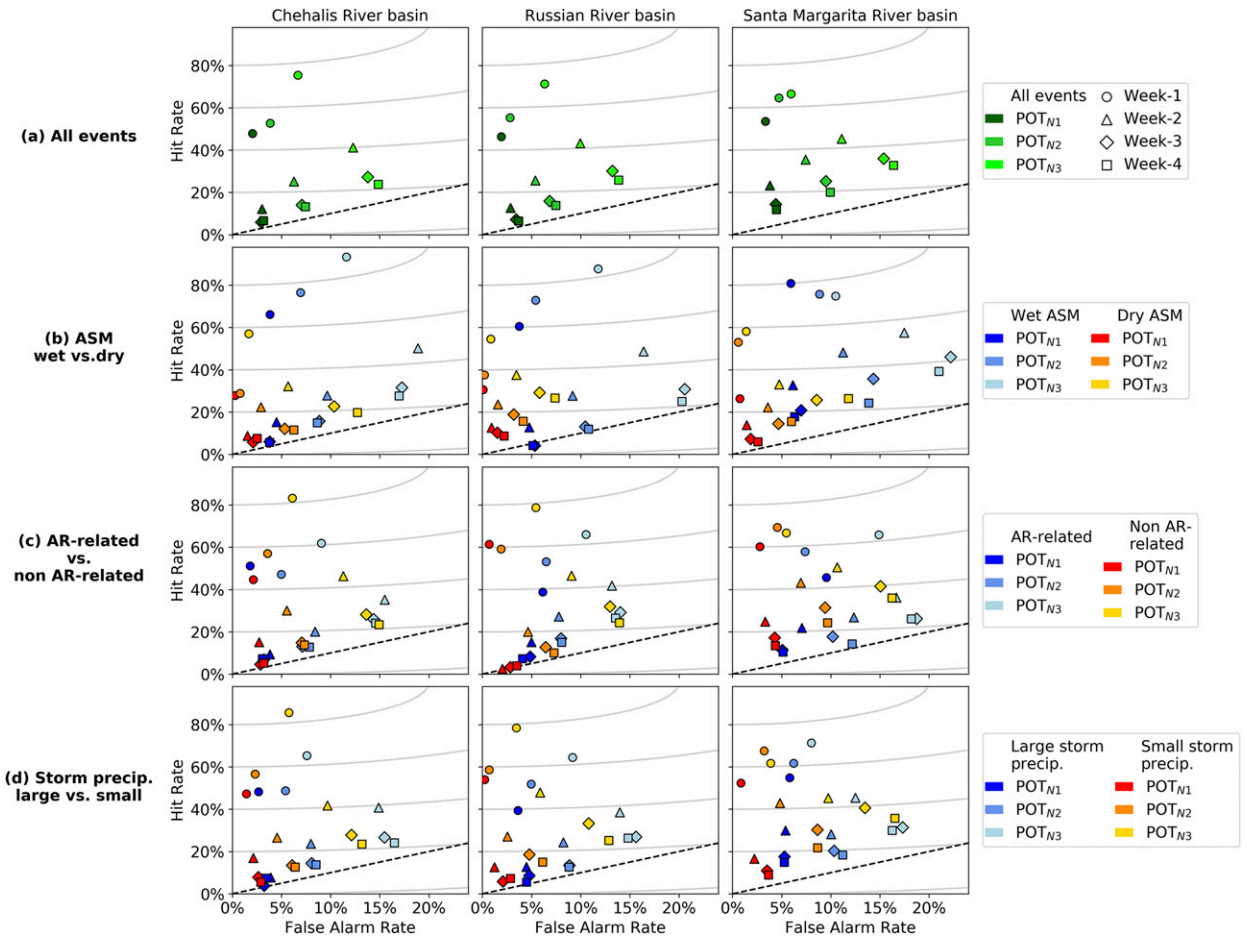

FIG. 8. ROC-like diagrams (ensemble mean hit rate vs false alarm rate) of POT_{N1}, POT_{N2}, and POT_{N3} extreme discharge events in three basins, for (a) all events; (b) events with wet antecedent soil moisture (ASM) vs events with dry ASM, separated by median values; (c) AR-related events vs non AR-related ones; and (d) events with large storm precipitation vs events with small storm precipitation, separated by median values. The shapes of symbols indicate different weeks.

skill starts to dominate streamflow forecast skill after no longer than lead 9 days.

5. Discussion

In section 4d we examined the relative influence of ASM and SubX reforecasts to streamflow (deterministic) forecast skill using RMSE and RMSE ratios following previous studies (e.g., Wood and Lettenmaier 2008; Li et al. 2009; Shukla and Lettenmaier 2011). For the largest (i.e., POT_{N1}) peak discharge events, we showed in section 4c(1) that there was only marginal forecast skill relative to climatology for lead times greater than about 2 weeks. We further examine here the relative influence of ASM and SubX reforecast skill on flood

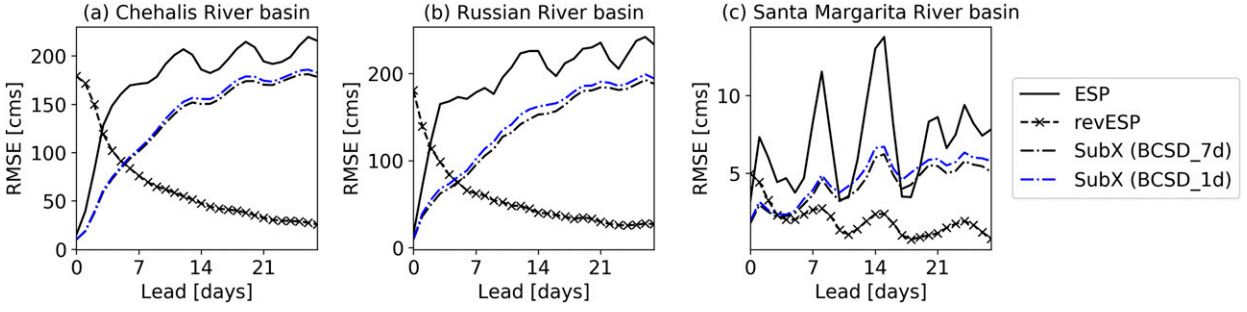

FIG. 9. RMSE of streamflow forecasts for ESP, revESP, and SubX-based forecasts.

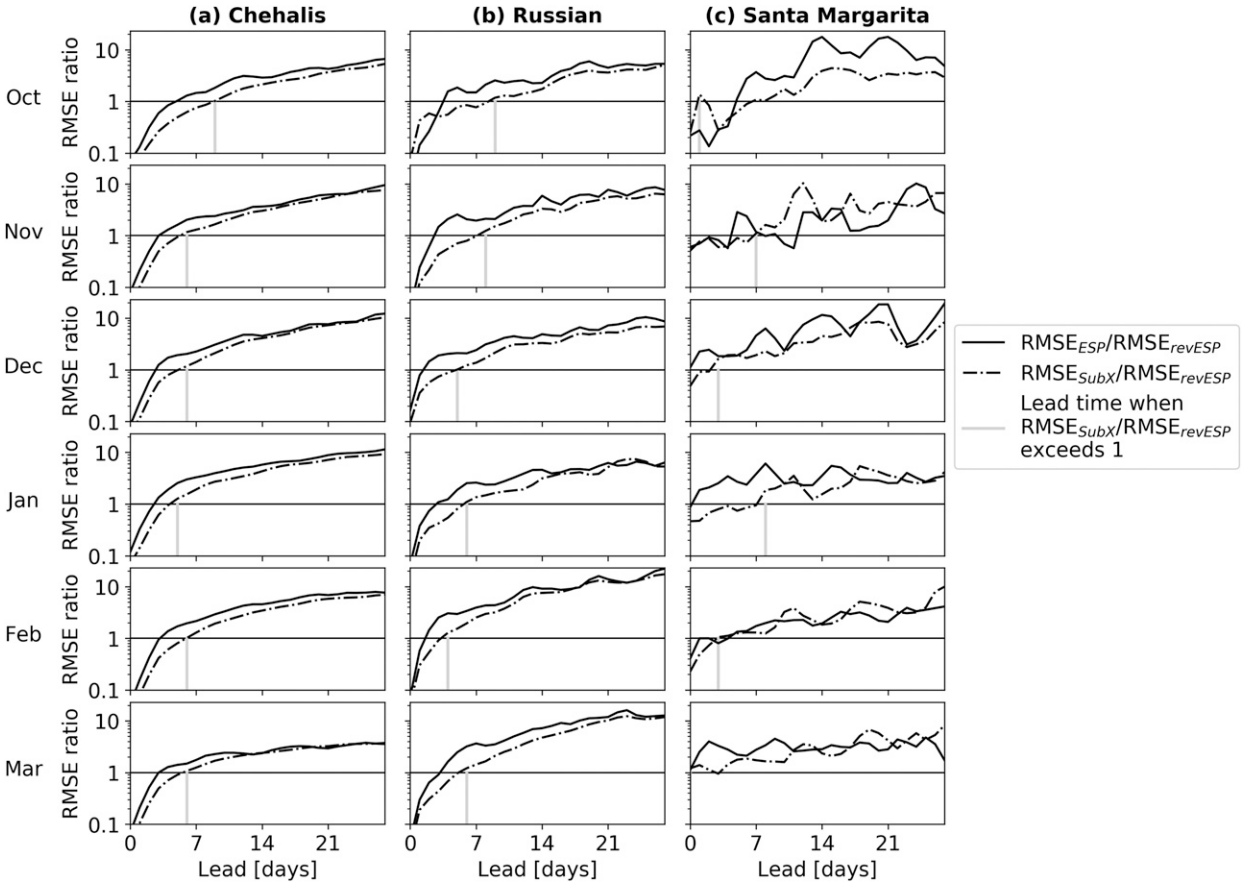

FIG. 10. Variation of RMSE ratios ($RMSE_{ESP}/RMSE_{revESP}$ and $RMSE_{SubX}/RMSE_{revESP}$) with lead time in three river basins in each month from October to March. The lead time when $RMSE_{SubX}/RMSE_{revESP}$ exceeds one is marked by a vertical gray line.

probabilistic forecast skill using BSS as the evaluation metric (as in Fig. 7). We compared the BSS_{SubX} and revESP-based BSS (denoted as "BSS_{revESP}") (see Fig. 11). Similar to the RMSE ratios, in each week, if $BSS_{SubX} \leq BSS_{revESP}$, we consider that the SubX forecast skill dominates the flood probabilistic forecast skill and vice versa. We can see that the BSS_{SubX} is smaller than BSS_{revESP} in all weeks except for the POT_{N2} and POT_{N3} events in the Chehalis River basin in week 1, POT_{N3} events in the Russian River basin in week 1, and POT_{N2} and POT_{N3} events in the Santa Margarita River basin in week 1, indicating that for most large flood events (i.e., POT_{N1}) in the three basins, the SubX reforecast skill dominates the flood probabilistic forecast skill in all weeks.

In previous sections, we evaluated the flood forecast skill using BSS values and hit-miss rates for flood events that occurred. However, another important aspect of forecasting is skill in predicting when an AR-driven flood *will not* occur, which could permit water managers to alter their operations. Hence, we examined the hit-miss rates for times when an extreme precipitation event occurs, but an extreme discharge event does not (denoted as "$POT_{NP}\backslash POT_{ND}$" events; see Fig. S9a). The skill in forecasting no flood occurrences is generally higher for large than small events, a pattern that is most clear in the Santa Margarita River basin. In comparison with Fig. 8a which shows flood events that occurred, we can see that the skill in forecasting flood occurrences (versus no occurrences) is higher for small than large events. We also examined the impact of wet/dry ASM conditions (see Fig. S9b). For no-flood-occurrence events, when ASM is dry, the forecast skill is generally higher than when ASM is wet.

We also compared the performance of daily BCSD and LOCA for the downscaling of precipitation of one SubX model, the ECCC-GEPS5. Figure S10a shows the precipitation skill (as measured by ACC) before and after applying daily BCSD and LOCA, the difference of which is very small. Figure S10b shows the basin-average relative bias of precipitation. The difference in the effect of removing precipitation bias (in the examination on a weekly basis) between BCSD_7d and LOCA_7d, or BCSD_1d and LOCA_1d, is generally smaller than the difference between BCSD_7d and BCSD_1d, or LOCA_7d and LOCA_1d. However, the difference in BSS values is larger between two methods than choices in the pooling of days (i.e., 7 days versus 1 day; see Fig. S11), possibly because the measures of precipitation bias removal do not consider precipitation intermittency and intensity. The LOCA method showed potential benefit under some circumstances. The LOCA method differs from daily BCSD in that it considers the relationship between local precipitation and large-scale precipitation patterns (see Text S1 for details). LOCA

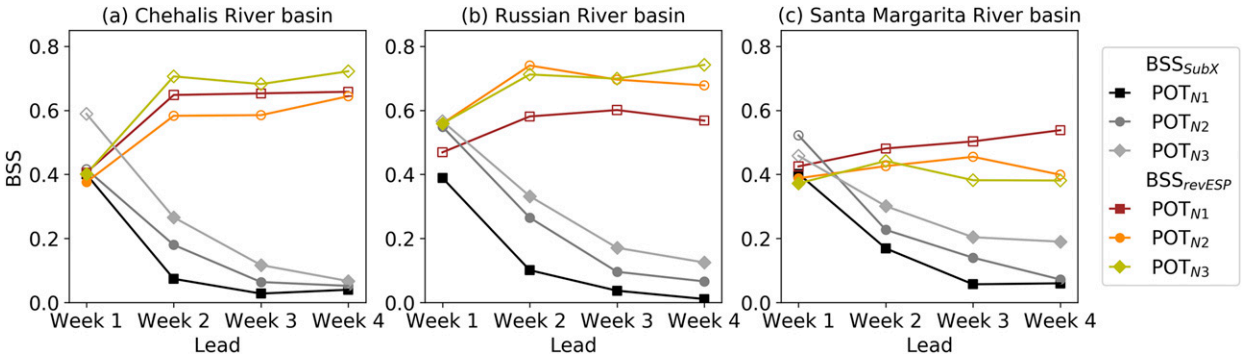

FIG. 11. SubX-based Brier skill score (denoted as "BSS_{SubX}") and revESP-based BSS (denoted as "BSS_{revESP}"). If $BSS_{revESP} \geq BSS_{SubX}$, the marker of former is shown as a hollow symbol while the latter is shown as a solid symbol, which indicates the SubX dominates the forecast skill, and vice versa.

has been shown to produce good estimates of extreme and zero-precipitation days when applied to global climate model (GCM) output (Pierce et al. 2014). Here we tested it for subseasonal forecasts, whose performance may be influenced by the lead-dependent forecast skill. Given that the computational time and effort of LOCA are several times higher than BCSD, we reported the flood forecast skill only for BCSD here, but we intend to investigate the performance of LOCA for downscaling of subseasonal forecasts in our future work.

Our results for precipitation and temperature skill are consistent with Pegion et al. (2019) regarding intermodel comparisons. For precipitation skill in weeks 3–4, we showed that GMAO-GEOS generally performed better than other models (although none of the models are statistically significant at these lead times) across the three basins. This is consistent with Pegion et al. (2019) in who showed that the GMAO-GEOS is most skillful among the SubX models in the simulation and prediction of MJO (which strongly modulates the hindcast skill of ARs and AR-related precipitation) at weeks 3–4. We also showed that the SubX precipitation had no statistically significant skill at weeks 3–4 in this region. For temperature skill in weeks 3–4, we showed that EMC-GEFS performed best, which was also found by Pegion et al. (2019).

More importantly, our study focused on the evaluation of the hindcast skill of AR-related flood events. Because WWRP/WCRP S2S and SubX databases have two models in common, the NCEP and the ECCC models, we can infer the performance of SubX models compared to others in the WWRP/WCRP S2S (e.g., ECMWF). We showed that the NCEP-CFSv2 model (with initialization at a daily interval) had only marginal deterministic flood forecast skill relative to climatology beyond around 2 weeks lead time in each of the three basins. This is generally consistent with Pan et al. (2019) who examined the WWRP/WCRP S2S precipitation skill over the western United States. They found that the models have little deterministic precipitation skill beyond 2 weeks. For weeks 1–2, models with higher resolution tend to have better deterministic skill (e.g., ECCC and ECMWF). The generally better performance of NCEP relative to the other SubX models including ECCC in our study could be partly due to the fact that we used fixed dates each month (see sections 3a and 3b) in our evaluation, which may favor models with finer initialization intervals like NCEP. Pan et al. (2019) also showed that despite the lack of deterministic skill at weeks 3–4 for all models, ECMWF still shows an advantage over the other hindcast systems. In our future work, we will consider the examination of the ECMWF model for flood forecasting. The ongoing development of subseasonal forecast databases such as the NOAA's SubX and the WWRP/WCRP S2S could lead to substantial improvement on subseasonal flood forecast skill, especially with studies on ARs along the U.S. West Coast (e.g., Baggett et al. 2017; DeFlorio et al. 2019a,b; Mundhenk et al. 2018; Nardi et al. 2018), which would better inform reservoir operation plans with forecasts at weeks 3–4. In addition, hybrid statistical–dynamical prediction systems for streamflow have the potential to outperform the dynamical ensembles by themselves, and could be designed and evaluated in future studies.

In terms of real-time forecasts, SubX as well as the WWRP/WCRP S2S project provides a testbed for research and a foundation for operational use (Pegion et al. 2019). The reforecast datasets have implications for operational forecasting by providing the basis for postprocessing methods (e.g., bias correction and calibration techniques) that provide adjustments to real-time predictions. Primary factors that impact forecast quality and ability to evaluate the performance of the hindcast in a forecast system configuration include hindcast period, ensemble size and ensemble strategy (e.g., initial times) (Merryfield et al. 2020). However, there are tradeoffs in the system configuration due to practical constraints. A few SubX models use more ensemble members in real-time forecasts than in reforecasts, which may enhance their real-time forecast skill. The planned second phase of the SubX project will adopt a more strict protocol to align forecast initialization dates for different models, which will produce a longer reforecast period and include a larger ensemble (Pegion et al. 2019). These improvements should benefit real-time operations as well. Our planned future work will continue to investigate applications of real-time subseasonal forecasts to flood prediction, likely using the second phase of SubX.

6. Conclusions

We examined the performance of SubX-driven forecasts of AR-related flooding in three watersheds along the coastal western

United States with leads from 1 to 4 weeks. We first evaluated SubX reforecasts of precipitation and temperature. After the statistical downscaling and bias correction of the forcings, we ran the DHSVM hydrologic model in each of the three basins, with a focus on reforecasting peak-over-threshold (POT) extreme discharge events in the period 1999–2016. We then evaluated both the deterministic and probabilistic skill of SubX-based flood forecasts. We further evaluated the relative influence of ASM and SubX reforecast skill to flood forecast skill using ESP and revESP experiments. Based on our analysis, we find the following:

1) SubX precipitation and temperature skill
   - Over all months (October–March), SubX precipitation forecast skill (as measured by anomaly correlation coefficient) drops quickly after week 1 lead, but still has usable skill at week 2, while at week 3–4, models show minimal skill (with positive ACC but not statistically different from zero). Generally, there is higher skill in temperature than precipitation forecasts, with all models showing usable skill through lead 4 weeks.
   - Across models, NCEP-CFSv2 performed best for both precipitation and temperature in weeks 1–2, with performance that is comparable with MME, while in weeks 3–4 EMC-GEFS performs best for temperature across the three basins.
2) SubX-based flood forecast skill
   - The deterministic forecast skill of NCEP-CFSv2 drops quickly with lead time, with little skill by lead days 14, 17, and 13 in the Chehalis, Russian, and Santa Margarita River basins, respectively, for the largest (POT_{N1}) events, and several days longer for small events (i.e., lower POT thresholds) in the first two basins.
   - SubX-based probabilistic forecast skill for extreme discharge events drops quickly after week 1, with minimal forecast skill by week 3 for the largest (POT_{N1}) events. Forecast skill is slightly higher for small events (i.e., lower POT thresholds), with minimal forecast skill by week 4.
3) Role of ASM in flood forecast
   - Comparing the influencing factors in flood forecast skill, SubX-based probabilistic flood forecast skill is generally lower for AR storms than for non-AR storms due to the generally larger magnitude of AR storms. However, forecast skill is influenced more strongly by ASM than by storm magnitude with lower skill when ASM is low, especially at shorter lead times and with lower POT thresholds (i.e., forecast skill is diminished less by ASM for large as contrasted with small storms).
   - In terms of the relative influence of ASM and SubX reforecast skill, ASM dominates streamflow deterministic forecast skill at leads up to 9 days with the maximum lead length occurring in October (following generally dry summers). ASM dominates flood probabilistic forecast skill only for small flood events in the three basins at week 1. For most large flood events in the three basins, the SubX reforecast skill dominates the flood probabilistic forecast skill at all weeks.

Acknowledgments. We acknowledge NOAA, NASA, and the U.S. Navy, the agencies that supported the SubX experiment and archive, and we thank the climate modeling groups (Environment Canada, NASA, NOAA/NCEP, NRL, and University of Miami) for producing and making available their model output. NOAA's Modeling, Analysis, Predictions, and Projections (MAPP) Program, the Office of Naval Research, NASA, and the NOAA National Weather Service jointly provided coordinating support and led development of the SubX system. We thank Mingyue Chen at NOAA for her help with our questions related to the NCEP-CFSv2 data. Access to the AR catalog was provided by Dr. Bin Guan of NASA's Jet Propulsion Laboratory via https://ucla.box.com/ARcatalog. Development of the AR detection algorithm used to produce the AR catalog (by Dr. Guan and colleagues) was supported by NASA. The authors thank Massimiliano Zappa at the Swiss Federal Research Institute WSL, and two anonymous reviewers whose comments improved the quality of our manuscript. The research support to coauthors of this paper was provided by the Center for Western Weather and Water Extremes (CW3E) at the Scripps Institution of Oceanography UC San Diego via AR Program Phase II, Grant 4600013361, sponsored by the California Department of Water Resources, and NOAA Regional Integrated Sciences and Assessments (RISA) support through the California–Nevada Applications Program (Grant NA17OAR4310284).

REFERENCES

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

Addor, N., S. Jaun, F. Fundel, and M. Zappa, 2011: An operational hydrological ensemble prediction system for the city of Zurich (Switzerland): Skill, case studies and scenarios. *Hydrol. Earth Syst. Sci.*, **15**, 2327–2347, https://doi.org/10.5194/hess-15-2327-2011.

Baggett, C. F., E. A. Barnes, E. D. Maloney, and B. D. Mundhenk, 2017: Advancing atmospheric river forecasts into subseasonal-to-seasonal time scales. *Geophys. Res. Lett.*, **44**, 7528–7536, https://doi.org/10.1002/2017GL074434.

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.

Barth, N. A., G. Villarini, M. A. Nayak, and K. White, 2017: Mixed populations and annual flood frequency estimates in the western United States: The role of atmospheric rivers. *Water Resour. Res.*, **53**, 257–269, https://doi.org/10.1002/2016WR019064.

Bartholmes, J. C., J. Thielen, M. H. Ramos, and S. Gentilini, 2009: The European Flood Alert System EFAS-Part 2: Statistical skill assessment of probabilistic and deterministic operational forecasts. *Hydrol. Earth Syst. Sci.*, **13**, 141–153, https://doi.org/10.5194/hess-13-141-2009.

Bohn, T. J., B. Livneh, J. W. Oyler, S. W. Running, B. Nijssen, and D. P. Lettenmaier, 2013: Global evaluation of MTCLIM and related algorithms for forcing of ecological and hydrological models. *Agric. For. Meteor.*, **176**, 38–49, https://doi.org/10.1016/j.agrformet.2013.03.003.

Cao, Q., A. Mehran, F. M. Ralph, and D. P. Lettenmaier, 2019: The role of hydrological initial conditions on atmospheric river floods in the Russian River basin. *J. Hydrometeor.*, **20**, 1667–1686, https://doi.org/10.1175/JHM-D-19-0030.1.

——, A. Gershunov, T. Shulgina, F. M. Ralph, N. Sun, and D. P. Lettenmaier, 2020: Floods due to atmospheric rivers along the

U.S. West Coast: The role of antecedent soil moisture in a warming climate. *J. Hydrometeor.*, **21**, 1827–1845, https://doi.org/10.1175/JHM-D-19-0242.1.

Day, G. N., 1985: Extended streamflow forecasting using NWSRFS. *Water Resour. Plann. Manage.*, **111**, 157–170, https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

DeFlorio, M. J., D. E. Waliser, B. Guan, D. A. Lavers, F. M. Ralph, and F. Vitart, 2018: Global assessment of atmospheric river prediction skill. *J. Hydrometeor.*, **19**, 409–426, https://doi.org/10.1175/JHM-D-17-0135.1.

——, ——, ——, F. M. Ralph, and F. Vitart, 2019a: Global evaluation of atmospheric river subseasonal prediction skill. *Climate Dyn.*, **52**, 3039–3060, https://doi.org/10.1007/s00382-018-4309-x.

——, and Coauthors, 2019b: Experimental subseasonal-to-seasonal (S2S) forecasting of atmospheric rivers over the western United States. *J. Geophys. Res. Atmos.*, **124**, 11 242–11 265, https://doi.org/10.1029/2019JD031200.

Dettinger, M., F. Ralph, T. Das, P. Neiman, and D. Cayan, 2011: Atmospheric rivers, floods and the water resources of California. *Water*, **3**, 445–478, https://doi.org/10.3390/w3020445.

Gershunov, A., T. Shulgina, F. M. Ralph, D. A. Lavers, and J. J. Rutz, 2017: Assessing the climate-scale variability of atmospheric rivers affecting western North America. *Geophys. Res. Lett.*, **44**, 7900–7908, https://doi.org/10.1002/2017GL074175.

——, and Coauthors, 2019: Precipitation regime change in western North America: The role of atmospheric rivers. *Sci. Rep.*, **9**, 9944, https://doi.org/10.1038/s41598-019-46169-w.

Gibson, P. B., D. E. Waliser, A. Goodman, M. J. DeFlorio, L. Delle Monache, and A. Molod, 2020: Subseasonal-to-seasonal hindcast skill assessment of ridging events related to drought over the western United States. *J. Geophys. Res. Atmos.*, **125**, e2020JD033655, https://doi.org/10.1029/2020JD033655.

Guan, B., and D. E. Waliser, 2015: Detection of atmospheric rivers: Evaluation and application of an algorithm for global studies. *J. Geophys. Res. Atmos.*, **120**, 12 514–12 535, https://doi.org/10.1002/2015JD024257.

Guan, H., Y. Zhu, E. Sinsky, W. Li, X. Zhou, D. Hou, C. Melhauser, and R. Wobus, 2019: Systematic error analysis and calibration of 2-m temperature for the NCEP GEFS reforecast of the Subseasonal Experiment (SubX) Project. *Wea. Forecasting*, **34**, 361–376, https://doi.org/10.1175/WAF-D-18-0100.1.

Gupta, H. V., H. Kling, K. K. Yilmaz, and G. F. Martinez, 2009: Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modeling. *J. Hydrol.*, **377**, 80–91, https://doi.org/10.1016/j.jhydrol.2009.08.003.

Hanley, J. A., and B. J. McNeil, 1982: The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiol.*, **143**, 29–36, https://doi.org/10.1148/radiology.143.1.7063747.

Infanti, J. M., and B. P. Kirtman, 2016: Prediction and predictability of land and atmosphere initialized CCSM4 climate forecasts over North America. *J. Geophys. Res. Atmos.*, **121**, 12 690–12 701, https://doi.org/10.1002/2016JD024932.

Kalnay, E., and Coauthors, 1996: The NCEP/NCAR 40-Year Reanalysis Project. *Bull. Amer. Meteor. Soc.*, **77**, 437–471, https://doi.org/10.1175/1520-0477(1996)077<0437:TNYRP>2.0.CO;2.

Knoben, W. J. M., J. E. Freer, and R. A. Woods, 2019: Technical note: Inherent benchmark or not? Comparing Nash Sutcliffe and Kling-Gupta efficiency scores. *Hydrol. Earth Syst. Sci.*, **23**, 4323–4331, https://doi.org/10.5194/hess-23-4323-2019.

——, ——, M. C. Peel, K. J. A. Fowler, and R. A. Woods, 2020: A brief analysis of conceptual model structure uncertainty using 36 models and 559 catchments. *Water Resour. Res.*, **56**, e2019WR025975, https://doi.org/10.1029/2019WR025975.

Konrad, C. P., and M. D. Dettinger, 2017: Flood runoff in relation to water vapor transport by atmospheric rivers over the western United States, 1949–2015. *Geophys. Res. Lett.*, **44**, 11 456–11 462, https://doi.org/10.1002/2017GL075399.

Koster, R. D., M. J. Suarez, A. Ducharne, M. Stieglitz, and P. Kumar, 2000: A catchment-based approach to modeling land surface processes in a general circulation model: 1. Model structure. *J. Geophys. Res.*, **105**, 24 809–24 822, https://doi.org/10.1029/2000JD900327..

Lang, M., T. Ouarda, and B. Bobee, 1999: Towards operational guidelines for over-threshold modeling. *J. Hydrol.*, **225**, 103–117, https://doi.org/10.1016/S0022-1694(99)00167-5.

Li, H., L. Luo, E. F. Wood, and J. Schaake, 2009: The role of initial conditions and forcing uncertainties in seasonal hydrologic forecasting. *J. Geophys. Res.*, **114**, D04114, https://doi.org/10.1029/2008JD010969.

Li, W., J. Chen, L. Li, H. Chen, B. Liu, C. Xu, and X. Li, 2019: Evaluation and bias correction of S2S precipitation for hydrological extremes. *J. Hydrometeor.*, **20**, 1887–1906, https://doi.org/10.1175/JHM-D-19-0042.1.

Liechti, K., L. Panziera, U. Germann, and M. Zappa, 2013: The potential of radar-based ensemble forecasts for flash-flood early warning in the southern Swiss Alps. *Hydrol. Earth Syst. Sci.*, **17**, 3853–3869, https://doi.org/10.5194/hess-17-3853-2013.

Lin, H., N. Gagnon, S. Beauregard, R. Muncaster, M. Markovic, B. Denis, and M. Charron, 2016: GEPS-based monthly prediction at the Canadian Meteorological Centre. *Mon. Wea. Rev.*, **144**, 4867–4883, https://doi.org/10.1175/MWR-D-16-0138.1.

——, R. Mo, F. Vitart, and C. Stan, 2018: Eastern Canada flooding 2017 and its subseasonal predictions. *Atmos.–Ocean*, **57**, 195–207, https://doi.org/10.1080/07055900.2018.1547679.

Livneh, B., E. A. Rosenberg, C. Lin, B. Nijssen, V. Mishra, K. M. Andreadis, E. P. Maurer, and D. P. Lettenmaier, 2013: A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *J. Climate*, **26**, 9384–9392, https://doi.org/10.1175/JCLI-D-12-00508.1.

Mahanama, S. P. P., R. D. Koster, R. H. Reichle, and L. Zubair, 2008: The role of soil moisture initialization in subseasonal and seasonal streamflow prediction: A case study in Sri Lanka. *Adv. Water Resour.*, **31**, 1333–1343, https://doi.org/10.1016/j.advwatres.2008.06.004.

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.

——, 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.

Merryfield, W. J., and Coauthors, 2020: Current and emerging developments in subseasonal to decadal prediction. *Bull. Amer. Meteor. Soc.*, **101**, E869–E896, https://doi.org/10.1175/BAMS-D-19-0037.1.

Molod, A., L. Takacs, M. J. Suarez, J. Bacmeister, I.-S. Song, and A. Eichmann, 2012: The GEOS-5 atmospheric general circulation model: Mean climate and development from MERRA to Fortuna. Tech. Memo. NASA/TM-2012-104606, 115 pp., https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20120011790.pdf.

Monhart, S., C. Spirig, J. Bhend, K. Bogner, C. Schär, and M. A. Liniger, 2018: Skill of subseasonal forecasts in Europe: Effect of bias correction and downscaling using surface observations. *J. Geophys. Res. Atmos.*, **123**, 7999–8016, https://doi.org/10.1029/2017JD027923.

——, M. Zappa, C. Spirig, C. Schär, and K. Bogner, 2019: Subseasonal hydrometeorological ensemble predictions in small- and medium-sized mountainous catchments: Benefits of the NWP approach. *Hydrol. Earth Syst. Sci.*, **23**, 493–513, https://doi.org/10.5194/hess-23-493-2019.

Mundhenk, B. D., E. A. Barnes, E. D. Maloney, and C. F. Baggett, 2018: Skillful empirical subseasonal prediction of landfalling atmospheric river activity using the Madden-Julian oscillation and quasi-biennial oscillation. *npj Climate Atmos. Sci.*, **1**, 20177, https://doi.org/10.1038/s41612-017-0008-2.

Nardi, K. M., E. A. Barnes, and F. M. Ralph, 2018: Assessment of numerical weather prediction model reforecasts of the occurrence, intensity, and location of atmospheric rivers along the West Coast of North America. *Mon. Wea. Rev.*, **146**, 3343–3362, https://doi.org/10.1175/MWR-D-18-0060.1.

Neiman, P. J., L. J. Schick, F. M. Ralph, M. Hughes, and G. A. Wick, 2011: Flooding in western Washington: The connection to atmospheric rivers. *J. Hydrometeor.*, **12**, 1337–1358, https://doi.org/10.1175/2011JHM1358.1.

Pan, B., K. Hsu, A. AghaKouchak, S. Sorooshian, and W. Higgins, 2019: Precipitation prediction skill for the West Coast United States: From short to extended range. *J. Climate*, **32**, 161–182, https://doi.org/10.1175/JCLI-D-18-0355.1.

Pegion, K., and Coauthors, 2019: The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. *Bull. Amer. Meteor. Soc.*, **100**, 2043–2060, https://doi.org/10.1175/BAMS-D-18-0270.1.

Pierce, D. W., D. R. Cayan, and B. L. Thrasher, 2014: Statistical downscaling using Localized Constructed Analogs (LOCA). *J. Hydrometeor.*, **15**, 2558–2585, https://doi.org/10.1175/JHM-D-14-0082.1.

Ralph, F. M., P. J. Neiman, G. Wick, S. Gutman, M. Dettinger, D. Cayan, and A. White, 2006: Flooding on California's Russian River: Role of atmospheric rivers. *Geophys. Res. Lett.*, **33**, L13801, https://doi.org/10.1029/2006GL026689.

——, J. J. Rutz, J. M. Cordeira, M. D. Dettinger, M. L. Anderson, D. Reynolds, L. J. Schick, and C. Smallcomb, 2019: A Scale to characterize the strength and impacts of atmospheric rivers. *Bull. Amer. Meteor. Soc.*, **100**, 269–289, https://doi.org/10.1175/BAMS-D-18-0023.1.

Reichle, R., and Q. Liu, 2014: Observation-corrected precipitation estimates in GEOS-5. Tech. Memo. NASA/TM-2014-104606, 18 pp., https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20150000725.pdf.

Rienecker, M. M., and Coauthors, 2008: The GEOS-5 Data assimilation system-documentation of versions 5.0.1, 5.1.0, and 5.2.0. Tech. Memo. NASA/TM-2008-104606, 97 pp., https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20120011955.pdf.

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.

Schick, S., O. Rössler, and R. Weingartner, 2019: An evaluation of model output statistics for subseasonal streamflow forecasting in European catchments. *J. Hydrometeor.*, **20**, 1399–1416, https://doi.org/10.1175/JHM-D-18-0195.1.

Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States: Understanding the role of initial hydrologic conditions and seasonal climate forecast skill. *Hydrol. Earth Syst. Sci.*, **15**, 3529–3538, https://doi.org/10.5194/hess-15-3529-2011.

——, N. Voisin, and D. P. Lettenmaier, 2012: Value of medium range weather forecasts in the improvement of seasonal hydrologic prediction skill. *Hydrol. Earth Syst. Sci.*, **16**, 2825–2838, https://doi.org/10.5194/hess-16-2825-2012.

Su, L., Q. Cao, M. Xiao, D.M. Mocko, M. Barlage, D. Li, C.D. Peters-Lidard, and D.P. Lettenmaier, 2021: Drought variability over the conterminous United States for the past century. *J. Hydrometeor.*, **22**, 1153–1168, https://doi.org/10.1175/JHM-D-20-0158.1.

Sun, S., R. Bleck, S. G. Benjamin, B. W. Green, and G. A. Grell, 2018a: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part I: Model overview and evaluation of systematic errors. *Mon. Wea. Rev.*, **146**, 1601–1617, https://doi.org/10.1175/MWR-D-18-0006.1.

——, B. W. Green, R. Bleck, and S. G. Benjamin, 2018b: Subseasonal forecasting with an icosahedral, vertically quasi-Lagrangian coupled model. Part II: Probabilistic and deterministic forecast skill. *Mon. Wea. Rev.*, **146**, 1619–1639, https://doi.org/10.1175/MWR-D-18-0007.1.

USWRC, 1982: Guidelines for determining flood flow frequency. Bulletin 17B of the Hydrology Subcommittee, 183 pp., https://water.usgs.gov/osw/bulletin17b/dl_flow.pdf.

Vitart, F., 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.

Weigel, A. P., M. A. Liniger, and C. Appenzeller, 2007: The discrete Brier and ranked probability skill scores. *Mon. Wea. Rev.*, **135**, 118–124, https://doi.org/10.1175/MWR3280.1.

Westra, S., R. Mehrotra, A. Sharma, and R. Srikanthan, 2012: Continuous rainfall simulation: 1—A regionalised sub-daily disaggregation approach. *Water Resour. Res.*, **48**, W01535, https://doi.org/10.1029/2011WR010489.

Wigmosta, M. S., L. W. Vail, and D. P. Lettenmaier, 1994: A distributed hydrology-vegetation model for complex terrain. *Water Resour. Res.*, **30**, 1665–1679, https://doi.org/10.1029/94WR00436.

Wilks, D. S., 2006: *Statistical Methods in the Atmospheric Sciences*. 2nd ed. International Geophysics Series, Vol. 100, Academic Press, 648 pp.

Wood, A. W., and D. P. Lettenmaier, 2008: An ensemble approach for attribution of hydrologic prediction uncertainty. *Geophys. Res. Lett.*, **35**, L14401, https://doi.org/10.1029/2008GL034648.

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

——, 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.

——, A. Kumar, and D. P. Lettenmaier, 2005: A retrospective assessment of National Centers for Environmental Prediction climate model–based ensemble hydrologic forecasting in the western United States. *J. Geophys. Res.*, **110**, D04105, https://doi.org/10.1029/2004JD004508.

Zhou, X., Y. Zhu, D. Hou, and D. Kleist, 2016: A comparison of perturbations from an ensemble transform and an ensemble Kalman filter for the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 2057–2074, https://doi.org/10.1175/WAF-D-16-0109.1.

——, ——, ——, Y. Luo, J. Peng, and R. Wobus, 2017: Performance of the new NCEP Global Ensemble Forecast System in a parallel experiment. *Wea. Forecasting*, **32**, 1989–2004, https://doi.org/10.1175/WAF-D-17-0023.1.

Zhu, Y., and Coauthors, 2018: Toward the improvement of subseasonal prediction in the national centers for environmental prediction global ensemble forecast system. *J. Geophys. Res.*, **123**, 6732–6745, https://doi.org/10.1029/2018JD028506.