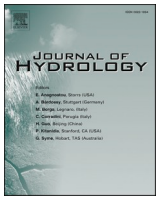
Journal of Hydrology 631 (2024) 130643

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



Adapting subseasonal-to-seasonal (S2S) precipitation forecast at watersheds for hydrologic ensemble streamflow forecasting with a machine learning-based post-processing approach

Lujun Zhang [a], Shang Gao [a,b], Tiantian Yang [a,*]

[a] School of Civil Engineering and Environmental Science, University of Oklahoma, Norman, OK, United States
[b] School of Natural Resources and the Environment, University of Arizona, Tucson, AZ, United States

ARTICLE INFO

This manuscript was handled by Yuefei Huang, Editor-in-Chief, with the assistance of Jing-Cheng Han, Associate Editor

Keywords:
Subseasonal-to-Seasonal
S2S
Precipitation
Forecast adaptation
Streamflow forecasting
Ensemble Streamflow Prediction

ABSTRACT

Accurate and reliable precipitation predictions made by dynamical forecast models could provide crucial information for human socioeconomic activities by enabling hydrologic forecasts at the Subseasonal-to-Seasonal (S2S) timescale. To utilize available S2S precipitation predictions for hydrologic forecasts, post-processing techniques have been applied to adapt the raw S2S precipitation to local watersheds. However, conventional statistical-based post-processing techniques are more focused on correcting the forecast bias, but rather limited in improving the predictive skill of available S2S precipitation forecasts. In this study, we combine the Random Forest classifiers (RF) with the Bias Correction and Spatial Disaggregation (BCSD) to adapt the 10-member ensemble precipitation forecast from the NASA Goddard Earth Observing System model version 5 (GEOS5) at 4 watersheds located in the NCEI South climate region of the United States. The adapted S2S precipitation is further applied for streamflow forecast by forcing a classical lumped hydrologic model. The performance of S2S precipitation as well as the corresponding streamflow predictions are benchmarked with the randomly resampled precipitation and the corresponding Ensemble Streamflow Prediction (ESP) framework-generated streamflow predictions. Evaluation statistics of Kling-Gupta Efficiency (KGE), Continuous Ranked Probability Skill Score (CRPSS), Reliability, Resolution, and Sharpness are employed to evaluate the predictive skill of precipitation and streamflow both deterministically and probabilistically. Our results indicate that dynamical S2S precipitation after forecast adaptation leads to consistently higher deterministic skill over ESP at all forecast lead times and across study watersheds. However, at longer forecast lead times beyond 10–15 days, S2S precipitation with a limited ensemble size does not present higher probabilistic skill than ESP. Our results shows that the joint application of RF and BCSD improves the predictive skill of the raw S2S precipitation at study watersheds in contrast to BCSD. Further, the added predictive skill of S2S precipitation brought by RF propagates into streamflow predictions, predominantly at longer forecast lead times exceeding 10 days. Overall, our results highlight the potential success of future work to apply other data-driven approaches to adapt the raw precipitation to local watersheds for more accurate and reliable streamflow forecasts at the S2S timescale.

1. Introduction

Accurate and reliable streamflow predictions at a Subseasonal-to-Seasonal (S2S) timescale, which spans from 10 to 30 days into the future (Vitart et al., 2017), could provide crucial information benefiting human socio-economic activities of various kinds (Chiew et al., 2003; White et al., 2017; Yang et al., 2021). The Ensemble Streamflow Prediction (ESP) framework is commonly adopted for streamflow forecasting at the S2S timescale (Day, 1985; Schaake and Larson, 1998). There are mainly two key steps to apply ESP for hydrologic forecasts at the S2S timescale. First, a calibrated hydrologic model is forced with measured hydrometeorological information up to the "current" time step to estimate the initial hydrologic conditions (IHCs). Then, with the estimated IHCs, the calibrated hydrologic model is further forced with multiple randomly resampled historical precipitation measurements to predict streamflow volumes probabilistically over the forecast horizon of interest. While exploiting our current understanding of hydrology through the application of a chosen hydrologic model, the ESP considers

* Corresponding author at: 202 W. Boyd St., Room 334, Norman, OK 73019, United States.
E-mail address: tiantian.yang@ou.edu (T. Yang).

https://doi.org/10.1016/j.jhydrol.2024.130643
Received 26 August 2023; Received in revised form 8 December 2023; Accepted 26 December 2023
Available online 22 January 2024
0022-1694/© 2024 Elsevier B.V. All rights reserved.

the predictability of streamflow from both the IHCs (i.e., initial soil moisture and/or snow conditions) and future hydrometeorological conditions. By incorporating an ensemble of resampled historical precipitation measurements, probabilistic predictions on future extreme events (i.e., floods) can be achieved (Delaney et al., 2020; Harrigan et al., 2018; Troin et al., 2021).

However, one common challenge persists when predicting streamflow at the S2S timescale under the ESP framework. That is when the forecast lead time is long enough (typically beyond 5–10 days), the predictability of streamflow coming from the IHCs would be lost, leaving the volumetric prediction of streamflow dominated by inaccurate resampled precipitation forecasts. As a result, the corresponding streamflow predictions normally present little deterministic predictive skill (Cao et al., 2021; Shukla et al., 2013; Wood and Lettenmaier, 2008).

Advancements in weather/climate predictions have provided an opportunity to advance ensemble streamflow forecasting at the S2S timescale. Many dynamical S2S precipitation forecast products have been made available by different mission agencies and/or through international collaborations over the globe in the recent decade (Kirtman et al., 2014; Pegion et al., 2019; Richter et al., 2022; Vitart et al., 2017). These dynamical S2S precipitation forecast products are generated by General Circulation Models (GCMs) coupled with dynamical land surface and oceanic components with a consideration of the "real-time" state of the Earth System. Therefore, available S2S precipitation forecast should be more reliable and accurate compared to the randomly resampled precipitation and could be used to force a hydrological model and enable climate-model-based hydrological forecasting for more accurate streamflow predictions (Liu et al., 2022; Ma and Yuan, 2023; Yuan, 2016; Yuan et al., 2013).

However, according to several previous evaluation studies, available S2S precipitation forecasts are commonly associated with a substantial amount of forecast bias and only present a marginal level of predictive skill when forecast lead time exceeds 2 weeks (de Andrade et al., 2021; Tian et al., 2017; Zhang et al., 2021). Combining the fact that the raw S2S precipitation forecasts often come with a spatial resolution coarser than ~ 32 km that exceeds the conventional hydrologic scale, post-processing techniques need to be applied to adapt the raw S2S precipitation to a typical watershed scale before any potential hydrologic applications.

To enhance the utilization of available S2S precipitation for hydrological predictions, various statistical-based forecast adaptation techniques are frequently applied (Wood et al., 2004). However, these available statistical-based techniques often prioritize the removal of forecast bias, potentially resulting in either unchanged or diminished overall predictive skill of S2S precipitation (Cao et al., 2021; Su et al., 2023; Wood and Lettenmaier, 2008). As a result, there remains a continuous demand for novel forecast adaptation techniques for S2S precipitation forecasts and the corresponding hydrological applications.

Machine Learning (ML)-based techniques are promising alternatives to conventional post-processing techniques such as BCSD to further improve the quality of available S2S forecasts through post-processing. Many recent studies have demonstrated the effectiveness of ML in the field of hydrometeorology for various forecast-related applications (Kim et al., 2021; Li et al., 2022; Liu et al., 2022; Zhu et al., 2023). Most recently, Zhang et al. (2023b) applied the Random Forest classifiers (RF) to combine additional forecast variables to improve the predictive skill of the S2S precipitation forecasts from NASA Goddard Earth Observing System Model Version 5 (i.e., GEOS5, one contributing model to the NMME-2 project) over the CONUS. According to Zhang et al. (2023b), the proposed RF improved the capability of the S2S forecasts from GEOS5 in forecasting weekly extreme precipitation over the CONUS. This highlights the potential success of various data-driven ML applications for more skillful S2S precipitation forecasts, which could enable more accurate and reliable streamflow forecasts at the S2S timescale.

However, the RF framework proposed by Zhang et al. (2023b) is due for a verification of its effectiveness at watersheds for climate-model-based hydrological forecasting approach for primarily two reasons. Firstly, Zhang et al. (2023b) carried out over the entire CONUS, which exceeds the conventional hydrologic scale. Further, in Zhang et al. (2023b), only categorical information was provided with the application of RF but without any volumetric information (which is critically needed for hydrologic forecasting). It is thus unclear whether the proposed RF can be applied at local watersheds for climate-model-based S2S streamflow predictions.

In this study, we aim to test and verify the effectiveness of ML in adapting raw S2S precipitation for daily streamflow predictions for the climate-model-based streamflow forecasting approach. We reckon that the successful application of ML-adapted S2S precipitation at the smallest water resource management unit could better demonstrate the potential effectiveness of available S2S precipitation in streamflow forecasting. Therefore, we carried out a series of streamflow hindcast experiments at several USGS HUC8 level watersheds, which are often considered to be the smallest spatial units of water resources management in the United States (Jones et al., 2022) via addressing research questions as follows: 1) Can an ML-based forecast adaptation technique improve the predictive skill of daily S2S precipitation forecasts at the HUC-8 scale? 2) Can the ML-adapted S2S precipitation lead to overall more skillful streamflow prediction than a baseline post-processing technique? And 3) If ML indeed shows improved predictive skills, how does the added skill of the S2S precipitation forecast propagate into streamflow predictions based on the climate-model-based streamflow forecasting approach at various forecast lead times?

To answer the above-mentioned research questions and extend the previous works by Zhang et al. (2023b), we apply RF at watersheds for climate-model-based streamflow forecasting. To be more specific, the RF framework proposed by Zhang et al. (2023b) is jointly employed with BCSD to adapt the raw S2S precipitation from GEOS5 to 4 watersheds in the NCEI South climate region where S2S precipitation appears to perform consistently worse than it is in other climate regions (Zhang et al., 2021). The underlying logic of selecting these 4 study watersheds in the NCEI South climate region is that if the proposed approach is proven to be effective at the selected watersheds, then it should be more likely to show at least the same level of effectiveness when transferred to other watersheds where S2S precipitation predictions are more accurate. The adapted S2S precipitation is further input to a lumped hydrologic model for streamflow forecasting experiments at the selected study watersheds to verify its hydrologic performances. In this study, the classical ESP enabled by the randomly resampled precipitation as well as BCSD-adapted S2S precipitation are treated as two baseline approaches to benchmark the performance of the proposed RF in S2S precipitation forecast adaptation and the corresponding streamflow forecasting. In addition, the "elasticities" between the added skill of precipitation and the corresponding added skill of streamflow forecasts brought by RF are computed for quantification and interpretation. This study covers a period from 1982 to 2011 (30 years).

In this study, multiple RFs are trained for categorical predictions on the occurrence of high-percentile extreme precipitation. The BCSD is applied to further correct the volumetric information of the S2S precipitation after RF-based corrections. There are primarily two reasons for such a design. The first reason is that previous works show ML techniques tend to underestimate high percentile rare values (Akbari Asanjan et al., 2018; Kim et al., 2022). Another perhaps more important reason is that we do not have a sufficient number of input-measurement pairs but only a rather limited number of hindcast experiments. We believe that, under this circumstance, the application of RF could be prone to overfitting if directly targeted on numerical values during the training of RF (Caruana et al., 2000; Karras et al., 2020). Therefore, we train RF for categorical predictions on the occurrence of high-percentile extreme precipitation first and further correct the volumetric information of precipitation with BCSD.

The remainder of this paper contains the following contents: Section 2 introduces datasets and study watersheds. Section 3 describes the

general experiment design, methodology, and evaluation statistics of this study. The results are presented in Section 4. Section 5 and Section 6 present and summarize the discussions and main conclusions, respectively.

2. Datasets and study regions

2.1. Datasets

Three measurement/reanalysis hydrometeorological datasets and one S2S forecast product used in this study including 1) the AN81d precipitation from the Parameter-elevation Regressions on the Independent Slopes Model (PRISM), 2) Daily potential evapotranspiration (PET) reanalysis dataset from the North American Regional Reanalysis (NARR), 3) Daily streamflow measurements from the National Water Information System at the United States Geological Survey (USGS), and 4) Multiple S2S hindcast variables from the NASA Goddard Earth Observing System Model version 5 (GEOS5), as one model member in the NMME-2 data archive. All datasets used in this study cover a common, 30-year period from 01/01/1982 to 12/31/2011.

Multiple S2S hindcast variables of precipitation, temperature, and 500-hPa and 850-hPa geopotential height from GEOS5 are used in this study as well. The GEOS5 produces ensemble forecasts/hindcasts containing 10 members by perturbing initial conditions (Borovikov et al., 2019). The collected S2S hindcasts have a monthly frequency that was initialized on the first day of the month throughout the entire study period. Although the collected S2S hindcasts span up to 274 days over the forecast horizon, we focus on the first 28 days of each collected hindcast as this study is specifically targeted at the S2S timescale. More details about the datasets used in this study are presented in Table 1.

2.2. Target watersheds

In this study, the same 4 watersheds in the South climate region of the United States (per NCEI definition) studied by Zhang et al., (2023a) are selected as the study watersheds. Fig. 1 shows the general information of the study watershed including locations, elevation, river networks, and streamflow gauge locations. The selected watersheds are considered less affected by human activities, which makes them suitable study watersheds in the region (Newman et al., 2015).

The selected study watersheds are dry. According to the collected PRSIM dataset, the 30-year averaged daily precipitation ranges from 3.2 mm/day to 4.2 mm/day across the study watersheds without strong seasonal patterns. In brief, precipitation in the summer seasons of June, July, and August accounts for less annual rainfall than that in other seasons at the BC, SC, and CR watersheds. Whereas precipitation at the BP watershed shows slightly higher values in the Winter and Spring seasons from December through May. The observed precipitation is distributed quite uniformly over the 4 study watersheds without significant spatial clustering (More detailed information in terms of the spatial and seasonal patterns of the climatology of precipitation at the 4 study watersheds is presented in the Supplementary material). Other relevant information about the study watersheds is presented in Table 2.

3. Experiment design, methodology, and evaluation statistics

3.1. General experiment design

Two forecast adaptation schemes are tested to adapt the raw S2S precipitation forecasts at the 4 selected watersheds and evaluated against the randomly resampled precipitation under the classical ESP. The first post-processing scheme (i.e., the baseline scheme) is the BCSD post-processing. To apply BCSD, the raw precipitation forecasts are bilinearly interpolated into 0.07-degree pixels first (i.e., the spatial resolution of PRISM). Then, bias corrections are conducted at each interpolated pixel using Multi-Segmented Quantile Mapping (i.e., MSQM, originally introduced by Grillakis et al. (2013)). The second post-processing scheme is a combination of BCSD and RF (hereafter referred to as "RF-BCSD"). Under the second post-processing scheme, the raw precipitation forecasts are also bilinearly interpolated into 0.07-degree pixels first. Only this time the RF is applied to correct the interpolated precipitation forecasts at each interpolated pixel first before the application of MSQM to remove forecast bias. Precipitation adapted under both schemes is eventually transformed into mean areal values over the study watersheds for evaluation as well as next-step experiments. More details about the two post-processing schemes of BCSD and RF-BCSD are introduced in the following section 3.2.

The adapted S2S precipitation is then applied for streamflow hindcast experiments following the classical ESP framework. The performance of streamflow predictions associated with the adapted S2S precipitation is benchmarked with the randomly resampled precipitation-enabled ESP streamflow predictions. The inclusion of the streamflow hindcasts generated by ESP could enable more informative comparisons between S2S precipitation-forced and conventional ESP-generated streamflow forecasting. To keep this study focused, more detailed information about the technical steps of executing S2S precipitation forced streamflow forecasts and ESP is included in the Supplementary material for interesting readers.

The streamflow hindcasts resulted from ESP and the adapted S2S precipitation are evaluated against the proxies of streamflow measurements (i.e., streamflow simulations generated by forcing the calibrated hydrologic model with precipitation and PET measurements). The usage of such streamflow proxies instead of measured streamflow during the evaluation is for later quantification of the propagation of the added skill from precipitation to streamflow, without mixing the effects coming from other sources, e.g., the imperfect structure of the hydrologic model, uncertainty arises from parameters of the hydrologic model, etc. The overall experiment design is depicted in Fig. 2.

In this study, we choose the Sacramental Soil Moisture Accounting (Sac-SMA) as the hydrologic model for streamflow hindcast experiments. The Sac-SMA is calibrated at each study watershed using the SCE-SAHEL algorithm developed by Naeini et al. (2018). Since there are three groups of precipitation inputs (S2S precipitation from two post-processing scenarios, as well as the resampled precipitation) to be

Table 1
Hydrometeorological datasets used in this study.

	Name	Temporal resolution	Spatial resolution	Data sources
Measurements /Reanalysis	Precipitation (PRISM)	Daily	~0.07°	https://prism.oregonstate.edu/
	PET (NARR)		~0.3°	https://www.emc.ncep.noaa.gov/mmb/rreanl/
	Streamflow (USGS)		N/A	https://waterdata.usgs.gov/nwis
GEOS5 S2S hindcasts	Precipitation		1°	https://www.cpc.ncep.noaa.gov/products/NMME/data.html
	Temperature		1°	
	500 hPa Geopotential		1°	
	850 hPa Geopotential		1°	

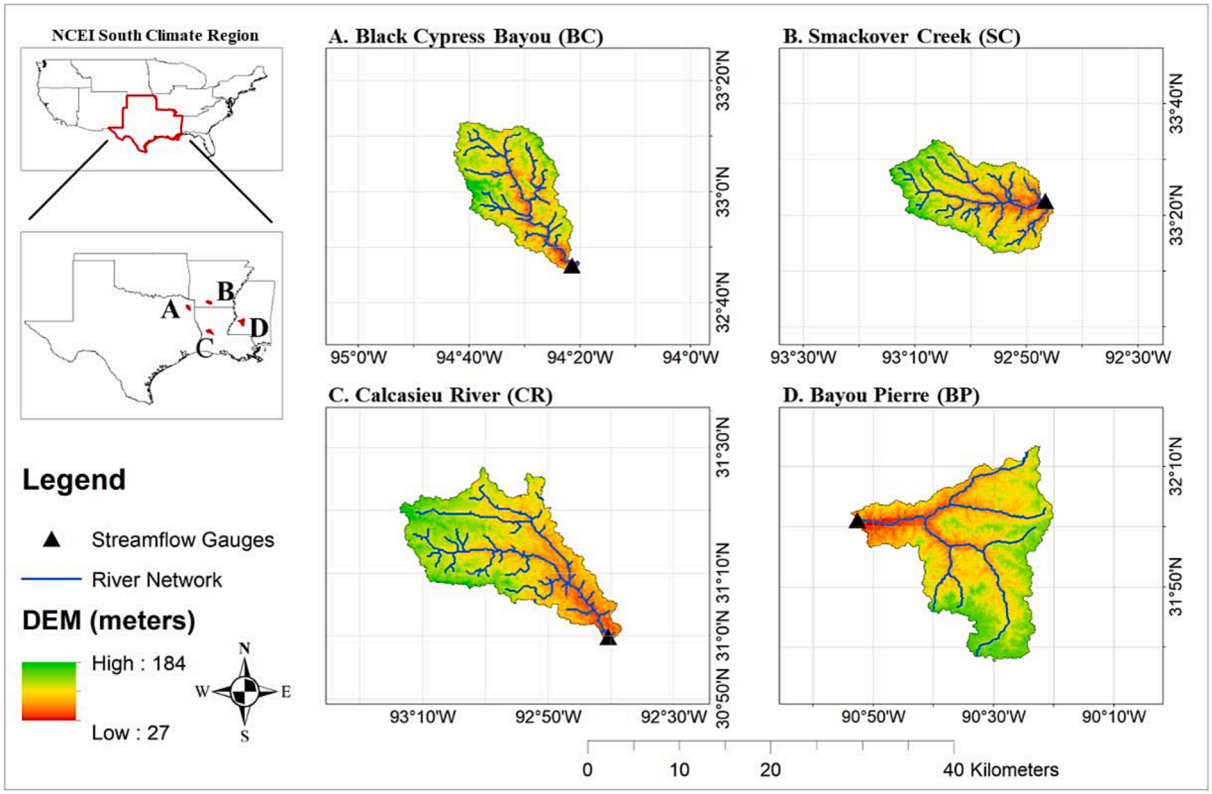


Fig. 1. Map of the 4 study watersheds (BC is in Texas; SC is in Arkansas; CR is in Louisiana; BP is in Missouri).

Table 2
Topography and hydrometeorology information of the study watersheds.

Watershed Names	Drainage Area (Square Kilometers)	Mean Elevation (Meters)	Aridity (PET/P)	Snow Fraction	USGS Gauge ID
Black Cypress Bayou (BC)	960	105.07	0.88	0.017	7,346,045
Smackover Creek (SC)	996	79.29	0.80	0.018	7,362,100
Calcasieu River (CR)	1295	81.79	0.75	0.003	8,013,000
Bayou Pierre (BP)	1689	102.05	0.82	0.004	7,290,650

applied for S2S streamflow forecasting, a total number of 1080 streamflow hindcast experiments, i.e., 3 groups of precipitation × 12 months × 30 years, are conducted at each of the study watersheds. Subject to the frequency of the collected S2S precipitation forecasts, the resulting streamflow hindcasts also have a monthly frequency and are initialized on the 1st day of the months during the entire study period.

The evaluation of precipitation in this study is carried out with Kling-Gupta Efficiency (KGE) and the Continuous Ranked Probability Skill Score (CRPSS) at different forecast lead times and across different study watersheds. In addition, the reliability, resolution, and sharpness of the adapted S2S precipitation were computed to gain a better understanding of the corrected ensemble forecasting system. The evaluation of the resulting streamflow forecasts in this study is carried out with the Kling-Gupta Efficiency (KGE) and the Continuous Ranked Probability Skill (CRPS). Since it is expected that the proposed RF-BCSD should bring overall more skillful precipitation as well as streamflow forecasts than BCSD, the "elasticity" values between the added skill of precipitation and corresponding streamflow prediction at different forecast lead times are computed. More details about the aforementioned evaluation statistics are described in section 3.3.

3.2. S2S forecast adaptation schemes

In this study, we employ two S2S forecast adaptation schemes. For both forecast adaptation schemes, we adopted the same cross-correction strategy to correct the entire record of the collected S2S precipitation. To do that, all collected datasets are equally divided into three 10-year groups (i.e., 1982–1991, 1992–2001, and 2002–2011). Each 10-year period is treated as the correction period in rotation, while the corresponding remaining 20-year period is treated as a reference period to train correction models under different forecast adaptation schemes (i.e., RF-BCSD or BCSD). During the post-processing/correction under both schemes, seasonality and different forecast lead times are considered as well. Specifically, the S2S data and PRISM from a certain reference period are divided into different groups based on 4 weeks of forecast lead times (i.e., 1 to 7 days, 8 to 14 days, 15 to 21 days, and 22 to 28 days) and 4 different seasons (i.e., December, January, February; March, April, May; June July August; and September, October, November) to train different correction models. The obtained correction model is then applied to correct the raw S2S precipitation from the correction period for a certain forecast lead time and within a certain season. The post-processed S2S precipitation under both schemes is eventually transformed into mean areal values for evaluation and the subsequent streamflow hindcast experiments.

Under the first forecast adaptation scheme, the slightly modified Bias Correction and Spatial Disaggregation (BCSD) was applied. There are generally two steps when applying BCSD to post-process precipitation forecasts. Firstly, the raw S2S precipitation is first bilinearly interpolated into 0.07° pixels (i.e., the spatial resolution of PRISM). Then, bias corrections are conducted at each interpolated 0.07° pixel using the Multi-

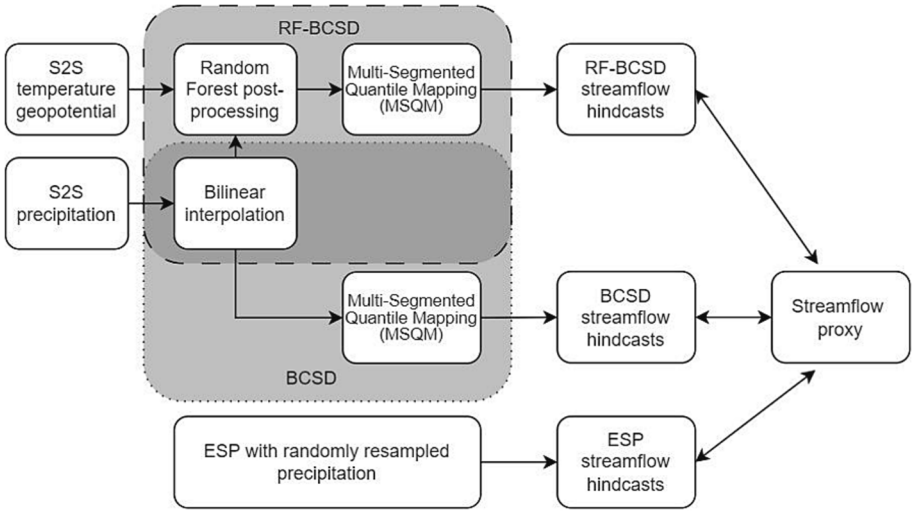


Fig. 2. A schematic diagram of the experiment design.

Segmented Quantile Mapping (MSQM). Given the popularity of the MSQM technique, more detailed technical steps of the MSQM bias correction are described in the Supplementary material to keep the paper concise.

In the second forecast adaptation scheme, a combination of RF and BCSD (i.e., RF-BCSD) was jointly employed. To be more specific, the raw S2S precipitation forecast is also bilinearly interpolated into 0.07-degree pixels first. However, before the MSQM bias correction, the RF is employed to correct the raw S2S precipitation for better categorical predictions on the high-percentile extreme events. The RF employed under the second forecast adaptation scheme is for the improvement of the predictive performance of S2S precipitation in forecasting the occurrence of daily extreme values. To be more specific, the RF is trained to make categorical predictions of either (1) no extreme precipitation, (2) extreme precipitation from 97 % to 98.5 %, (3) extreme precipitation from 98.5 to 99.5 %, (4) extreme precipitation from 99.5 % to 99.9 %, or (5) extreme precipitation above 99.9 %. The training/correction of RF was executed while considering different forecast lead times and with the same cross-validation strategy as explained in the previous section. When training the RF classifier, all collected S2S variables (i.e., precipitation, temperature, geopotential height at 500 hPa, and geopotential height at 850 hPa) are used as inputs. The application of RF in this study is realized through the Python-based open-source package of "scikit-learn". The hyperparameters of RF are set to values recommended by Zhang et al. (2023b), where the "max_depth", the "max_features", and the "n_estimations" are set to be 7, 0.6, and 200, respectively.

After the RF-based correction, the positive predictions on extreme precipitation are restored to numerical values based on the statistics of the raw S2S precipitation forecasts. For example, if the RF made a positive prediction on extreme precipitation within the percentile segment of 99.5 % to 99.9 % on date *X*, the median value of the quantile segments from 99.5 % to 99.9 % of the raw S2S precipitation will be placed on date *X* to replace the original forecast values. Finally, the RF-corrected S2S precipitation forecast with all numerical values is further corrected using the MSQM technique to thoroughly remove the forecast bias.

3.3. Evaluation statistics

In this study, evaluation is carried out from two perspectives to comprehensively quantify the predictive skill of the adapted S2S precipitation forecast both deterministically and probabilistic. Kling-Gupta Efficiency (KGE) and Correlation Coefficient (CC) are employed to evaluate the adapted S2S precipitation forecast as well as the randomly resampled precipitation deterministically. The Continuous Ranked Probability Skill Score (CRPSS) is employed to evaluate the adapted S2S precipitation from a probabilistic perspective. The reliability, resolution, and sharpness were further computed to specifically evaluate the performance of the adapted S2S precipitation on extreme precipitation events above 97 % subject to our RF-based corrections. For the resulting streamflow predictions associated with ESP and the adapted S2S precipitation, we applied KGE and CPRS for evaluation. More details about the methodology of employed evaluation statistics are described in the following sections 3.3.1, 3.3.2, and 3.3.3.

3.3.1. Kling-Gupta Efficiency (KGE) and correlation coefficient (CC)

The Kling-Gupta Efficiency (KGE) was originally introduced by Gupta et al. (2009) and has become a widely applied performance evaluation statistic in the field of hydrology ever since. The KGE is computed based on three distinct statistics of correlation coefficient (CC), Bias Ratio (BR) following equation (2): and relative variability (RV). The values of CC, BR, and RV can be computed with the following equations (2), (3), and (4):

$$CC = \frac{\sum_{i=1}^{n}\left(\left(Q_{Sim,i} - \overline{Q}_{Sim}\right) - \left(Q_{Obs,i} - \overline{Q}_{Obs}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(Q_{Sim,i} - \overline{Q}_{Sim}\right)^2 \sum_{i=1}^{n}\left(Q_{Obs,i} - \overline{Q}_{Obs}\right)^2}} \tag{2}$$

$$BR = \frac{\overline{Q}_{Sim}}{\overline{Q}_{Obs}} \tag{3}$$

$$RV = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{Sim,i} - \overline{Q}_{Sim}\right)^2}/\overline{Q}_{Sim}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(Q_{Obs,i} - \overline{Q}_{Obs}\right)^2}/\overline{Q}_{Obs}} \tag{4}$$

In equations (2), (3), and (4), n represents the total number of data points, $Q_{Sim,i}$ and $Q_{Obs,i}$ are the model simulated/predicted data and observed data, and $\overline{Q}_{Sim}$ and $\overline{Q}_{Obs}$ are the means of the model simulated/predicted and observed data series, respectively. With the computed values of CC, BR, and RV, KGE can be computed with the following equation (5):

$$KGE = 1 - \sqrt{(CC-1)^2 + (BR-1)^2 + (RV-1)^2} \tag{5}$$

The values of KGE have no unit and range from $-\infty$ to 1. The ideal value of KGE would be 1, indicating that model simulated/predicted values are perfectly aligned with measurements. KGE values below

−0.41 would indicate the simulated/predicted values are not skillful compared to climatology (Knoben et al., 2019).

3.3.2. Continuous Ranked Probability skill Score (CRPSS)

The Continuous Ranked Probability Skill Score (CRPSS) is a widely applied measure of how well the forecast probability of a prediction system matches with the observed outcomes. CRPSS is computed with the following equation (6):

$$CRPSS = 1 - CRPS_{fore}/CRPS_{clim} \tag{6}$$

In equation (6), $CRPS_{fore}$ is the Continuous Ranked Probability Score (CRPS) of a prediction system to be evaluated and the $CRPS_{clim}$ is the CRPS of the climatology. In this study, we compute $CRPS_{clim}$ based on the randomly resampled precipitation in ESP. The computed $CRPS_{clim}$ is then employed to compute CRPSS of the adapted S2S precipitation forecasts. CRPSS ranges from zero to one, with one indicating the forecast has perfect predictive skill in comparison to climatology.

The CRPS is calculated based on the empirical probability density function (PDF) of the ensemble forecast system X and the corresponding measurement Y following equation (7):

$$CRPS = \frac{1}{N}\sum_{n=1}^{N}\int [F_n(x) - F_n(y)]^2 dx \tag{7}$$

In equation (7), n represents a certain forecast case, $F_n(x)$ is the empirical probability density function (PDF) of the ensemble forecast system, $F_n(y)$ is the empirical PDF given a measured value. The unit of CRPS is subject to the variable to be evaluated.

3.3.3. Reliability, resolution, and sharpness of the adapted ensemble S2S precipitation forecast

It is well known that once forecast lead time exceeded the weather timescale, the dynamical S2S precipitation forecast would show very limited predictive skills. Therefore, it is important to quantify the usefulness of available ensemble predictions, especially for extreme precipitation events. In this study, we follow Yuan and Wood (2013) and Wilks (2011) to compute the reliability, resolution, and sharpness of the adapted ensemble S2S precipitation forecast. The calculation of the aforementioned statistics is specifically targeted at extreme precipitation events above 97 % subject to our previous RF-based forecast corrections. The reliability, resolution, and sharpness of the adapted S2S ensemble precipitation forecast are computed following equations (8), (9), and (10):

$$Reliability = \frac{1}{n}\sum_{i=1}^{E} N_i(f_i - \overline{y_i})^2 \tag{8}$$

$$Resolution = \frac{1}{n}\sum_{i=1}^{E} N_i(\overline{y_i} - \overline{y})^2 \tag{9}$$

$$Sharpness = \sqrt{\frac{1}{n}\sum_{i=1}^{E} N_i(f_i - \overline{y})^2} \tag{10}$$

In equations (8) (9) and (10), f_i is the overall forecast probability of the extreme event of the ensemble member i from an ensemble forecast system consisting of E ensemble members. $\overline{y_i}$ is the conditional probability of the extreme event that was observed given the forecast probability of ensemble member i. N_i is the number of cases where the ensemble member i has given a positive prediction on the extreme event. n is the total number of forecast instances. $\overline{y}$ is defined as $1/n\sum_{i=1}^{E} N_i\overline{y_i}$. Smaller reliability values and larger resolution values would indicate better probabilistic forecast. The sharpness values do not quantify the performance of the ensemble forecast system directly, but they are jointly considered with reliability and resolution values for a more comprehensive evaluation of the adapted S2S precipitation forecast.

3.3.4. The "elasticities" of the added skills between precipitation and streamflow

The "elasticities" are computed at different forecast lead times to quantity the propagation of the added skill from precipitation to streamflow hindcasts brought by RF-BCSD in contrast to BCSD. The computation of elasticity is defined as the following equation (11):

$$E = \Delta_{KGE}(Precipitation)/\Delta_{KGE}(Streamflow) \tag{11}$$

The skill elasticities are defined as the unit change of streamflow hindcasts skill in correspondence to the unit change of precipitation hindcast skill. Note that when computing skill elasticities, only KGE was used.

4. Results

4.1. Calibration of Sac-SMA

Fig. 3 presents the streamflow simulations at 4 study watersheds after the calibrations of Sac-SMA. The red lines indicate streamflow measurements, and the blue lines indicate the simulated streamflow. The vertical dashed lines separate the entire study period into the calibration and validation periods. The KGE values within calibration and validation periods at each study watershed are labeled on the figure as well.

From Fig. 3, it can be observed that after the calibration of Sac-SMA, the high-volume streamflow events seem to be consistently underestimated at all watersheds. But in general, KGE ranges from 0.84 to 0.69 at all watersheds during the calibration period. Although the performances of hydrologic simulations drop during the validation periods, the values of KGE remain above 0.6 at all study watersheds.

4.2. Predictive performance of precipitation

To examine the quality of the S2S precipitation resulting from the 2 forecast adaptation schemes, both forecast bias and predictive skill are examined. The forecast bias is examined with Quantile-Quantile plots (QQ-plots) and the predictive skill is quantified through the employment of KGE and CRPSS.

4.2.1. Forecast bias

In Fig. 4, the mean areal S2S precipitation hindcasts of all ensemble members from GEOS5 over the 4 study watersheds are plotted as scatter points against the reference values from the PRISM dataset at the same ranking/percentile (Quantile-to-Quantile plots). The 4 columns of Fig. 4 correspond to the 4 study watersheds. The 3 rows of Fig. 4 correspond to the raw GEOS5 S2S precipitation hindcasts (grey-colored dots), the BCSD-adapted S2S precipitation (black-colored dots), and the RF-BCSD adapted S2S precipitation (blue-colored dots), respectively. The ideal S2S precipitation hindcast values would lie perfectly on the red-colored 45-degree lines, which indicates that there's no bias of S2S precipitation at all magnitudes.

From Fig. 4, differences can be observed in terms of the climatology of precipitation. The maximum precipitation ranges from around 120 mm (BC) to 200 mm (CR) across the 4 study watersheds. The raw S2S hindcasts from GEOS5 underestimate precipitation across all study watersheds, as all the grey-colored scatter points lie above the 45-degree reference lines. The application of BCSD removed the forecast bias effectively at all magnitudes and across all study watersheds, as most of the black-colored scatter points aligned with the 45-degree reference line much better if compared to the grey-colored scatter points. On the other hand, the joint application of RF and BCSD also removes forecast bias effectively. No major differences can be told if comparing second and third-row panels, which indicates that the application of RF does not affect the removal of forecast bias.

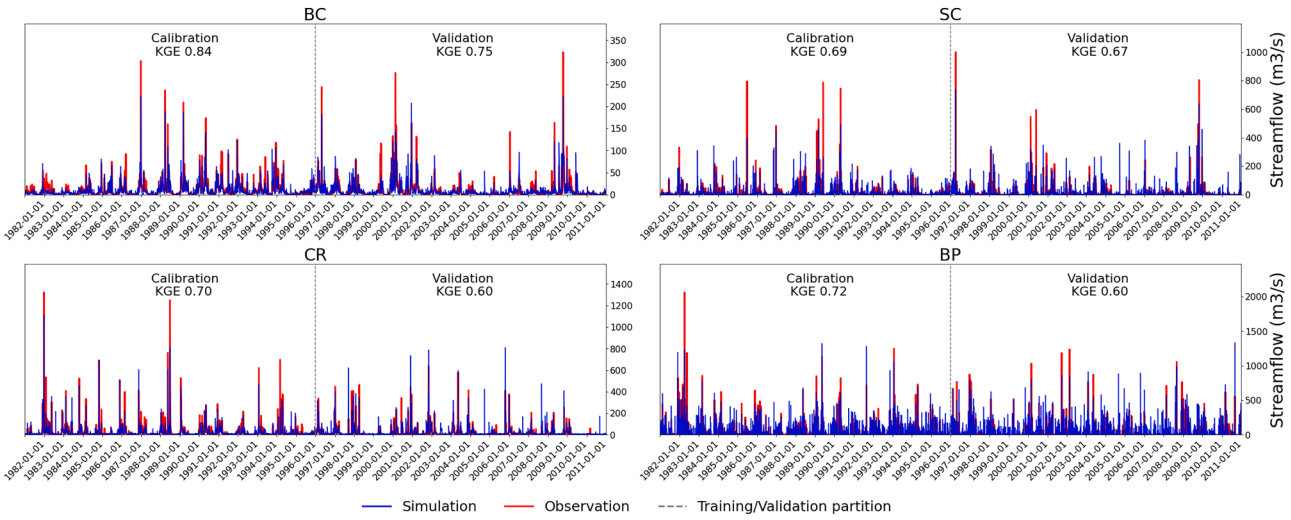


Fig. 3. Simulated and observed daily hydrographs at four study watersheds after the parameter calibration of the Sac-SMA.

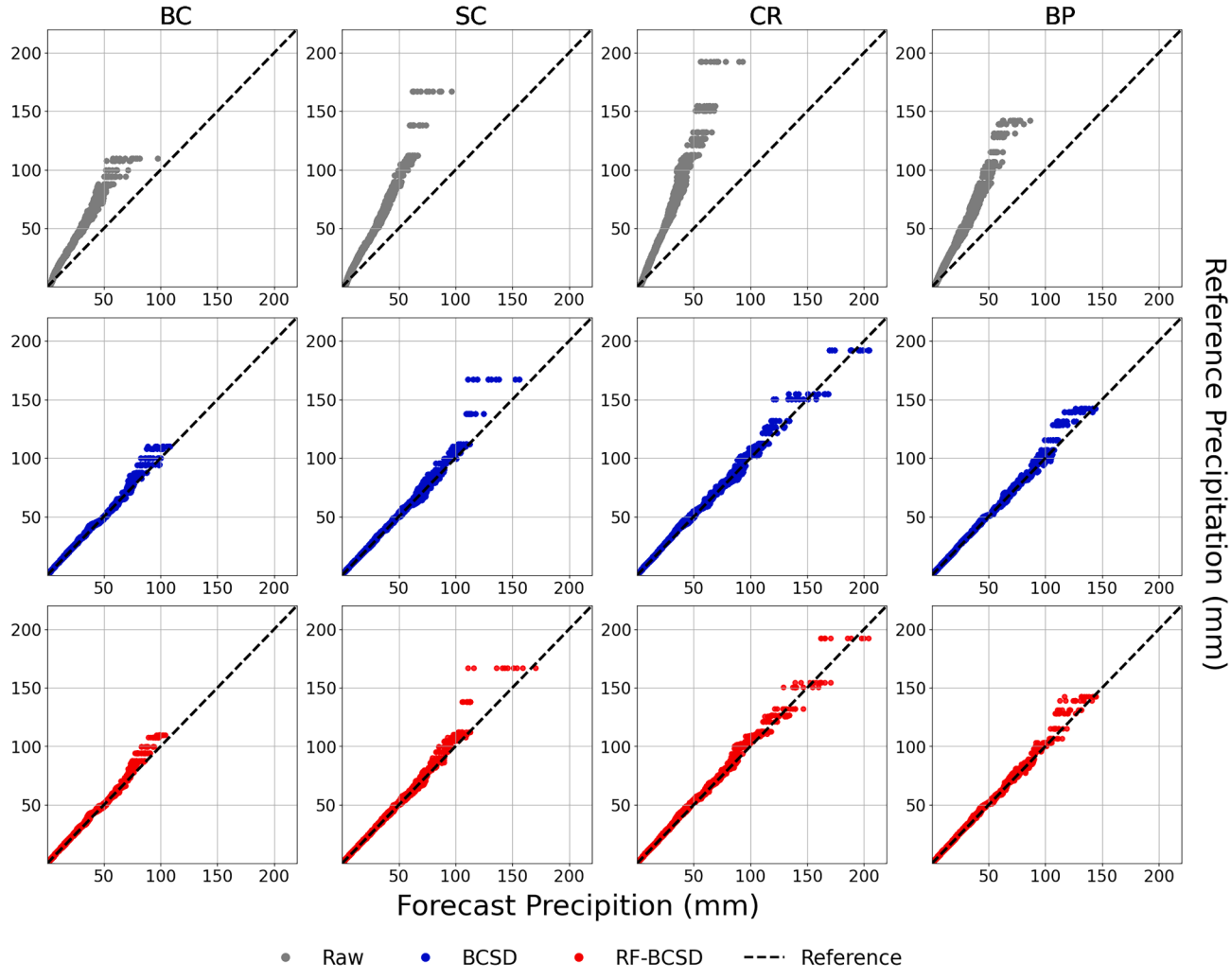


Fig. 4. QQ-plots of the mean areal S2S precipitation in raw condition, after BCSD adaptation, and after RF-BCSD adaptation.

4.2.2. Predictive skill

To further quantify the predictive skill of S2S precipitation resulting from BCSD and RF-BCSD adaptations, two evaluation statistics of the KGE and the CRPSS are computed. Fig. 5 presents the deterministic KGE values of the ensemble means of S2S precipitation at the 4 selected study watersheds over the forecast horizon of 28 days. The higher the KGE values the more skillful the S2S precipitation is. In Fig. 5, the grey-colored, blue-colored, and red-colored lines are the KGE skill of the

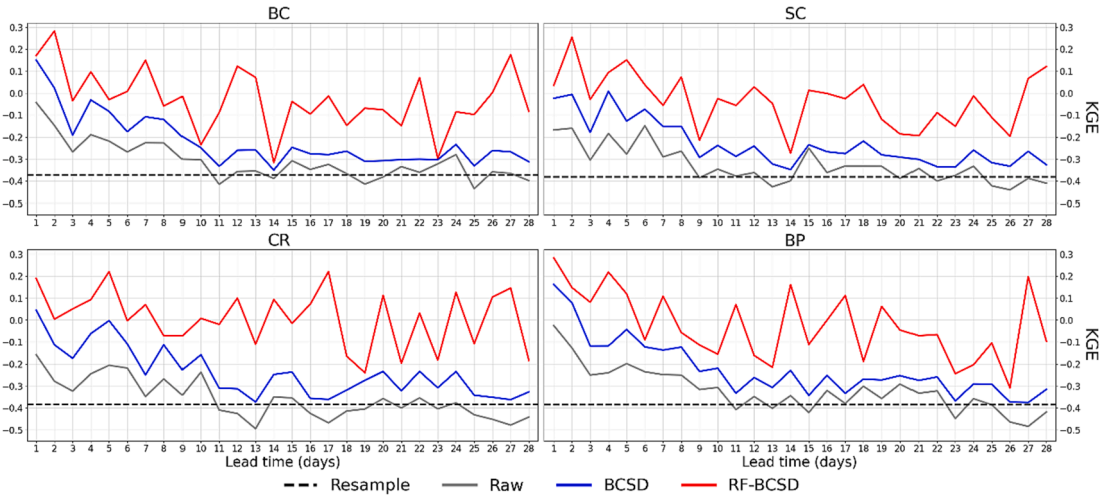


Fig. 5. KGE skill of the mean areal raw S2S precipitation, BCSD-adapted S2S precipitation, RF-BCSD-adapted S2S precipitation, and randomly-resampled precipitation at 4 study watersheds and over the 28-day forecast horizon.

raw S2S precipitation and the S2S precipitation adapted by BCSD and RF-BCSD respectively. The KGE skills of the randomly resampled precipitation are plotted in black dashed lines.

From Fig. 5, it can be observed that all S2S precipitation shows a similar decreasing trend across the forecast horizon and over the 4 study watersheds except for the randomly resampled precipitation as it reflects the constant climatology skills. Comparisons can be made between the S2S precipitation and the resampled precipitation first. For all study watersheds, S2S precipitation shows higher skills than the randomly resampled precipitation when forecast lead times are smaller ~ 10 days. Once the forecast lead exceeds ~ 10 days, the raw S2S precipitation shows an equivalent level of KGE skills to the randomly resampled precipitation whereas the BCSD and RF-BCSD adapted S2S precipitation consistently performs better than the randomly resampled precipitation.

Comparisons can also be made between different S2S precipitation. It can be observed that the S2S precipitation resulting from BCSD and RF-BCSD show consistently higher predictive skill than the raw S2S precipitation at all forecast lead times and across 4 study watersheds. Faburther, the RF-BCSD adaptation leads to the highest predictive skills over the entire forecast horizon and across all study watersheds. Over the entire forecast horizon and across different study watersheds, the RF-BCSD seems bringing more significant improvement in KGE values at longer forecast lead times. This can be told as RF-BCSD associated KGE values are sometimes equivalent to the BCSD associated KGE values at much shorter lead times (e.g., at the BC watershed, RF-BCSD at day 12 and BCSD at day 1; at the SC watershed, RF-BCSD at days 14–18 and BCSD at days 1–2; at the CR watershed, RF-BCSD at days 19–28 and BCSD at days 5–8; at the BP watershed, RF-BCSD at days 23–26 and BCSD at days 7–10).

Fig. 6 presents the probabilistic statistic of the CRPSS of S2S ensemble precipitation as well as the randomly resampled precipitation at the study watersheds over the forecast horizon of 28 days. The larger the CRPSS values the more skillful the precipitation forecast is. Similar to previous Fig. 5, the grey-colored, blue-colored, and red-colored lines are the CRPSS skills of the raw S2S precipitation and the S2S precipitation adapted by BCSD and RF-BCSD respectively. The CPRSS skills of the randomly resampled precipitation are constantly zero and plotted in black dashed lines.

From Fig. 6, it can be observed that both raw and adapted S2S precipitation present a slight decreasing trend of CRPSS. Specifically, S2S

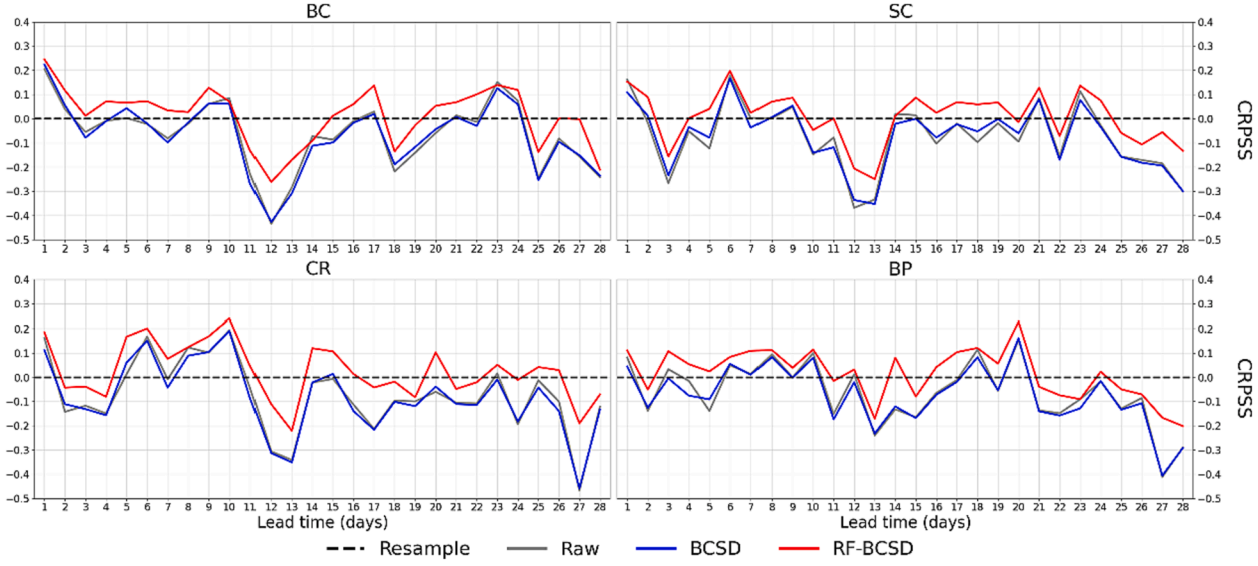


Fig. 6. CRPSS skill of the mean areal raw S2S precipitation, BCSD-adapted S2S precipitation, RF-BCSD-adapted S2S precipitation, and randomly-resampled precipitation at 4 study watersheds and over the 28-day forecast horizon.

precipitation does not present overall superior performance than the randomly resampled precipitation does. S2S precipitation sometimes presents CRPSS values smaller than zero (less skillful than climatology) at forecast lead times within a week (e.g., days 3 and 4 at the BC watershed; day 3 at the SC watershed; days 2, 3, and 4 at the CR watershed; day 2 at the BP watershed). Beyond a week, S2S precipitation also presents less consistent performance where a mixed behavior of both outperforming and underperforming of the climatology CRPSS is observed.

Comparing CRPSS values between different S2S precipitation, it can be observed that RF-BCSD-adapted S2S precipitation outperforms both raw and BCSD-adapted S2S precipitation. It can be observed that RF-BCSD brings consistently higher CRPSS values (more skillful) across different study watersheds at all forecast lead times. On the other hand, the BCSD-adapted S2S precipitation forecast does not show significant improvement over the raw S2S precipitation. The BCSD-adapted S2S precipitation generally presents very close CRPSS values to that from raw S2S precipitation at all forecast lead times and across different study watersheds.

4.2.3. *Probabilistic performance on extreme precipitation events*

The probabilistic evaluation statistics of reliability, resolution, and sharpness were computed to further examine the performance of the adapted S2S precipitation in predicting extreme precipitation events above 97 %. The values of the aforementioned statistics are listed in Table 3.

According to Table 3, the ML-BCSD presents overall less skillful ensemble predictions than BCSD does. The ML-BCSD presents less reliable ensemble forecasts than ML-BCSD does, given overall larger reliability values across the 4 study watersheds. Furthermore, ML-BCSD also presents worse (larger) resolution values than BCSD does, despite previously presented better determinisic KGE and CPRSS skills. Nevertheless, ML-BCSD presents sharper ensemble forecasts in comparison to BCSD as indicated by larger sharpness values. Such larger sharpness values indicate that when predicting extreme events, ML-BCSD tends to produce a narrower ensemble spread in comparison to BCSD.

4.3. *Predictive performance of streamflow*

To quantify how available S2S precipitation affects the overall quality of streamflow predictions while excluding other factors (e.g., the imperfect structure of the hydrologic model, uncertainty arising from parameters of the hydrologic model, etc.), the streamflow predictions resulting from S2S precipitation are evaluated against the proxy of the streamflow measurements (i.e., simulated streamflow from Sac-SMA using the calibrated parameter set and precipitation measurements from PRISM). Similar to the previous evaluation of the S2S precipitation, a deterministic evaluation metric KGE, and a probabilistic metric of CRPS are employed.

Fig. 7 presents the deterministic KGE values of the ensemble means of the streamflow hindcasts resulting from ESP and the adapted S2S precipitation over the forecast horizon of 28 days and across 4 study watersheds. In Fig. 7, the grey-colored lines are the KGE values of the baseline ESP-generated streamflow hindcasts. The blue and red color lines are the KGE values of streamflow hindcasts associated with BCSD and RF-BCSD-adapted S2S precipitation.

Table 3
Reliability, Resolution, and Sharpness of the adapted S2S precipitation forecast for extreme precipitation events above 97 %.

Watersheds	Reliability		Resolution		Sharpness	
	BCSD	ML-BCSD	BCSD	ML-BCSD	BCSD	ML-BCSD
BC	4.62e-5	7.15e-4	8.24e-5	7.44e-4	0.033	0.036
SC	3.77e-5	2.75e-4	7.39e-5	3.33e-4	0.035	0.035
CR	8.39e-5	1.20e-4	1.36e-4	1.20e-3	0.037	0.042
BP	6.13e-5	1.20e-4	1.17e-4	1.21e-3	0.040	0.046

The KGE of streamflow hindcasts presented in Fig. 7 shows a decreasing trend over the forecast horizon at all study watersheds. Comparing KGE at different study watersheds, similar patterns are observed. The KGE values are drastically higher (close to 1) at forecast lead times within 2–3 days than at longer forecast lead times. However, at the BP watershed, such a pattern appears to be less significant compared to that at other watersheds. Comparing the KGE of streamflow associated with ESP, BCSD, and RF-BCSD, ESP presents the lowest KGE values across all study watersheds across the entire forecast horizon, indicating the overall advantage of BCSD and RF-BCSD. Comparing BCSD and RF-BCSD-adapted S2S precipitation, RF-BCSD presents consistently higher KGE values across different study watersheds. However, unlike previous precipitation evaluation results, the differences in KGE values between ESP, BCSD, and RF-BCSDb appear to be marginal at forecast lead times within ~ 3 to ~ 7 days at all watersheds except at BP. At the BP watershed, the advantages of KGE brought by RF-BCSD are rather consistent and show less difference at different forecast lead times.

Fig. 8 presents the probabilistic CRPS values of the entire ensemble of the streamflow hindcasts resulting from ESP, as well as BCSD and RF-BCSD adapted S2S precipitation over the forecast horizon of 28 days and across 4 study watersheds with grey, blue, and red lines.

From Fig. 8, it can be observed that CRPS generally shows an increasing trend over forecast lead time across 4 study watersheds. Such an increase in CRPS indicates a decrease in the predictive skill of streamflow over forecast lead times, which is consistent with the KGE results presented in Fig. 7. One exception would be the BC watershed where such a decreasing trend of the predictive skill seems more chaotic.

Across the entire forecast horizon, no major differences can be observed between CRPS of ESP, BCSD, and RF-BCSD at very short forecast lead times within ~ 3 to ~ 5 days. However, as the forecast lead time increases, such differences become much more apparent and significant. The only exception is the BP watershed where the differences between the CRPS of ESP, BCSD, and RF-BCSD appear to be inconsistent where different CRPS skill curves intersect with each other more often.

Comparing ESP and BCSD, BCSD outperforms or at least provides an equivalent level of CRPS to that of ESP at very short forecast lead times within 7–10 days. However, once exceeding a certain forecast lead time (i.e., at BC watershed after day 9; at SC watershed after day 11; at CR watershed after day 7; at BP watershed after day 7), BCSD underperform ESP and presents higher CRPS values.

In contrast, RF-BCSD presents consistently lower CRPS values (more skillful) than BCSD does at all forecast lead times and across different study watersheds. However, RF-BCSD presents similar characteristics to BCSD when compared to the baseline ESP. That is the advantage of RF-BCSD over ESP becomes less obvious or no longer exists once exceeding a certain forecast lead time (i.e., at BC watershed after day 20; at SC watershed after day 15; at CR watershed after day 14; at BP watershed after day 9).

4.4. *The "elasticities" of the added skill between precipitation and streamflow*

To further quantify the changes in the predictive skill of streamflow hindcasts corresponding to the changes in the predictive skill of precipitation, the "elasticities" of KGE at different forecast lead times are computed and presented in Table 4.

According to the "elasticities" shown in Table 4, the 4 study watersheds can be divided into 2 groups in general. The firs group, being the BC and SC watersheds, show similar behavior, where the "elasticities" are relatively small at very short lead times within 8 days but become drastically larger after ~ 8 days. In the other group, i.e., the CR and BP watersheds, the "elasticities" show different behaviors. At the CR watershed, the elasticities fluctuate across the entire forecast horizon without showing an obvious pattern. The elasticity at the BP watershed

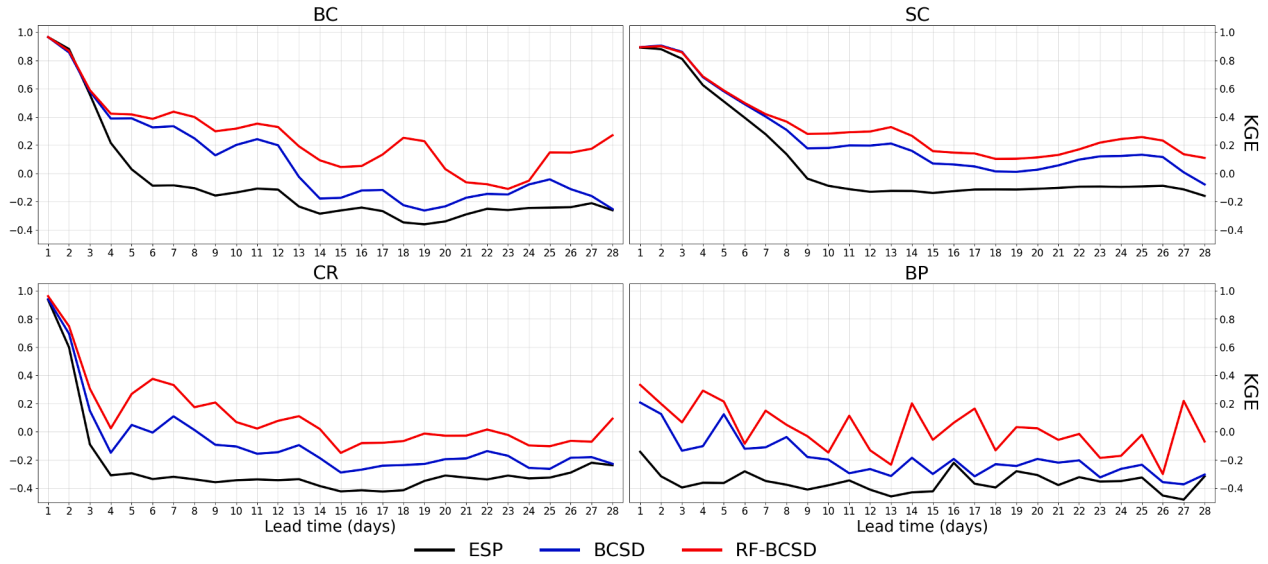


Fig. 7. KGE skill of the streamflow hindcasts resulted from BCSD-adapted S2S precipitation, RF-BCSD-adapted S2S precipitation, and randomly-resampled precipitation (ESP) at 4 study watersheds and over the 28-day forecast horizon.

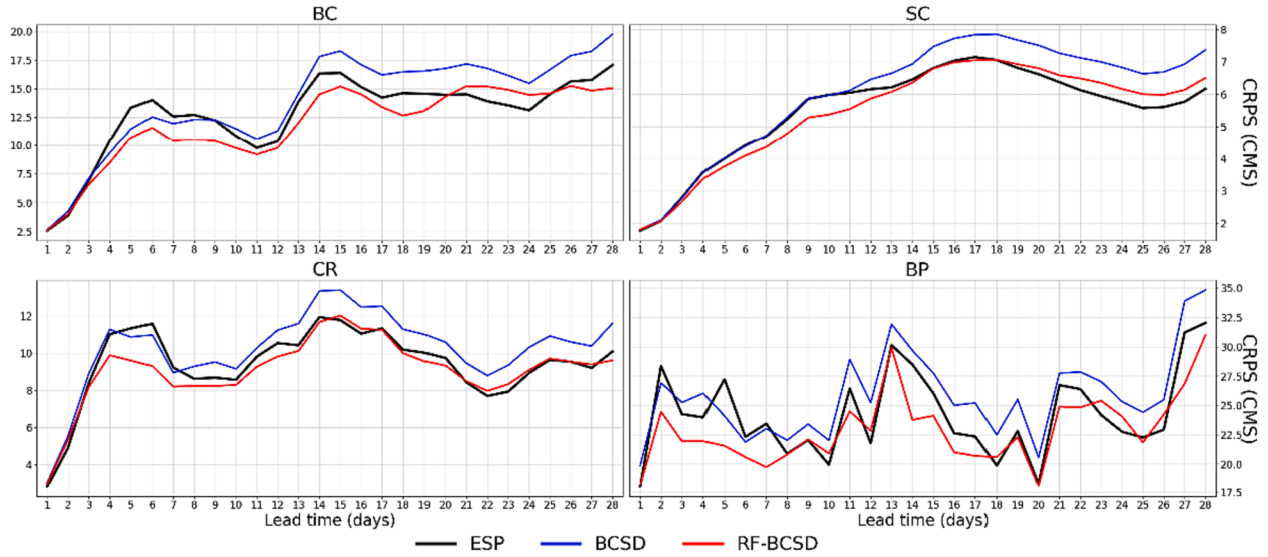


Fig. 8. CRPS skill of the streamflow hindcasts resulted from resulted from BCSD-adapted S2S precipitation, RF-BCSD-adapted S2S precipitation, and randomly-resampled precipitation (ESP) at 4 study watersheds and over the 28-day forecast horizon.

Table 4
KGE "elasticities" at different forecast lead times and different study watersheds.

Study watersheds	Forecast lead times (days)							Mean
	1 to 4	5 to 8	9 to 12	13 to 16	17 to 20	20 to 24	25 to 28	
BC	0.08	0.93	2.50	2.63	2.02	1.74	1.20	1.59
SC	0.00	0.13	0.63	0.62	0.51	0.48	0.56	0.42
CR	0.32	2.24	1.03	0.60	2.01	0.85	0.83	1.13
BP	1.06	0.95	1.02	0.97	1.03	0.97	1.00	1

remains rather stable around the values of 1 across the entire forecast horizon.

5. Discussion

In this study, we have employed two post-processing schemes to adapt the raw S2S precipitation at 4 local watersheds for climate-model-based streamflow predictions. The performance of the adapted S2S precipitation and the corresponding streamflow predictions is benchmarked with the randomly resampled precipitation and the corresponding ESP-generated streamflow predictions. The first post-processing scheme is a popular and standard BCSD while another is a combination of the RF and BCSD.

Our result from section 4.2 shows that both BCSD and RF-BCSD remove the original forecast bias very effectively. Further, the KGE of the ensemble means of randomly resampled precipitation and the adapted S2S precipitation indicate after BCSD and the proposed RF-BCSD forecast adaptation, dynamical S2S precipitation presents consistently higher deterministic skill than the randomly resampled precipitation under the classical ESP framework. However, the probabilistic evaluation metric of CRPSS indicates the dynamical S2S precipitation does not consistently present higher skill compared over climatology (i.e., randomly resampled precipitation).

We reckon that this inconsistency between KGE and CRPSS of the dynamical S2S precipitation could be due to a rather limited size of the ensemble members of the selected dynamical forecasting products. To

be more specific, the selected GEOS5 S2S precipitation forecast only has 10 ensemble members whereas the randomly resampled precipitation consists of 29 ensemble members. Such differences in the ensemble sizes could potentially cause the inconsistency between KGE and CRPSS results, as a small ensemble size could greatly impact the quality of probabilistic forecasts according to Mullen and Buizza (2002).

The proposed RF-BCSD has shown better statistics for most of the evaluation metrics in contrast to BCSD at all forecast lead times and across different study watersheds. In this regard, we reckon the proposed RF-BCSD is effective. However, RF-BCSD did produce inferior statistics in terms of reliability and resolution which are specifically targeted at extreme precipitation events above 97 %. Our explanation is that the proposed RF-BCSD is trained to identify positive extreme events and substitute (statistically) additional high percentile values into the origin ensemble forecasts, which disturbs and worsens reliability and resolution values. Such a post-processing strategy also leads to a narrower ensemble spread which is indicated by larger sharpness values. One potential way to address this issue is to utilize more advanced ML or Deep Learning techniques to jointly consider the entire ensemble forecast for post-processing (Ganaie et al., 2022; Grönquist et al., 2021; Wang et al., 2017a). The Bayesian Joint Probability (BJP) technique could be another alternative in this regard. Previous studies have demonstrated the effectiveness of BJP in post-processing ensemble precipitation forecasts which could reliably quantify forecast uncertainty (Li et al., 2021; Robertson et al., 2013; Yuan and Wood, 2012; Zhao et al., 2016). However, due to the length of this study, it will be a future effort to further explore and compare the effectiveness of other ML and BJP in adapting dynamical ensemble S2S precipitation forecasts.

Nevertheless, the major focus of this study is to introduce a novel and simple forecast adaptation technique in contrast to BCSD for ensemble streamflow predictions. In this regard, our proposed RF-BCSD has shown overall superior performance than the BCSD does. Expectedly, the added skill of the S2S precipitation brought by the application of RF has propagated into the streamflow predictions through the employed hydrologic model. According to the computed KGE and CRPS values in section 4.3, the RF-BCSD adapted S2S precipitation has led to consistently more skillful streamflow predictions over the entire forecast horizon of 28 days and across 4 different study watersheds.

To further quantify the propagation of the added skill of S2S precipitation to streamflow predictions, the "elasticities" of KGE at different forecast lead times are computed and presented in section 4.4. We found that the streamflow forecast seems unaffected by the quality of precipitation at very short forecast lead times at the BC and SC watersheds. However, similar behavior is not observed in the CR and BP watersheds. To explain such differences observed in "elasticities" at the 4 study watersheds, we took a look at the calibrated parameter sets of the Sac-SMA models at the 4 study watersheds. For BC and SC, the upper zone free water capacities (UZFWC) are 46 mm and 52 mm (3 folds averaged parameter values, below the same), which are much larger than upper zone tension water capacities (UZTWC) of 8.33 mm and 12.31 mm. Further, the upper zone water depletion coefficient (UZK, which describes the linear relationship between the stored water in UZFWC and its contributing runoff) at BC and SC are relatively small with values of 0.21 and 0.17. In contrast, the UZFWC at CR and BP (82 mm and 37 mm) are much smaller than the UZTWC (140 mm and 120 mm). In addition, the UZK at CR and BP (0.58 and 0.68) are much larger than that at BC and SC.

Given these parameters, while considering the computation logic of the Sac-SMA, we believe the observed difference in "elasticities" could at least be partially attributed to the differences in "hydrologic memories" at the 4 study watersheds. In Sac-SMA, the amount of input precipitation tries to fill UZTWC first before entering UZFWC. Therefore, a larger UZTWC would result in less water in UZFWC. However, only the water stored in UZFWC contributes to streamflow. Further, smaller UZK values could make such lagged effects last longer due to a slower draining speed. Therefore, at BC and SC, relatively more water in the UZFWC and relatively slower draining speed lead to relatively longer "hydrologic memories", and vice versa at the CR and BP watersheds.

The different drainage areas of the study watersheds could also contribute to the observed difference among "elasticities". That is, for watersheds with a larger drainage area, precipitation falls on the upper end of the watershed would naturally take longer time to manifest as streamflow compared to that at watersheds with smaller drainage areas. This characteristic of larger watersheds could potentially lead to a longer "hydrologic memory" in our previous analysis. However, our previous analysis of the "hydrologic memory" indicates that the largest study watershed BP actually presents a relatively shorter "hydrologic memory". Therefore, we reckon our analysis is not compromised by the difference in drainage areas between study watersheds in general. Nevertheless, to better quantify the impact of the "hydrologic memory" on streamflow prediction at different watersheds, more advanced distributed hydrologic models need to be employed for more detailed analysis for future studies.

Our interpretation of the different behaviors of "elasticity" values across study watersheds suggests that further advancements of streamflow predictions at the S2S timescale perhaps require efforts in multi-aspects. analysis of "elasticities" at different forecast lead times indicates that at shorter forecast lead times, the quality of streamflow predictions could be significantly affected by the uncertainty arising from the estimation of IHCs. Therefore, more advanced data assimilation techniques should be developed to consider the land-surface inertia of watersheds more accurately and comprehensively.

Previous studies suggest that accurate, reliable, and seamless streamflow predictions at the S2S timescale could greatly benefit many human activities including public health, disaster preparedness, hydropower generation scheduling, and irrigation planning (Graham et al., 2022; White et al., 2017; Yang et al., 2017; Yang et al., 2020). Our analysis indicates that the quality of precipitation forecasts at the S2S timescale is still the limiting factor for superior streamflow predictions. According to our result, at longer forecast lead times exceeding ~ 10 days, the predictive skill of streamflow is more likely to be dominated by the quality of precipitation forecasts. Efforts have been made to advance weather/climate predictions at the S2S timescale by identifying additional predictability sources, better assimilating measured atmospheric variables, and improving modeling tools (Domeisen et al., 2022; Mayer and Barnes, 2022; White et al., 2022; Yang et al., 2016). However, due to uncertainty arising from various sources, it is still extremely challenging to provide accurate and reliable weather/climate forecasts at the S2S timescale to this date (AghaKouchak et al., 2022; Krishnamurthy, 2019).

It is therefore important for hydrologists, or other forecast end-users, to make the best of available S2S precipitation forecasts through various post-processing techniques. In this study, we have demonstrated the effectiveness of a rather simple and popular ML-based approach (i.e., RF) at 4 study watersheds by combining it with a more conventional distribution-based forecast adaption technique (i.e., BCSD) for streamflow forecasting at the S2S timescale. Due to a rather limited training sample size subject to the availability of S2S precipitation hindcasts, we did not train RF for direct volumetric information. Instead, in this study, the RF is trained to give categorical predictions. The corresponding volumetrical information is further corrected by the BCSD. The effectiveness of such a framework suggests that more advanced data-driven techniques could also be effective for similar tasks, as long as enough training samples are provided. Most recently, Pan et al. (2021) have applied a deep learning technique, termed generative adversarial network (GAN), in correcting the bias of GCM-simulated precipitation over the entire CONUS. The successful application of GAN by Pan et al. (2021) shed some light on potential future works at a watershed scale, as GAN is known for its prone to over-fitting with limited training data (Creswell et al., 2018; Wang et al., 2017b). Therefore, we encourage future studies to further explore the effectiveness of more advanced data-driven approaches in adapting S2S precipitation at different watersheds for more accurate and reliable S2S streamflow forecasting.

In summary, the authors believe that to advance streamflow forecast at the S2S timescale, future studies could (1) develop novel or more advanced data assimilation techniques to better consider the "hydrologic memory" of watersheds; (2) improve weather/climate predictions from its core for more accurate and reliable S2S hydrometeorological forecasts; (3) apply other data-driven approaches to harness S2S forecast products in hydrology through various emerging deep learning techniques.

6. Conclusion

In this study, we investigated the effectiveness of an ML technique, termed RF, in adapting the raw S2S precipitation forecasts from NASA GEOS5 to 4 local watersheds in the NCEI South climate region. To address the commonly presented forecast bias as well as to improve the predictive skill of raw S2S forecast during the forecast adaptation, the RF is jointly employed with BCSD (i.e., RF-BCSD) in contrast to another scenario where only BCSD is employed. The adapted S2S precipitation under both schemes is further applied to force the lumped Sac-SMA model for streamflow hindcast experiments. The randomly resampled precipitation and the corresponding streamflow predictions generated with the classical ESP are also incorporated in this study to serve as the benchmark against the two sets of adapted dynamical S2S precipitation and their corresponding streamflow predictions.

According to our result, the adapted S2S precipitation presents higher deterministic skill (higher KGE) compared to the randomly resampled precipitation. However, the adapted S2S precipitation does not consistently present higher probabilistic skill compared to the randomly resampled precipitation as indicated by CRPSS. We reckon this inconsistency between KGE and CRPSS of the adapted S2S precipitation could be due to its limited ensemble size, compared to the randomly resampled precipitation. Comparing the proposed RF-BCSD and BCSD, RF-BCSD outperforms BCSD with not only higher KGE but also better CRPSS. However, one drawback of the proposed RF-BCSD is the worsened reliability and resolution of the resulting ensemble forecast.

The resulting streamflow predictions generally present consistent performance with precipitation forecasts. Specifically, the adapted S2S precipitation (RF-BCSD and BCSD) leads to overall higher deterministic skills (KGE) in contrast to ESP. However, the streamflow prediction associated with adapted S2S precipitation (RF-BCSD and BCSD) does not outperform ESP in terms of probabilistic skills (CRPS) when forecast lead times exceed 10–21 days at different watersheds. We reckon this issue may be mitigated by including more dynamical precipitation forecast products to form a larger ensemble size. Nevertheless, the proposed RF-BCSD outperforms BCSD in streamflow forecasting. In addition, the propagation of the added skill from precipitation to streamflow brought by RF-BCSD is quantified with our pre-defined "elasticity". Our results highlight future applications of other data-driven ML or Deep Learning techniques at watersheds to get the best use of available S2S forecast products for more accurate streamflow forecasts. Our major conclusions are summarized as follows:

1. With proper forecast adaptation, dynamical S2S precipitation from GEOS leads to consistently higher deterministic predictive skills in contrast to the randomly resampled precipitation. However, the adapted S2S precipitation does not present superior probabilistic skill metrics over ESP.
2. The resulting streamflow prediction generally presents consistent performance with the precipitation where adapted S2S precipitation leads to deterministically more skillful streamflow forecasting at all forecast lead times and across all study watersheds. However, once forecast lead time exceeds a certain forecast lead time, adapted S2S precipitation presents inferior probabilistic skill in contrast to ESP.
3. Comparing the proposed RF-BCSD and BCSD, RF-BCSD leads to overall more skillful precipitation as well as streamflow predictions as suggested by different evaluation statistics. However, under the current design of RF-BCSD, the reliability and resolution of the adapted S2S ensemble recitation are worsened compared to BCSD.
4. Our analysis of the "elasticity" of the added skill between precipitation and streamflow suggests that the "memory" of the hydrologic system could play an important role in terms of the accuracy of streamflow forecasts, especially at relatively shorter forecast lead times.

CRediT authorship contribution statement

Lujun Zhang: Conceptualization, Formal analysis, Methodology, Software, Writing – original draft. **Shang Gao:** Formal analysis, Writing – review & editing, Methodology. **Tiantian Yang:** Funding acquisition, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All the datasets used in this study are available in https://prism.oregonstate.edu/, https://www.emc.ncep.noaa.gov/mmb/rreanl/, https://waterdata.usgs.gov/nwis, https://www.cpc.ncep.noaa.gov/products/NMME/data.html.

Acknowledgments

The financial support of this work is from the National Science Foundation (NSF) CAREER Award (No. 2236926) and the EPSCoR Track-1 Project under Grant No. OIA-1946093 and its subaward No. EPSCoR-2020-3. This work is also partially supported by the U.S. Department of Defense, Army Corps of Engineers (DOD-COR) Engineering With Nature (EWN) Program (Award No. W912HZ-21-2-0038), and U.S. Dept. of Commerce, National Oceanic and Atmospheric Administration (NOAA) agreement No. NA23OAR4310459.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2024.130643.

References

AghaKouchak, A., Pan, B., Mazdiyasni, O., Sadegh, M., Jiwa, S., Zhang, W., Love, C., Madadgar, S., Papalexiou, S., Davis, S., 2022. Status and prospects for drought forecasting: opportunities in artificial intelligence and hybrid physical–statistical forecasting. Phil. Trans. R. Soc. A 380 (2238), 20210288.

Akbari Asanjan, A., Yang, T., Hsu, K., Sorooshian, S., Lin, J., Peng, Q., 2018. Short-term precipitation forecast based on the PERSIANN system and LSTM recurrent neural networks. J. Geophys. Res. Atmos. 123 (22).

Borovikov, A., Cullather, R., Kovach, R., Marshak, J., Vernieres, G., Vikhliaev, Y., Zhao, B., Li, Z., 2019. GEOS-5 seasonal forecast system. Clim. Dyn. 53, 7335–7361.

Cao, Q., Shukla, S., DeFlorio, M.J., Ralph, F.M., Lettenmaier, D.P., 2021. Evaluation of the subseasonal forecast skill of floods associated with atmospheric rivers in coastal Western US watersheds. J. Hydrometeorol. 22 (6), 1535–1552.

Caruana, R., Lawrence, S. and Giles, C. 2000. Overfitting in neural nets: Backpropagation, conjugate gradient, and early stopping. Advances in neural information processing systems 13.

Chiew, F., Zhou, S., McMahon, T., 2003. Use of seasonal streamflow forecasts in water resources management. J. Hydrol. 270 (1–2), 135–144.

Creswell, A., White, T., Dumoulin, V., Arulkumaran, K., Sengupta, B., Bharath, A.A., 2018. Generative adversarial networks: an overview. IEEE Signal Process Mag. 35 (1), 53–65.

Day, G.N., 1985. Extended streamflow forecasting using NWSRFS. J. Water Resour. Plan. Manag. 111 (2), 157–170.

de Andrade, F.M., Young, M.P., MacLeod, D., Hirons, L.C., Woolnough, S.J., Black, E., 2021. Subseasonal precipitation prediction for Africa: Forecast evaluation and sources of predictability. Weather Forecast. 36 (1), 265–284.

Delaney, C.J., Hartman, R.K., Mendoza, J., Dettinger, M., Delle Monache, L., Jasperse, J., Ralph, F.M., Talbot, C., Brown, J., Reynolds, D., 2020. Forecast informed reservoir operations using ensemble streamflow predictions for a multipurpose reservoir in Northern California. Water Resour. Res. 56 (9).
Domeisen, D.I., White, C.J., Afargan-Gerstman, H., Muñoz, Á.G., Janiga, M.A., Vitart, F., Wulff, C.O., Antoine, S., Ardilouze, C., Batté, L., 2022. Advances in the subseasonal prediction of extreme events: Relevant case studies across the globe. Bull. Am. Meteorol. Soc. 103 (6), E1473–E1501.
Ganaie, M.A., Hu, M., Malik, A., Tanveer, M., Suganthan, P., 2022. Ensemble deep learning: A review. Eng. Appl. Artif. Intel. 115, 105151.
Graham, R.M., Browell, J., Bertram, D., White, C.J., 2022. The application of sub-seasonal to seasonal (S2S) predictions for hydropower forecasting. Meteorol. Appl. 29 (1), e2047.
Grillakis, M.G., Koutroulis, A.G., Tsanis, I.K., 2013. Multisegment statistical bias correction of daily GCM precipitation output. J. Geophys. Res. Atmos. 118 (8), 3150–3162.
Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., Hoefler, T., 2021. Deep learning for post-processing ensemble weather forecasts. Phil. Trans. R. Soc. A 379 (2194), 20200092.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377 (1–2), 80–91.
Harrigan, S., Prudhomme, C., Parry, S., Smith, K., Tanguy, M., 2018. Benchmarking ensemble streamflow prediction skill in the UK. Hydrol. Earth Syst. Sci. 22 (3), 2023–2039.
Jones, K.A., Niknami, L.S., Buto, S.G., Decker, D., 2022. Federal Standards and Procedures for the National Watershed Boundary Dataset (WBD): Chapter 3 of Section A, Federal Standards, Book 11. Collection and Delineation of Spatial Data, US Geological Survey.
Karras, T., Aittala, M., Hellsten, J., Laine, S., Lehtinen, J., Aila, T., 2020. Training generative adversarial networks with limited data. Adv. Neural Inf. Proces. Syst. 33, 12104–12114.
Kim, T., Yang, T., Gao, S., Zhang, L., Ding, Z., Wen, X., Gourley, J.J., Hong, Y., 2021. Can artificial intelligence and data-driven machine learning models match or even replace process-driven hydrologic models for streamflow simulation?: A case study of four watersheds with different hydro-climatic regions across the CONUS. J. Hydrol. 598, 126423.
Kim, T., Yang, T., Zhang, L., Hong, Y., 2022. Near real-time hurricane rainfall forecasting using convolutional neural network models with Integrated Multi-satellitE Retrievals for GPM (IMERG) product. Atmos. Res. 270, 106037.
Kirtman, B.P., Min, D., Infanti, J.M., Kinter, J.L., Paolino, D.A., Zhang, Q., Van Den Dool, H., Saha, S., Mendez, M.P., Becker, E., 2014. The North American multimodel ensemble: phase-1 seasonal-to-interannual prediction; phase-2 toward developing intraseasonal prediction. Bull. Am. Meteorol. Soc. 95 (4), 585–601.
Knoben, W.J., Freer, J.E., Woods, R.A., 2019. Inherent benchmark or not? Comparing Nash-Sutcliffe and Kling-Gupta efficiency scores. Hydrol. Earth Syst. Sci. 23 (10), 4323–4331.
Krishnamurthy, V., 2019. Predictability of weather and climate. Earth Space Sci. 6 (7), 1043–1056.
Li, Y., Wu, Z., He, H., Wang, Q.J., Xu, H., Lu, G., 2021. Post-processing sub-seasonal precipitation forecasts at various spatiotemporal scales across China during boreal summer monsoon. J. Hydrol. 598, 125742.
Li, H., Wu, Z., Yuan, X., Yang, Y., He, X., Duan, H., 2022. The research on modeling and application of dynamic grey forecasting model based on energy price-energy consumption-economic growth. Energy 257, 124801.
Liu, J., Yuan, X., Zeng, J., Jiao, Y., Li, Y., Zhong, L., Yao, L., 2022. Ensemble streamflow forecasting over a cascade reservoir catchment with integrated hydrometeorological modeling and machine learning. Hydrol. Earth Syst. Sci. 26 (2), 265–278.
Ma, R., Yuan, X., 2023. Subseasonal Ensemble Prediction of Flash Droughts over China. J. Hydrometeorol. 24 (5), 897–910.
Mayer, K.J., Barnes, E.A., 2022. Quantifying the effect of climate change on midlatitude subseasonal prediction skill provided by the tropics. Geophys. Res. Lett. 49 (14).
Mullen, S.L., Buizza, R., 2002. The impact of horizontal resolution and ensemble size on probabilistic forecasts of precipitation by the ECMWF ensemble prediction system. Weather Forecast. 17 (2), 173–191.
Naeini, M.R., Yang, T., Sadegh, M., AghaKouchak, A., Hsu, K.-L., Sorooshian, S., Duan, Q., Lei, X., 2018. Shuffled complex-self adaptive hybrid evolution (SC-SAHEL) optimization framework. Environ. Model. Softw. 104, 215–235.
Newman, A., Clark, M., Sampson, K., Wood, A., Hay, L., Bock, A., Viger, R., Blodgett, D., Brekke, L., Arnold, J., 2015. Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: data set characteristics and assessment of regional variability in hydrologic model performance. Hydrol. Earth Syst. Sci. 19 (1), 209–223.
Pan, B., Anderson, G.J., Goncalves, A., Lucas, D.D., Bonfils, C.J., Lee, J., Tian, Y., Ma, H. Y., 2021. Learning to correct climate projection biases. J. Adv. Model. Earth Syst. 13 (10).
Pegion, K., Kirtman, B.P., Becker, E., Collins, D.C., LaJoie, E., Burgman, R., Bell, R., DelSole, T., Min, D., Zhu, Y., 2019. The Subseasonal Experiment (SubX): A multimodel subseasonal prediction experiment. Bull. Am. Meteorol. Soc. 100 (10), 2043–2060.
Richter, J.H., Glanville, A.A., Edwards, J., Kauffman, B., Davis, N.A., Jaye, A., Kim, H., Pedatella, N.M., Sun, L., Berner, J., 2022. Subseasonal Earth system prediction with CESM2. Weather Forecast. 37 (6), 797–815.
Robertson, D., Shrestha, D., Wang, Q., 2013. Post-processing rainfall forecasts from numerical weather prediction models for short-term streamflow forecasting. Hydrol. Earth Syst. Sci. 17 (9), 3587–3603.
Schaake, J. and Larson, L. 1998 Ensemble streamflow prediction (ESP): Progress and research needs, pp. J19-J24.
Shukla, S., Sheffield, J., Wood, E.F., Lettenmaier, D.P., 2013. On the sources of global land surface hydrologic predictability. Hydrol. Earth Syst. Sci. 17 (7), 2781–2796.
Su, L., Cao, Q., Shukla, S., Pan, M., Lettenmaier, D.P., 2023. Evaluation of Subseasonal Drought Forecast Skill over the Coastal Western United States. J. Hydrometeorol. 24 (4), 709–726.
Tian, D., Wood, E.F., Yuan, X., 2017. CFSv2-based sub-seasonal precipitation and temperature forecast skill over the contiguous United States. Hydrol. Earth Syst. Sci. 21 (3), 1477–1490.
Troin, M., Arsenault, R., Wood, A.W., Brissette, F., Martel, J.-L., 2021. Generating ensemble streamflow forecasts: A review of methods and approaches over the past 40 years. Water Resour. Res. 57 (7).
Vitart, F., Ardilouze, C., Bonet, A., Brookshaw, A., Chen, M., Codorean, C., Déqué, M., Ferranti, L., Fucile, E., Fuentes, M., 2017. The subseasonal to seasonal (S2S) prediction project database. Bull. Am. Meteorol. Soc. 98 (1), 163–173.
Wang, K., Gou, C., Duan, Y., Lin, Y., Zheng, X., Wang, F.-Y., 2017b. Generative adversarial networks: introduction and outlook. IEEE/CAA J. Autom. Sin. 4 (4), 588–598.
Wang, H.-Z., Li, G.-Q., Wang, G.-B., Peng, J.-C., Jiang, H., Liu, Y.-T., 2017a. Deep learning based ensemble approach for probabilistic wind power forecasting. Appl. Energy 188, 56–70.
White, C.J., Carlsen, H., Robertson, A.W., Klein, R.J., Lazo, J.K., Kumar, A., Vitart, F., Coughlan de Perez, E., Ray, A.J., Murray, V., 2017. Potential applications of subseasonal-to-seasonal (S2S) predictions. Meteorol. Appl. 24 (3), 315–325.
White, C.J., Domeisen, D.I., Acharya, N., Adefisan, E.A., Anderson, M.L., Aura, S., Balogun, A.A., Bertram, D., Bluhm, S., Brayshaw, D.J., 2022. Advances in the application and utility of subseasonal-to-seasonal predictions. Bull. Am. Meteorol. Soc. 103 (6), E1448–E1472.
Wilks, D.S., 2011. Statistical methods in the atmospheric sciences. Academic press.
Wood, A.W., Lettenmaier, D.P., 2008. An ensemble approach for attribution of hydrologic prediction uncertainty. Geophys. Res. Lett. 35 (14).
Wood, A.W., Leung, L.R., Sridhar, V., Lettenmaier, D., 2004. Hydrologic implications of dynamical and statistical approaches to downscaling climate model outputs. Clim. Change 62 (1–3), 189–216.
Yang, T., Gao, X., Sorooshian, S., Li, X., 2016. Simulating California reservoir operation using the classification and regression-tree algorithm combined with a shuffled cross-validation scheme. Water Resour. Res. 52 (3), 1626–1651.
Yang, T., Asanjan, A.A., Welles, E., Gao, X., Sorooshian, S., Liu, X., 2017. Developing reservoir monthly inflow forecasts using artificial intelligence and climate phenomenon information. Water Resour. Res. 53 (4), 2786–2812.
Yang, T., Liu, X., Wang, L., Bai, P., Li, J., 2020. Simulating hydropower discharge using multiple decision tree methods and a dynamical model merging technique. J. Water Resour. Plan. Manag. 146 (2), 04019072.
Yang, T., Zhang, L., Kim, T., Hong, Y., Zhang, D., Peng, Q., 2021. A large-scale comparison of Artificial Intelligence and Data Mining (AI&DM) techniques in simulating reservoir releases over the Upper Colorado Region. J. Hydrol. 602, 126723.
Yuan, X., 2016. An experimental seasonal hydrological forecasting system over the Yellow River basin–Part 2: The added value from climate forecast models. Hydrol. Earth Syst. Sci. 20 (6), 2453–2466.
Yuan, X., Wood, E.F., 2012. Downscaling precipitation or bias-correcting streamflow? Some implications for coupled general circulation model (CGCM)-based ensemble seasonal hydrologic forecast. Water Resour. Res. 48 (12).
Yuan, X., Wood, E.F., 2013. Multimodel seasonal forecasting of global drought onset. Geophys. Res. Lett. 40 (18), 4900–4905.
Yuan, X., Wood, E.F., Roundy, J.K., Pan, M., 2013. CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. J. Clim. 26 (13), 4828–4847.
Zhang, L., Yang, T., Gao, S., Hong, Y., Fan, M., Lu, D., Xu, H. and Xiao, C. 2023a. An alternative Ensemble Streamflow Prediction (ESP) approach using improved Subseasonal-to-Seasonal (S2S) precipitation forecasts from the North America Multi-Model Ensemble Phase II (NMME-2) dataset. Journal of Hydrology (under review).
Zhang, L., Kim, T., Yang, T., Hong, Y., Zhu, Q., 2021. Evaluation of Subseasonal-to-seasonal (S2S) precipitation forecast from the North American Multi-Model ensemble phase II (NMME-2) over the contiguous US. J. Hydrol. 603, 127058.
Zhang, L., Yang, T., Gao, S., Hong, Y., Zhang, Q., Wen, X., Cheng, C., 2023b. Improving Subseasonal-to-Seasonal forecasts in predicting the occurrence of extreme precipitation events over the contiguous U.S. using machine learning models. Atmos. Res. 281, 106502.
Zhao, T., Schepen, A., Wang, Q., 2016. Ensemble forecasting of sub-seasonal to seasonal streamflow by a Bayesian joint probability modelling approach. J. Hydrol. 541, 839–849.
Zhu, E., Wang, Y., Yuan, X., 2023. Changes of terrestrial water storage during 1981–2020 over China based on dynamic-machine learning model. J. Hydrol. 621, 129576.