Investigating the Role of Snow Water Equivalent on Streamflow Predictability during Drought

PARTHKUMAR A. MODI,[a] ERIC E. SMALL,[b] JOSEPH KASPRZYK,[a] AND BEN LIVNEH[a,c]

[a] *Department of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, Colorado*
[b] *Department of Geological Sciences, University of Colorado Boulder, Boulder, Colorado*
[c] *Cooperative Institute for Research in Environmental Science, University of Colorado Boulder, Boulder, Colorado*

(Manuscript received 23 November 2021, in final form 27 April 2022)

ABSTRACT: Snowpack provides the majority of predictive information for water supply forecasts (WSFs) in snow-dominated basins across the western United States. Drought conditions typically accompany decreased snowpack and lowered runoff efficiency, negatively impacting WSFs. Here, we investigate the relationship between snow water equivalent (SWE) and April–July streamflow volume (AMJJ-V) during drought in small headwater catchments, using observations from 31 USGS streamflow gauges and 54 SNOTEL stations. A linear regression approach is used to evaluate forecast skill under different historical climatologies used for model fitting, as well as with different forecast dates. Experiments are constructed in which extreme hydrological drought years are withheld from model training, that is, years with AMJJ-V below the 15th percentile. Subsets of the remaining years are used for model fitting to understand how the climatology of different training subsets impacts forecasts of extreme drought years. We generally report overprediction in drought years. However, training the forecast model on drier years, that is, below-median years (P_{15}, $P_{57.5}$], minimizes residuals by an average of 10% in drought year forecasts, relative to a baseline case, with the highest median skill obtained in mid- to late April for colder regions. We report similar findings using a modified National Resources Conservation Service (NRCS) procedure in nine large Upper Colorado River basin (UCRB) basins, highlighting the importance of the snowpack–streamflow relationship in streamflow predictability. We propose an "adaptive sampling" approach of dynamically selecting training years based on antecedent SWE conditions, showing error reductions of up to 20% in historical drought years relative to the period of record. These alternate training protocols provide opportunities for addressing the challenges of future drought risk to water supply planning.

SIGNIFICANCE STATEMENT: Seasonal water supply forecasts based on the relationship between peak snowpack and water supply exhibit unique errors in drought years due to low snow and streamflow variability, presenting a major challenge for water supply prediction. Here, we assess the reliability of snow-based streamflow predictability in drought years using a fixed forecast date or fixed model training period. We critically evaluate different training protocols that evaluate predictive performance and identify sources of error during historical drought years. We also propose and test an "adaptive sampling" application that dynamically selects training years based on antecedent SWE conditions providing to overcome persistent errors and provide new insights and strategies for snow-guided forecasts.

KEYWORDS: Streamflow; Drought; Snowpack; Seasonal forecasting; Statistical forecasting

1. Introduction

In mountainous regions of the western United States, the majority of annual runoff originates as snowmelt, despite only an estimated 37% of precipitation falling as snow (Palmer 1988; Doesken and Judson 1996; Daly et al. 2000; Li et al. 2017). Water supply forecasts (WSFs; Garen 1992) predict seasonal streamflow volume to support a broad array of natural resource decisions (Pagano et al. 2004). The recurring cycle of snowpack accumulating in colder months and subsequent snowmelt producing streamflow has been one of the fundamental relationships facilitating WSFs. However, in recent decades, warmer climate across the western United States has been accompanied by declines in mountain snowpack (Barnett et al. 2005; Mote et al. 2018) and increased interannual streamflow variability (Pagano and Garen 2005; Abatzoglou et al. 2014). These changes have exacerbated forecast errors and have challenged assumptions of stationarity that underpin contemporary operational WSFs (Sturtevant and Harpold 2019). While it has been established that climate warming will impact WSFs in general (He et al. 2016) and categorical drought prediction in particular (Livneh and Badger 2020), quantifying the sensitivity of historic forecast skill at different forecast dates is arguably most valuable for water management during drought years when allocation shortfalls may occur. This assessment is crucial given the elevated need for reliable water supply information during drought to

Denotes content that is immediately available upon publication as open access.

Supplemental information related to this paper is available at the Journals Online website: https://doi.org/10.1175/JHM-D-21-0229.s1.

Corresponding author: Parthkumar Modi, parthkumar.modi@colorado.edu

DOI: 10.1175/JHM-D-21-0229.1

support municipal, agricultural, industrial water supply planning, trade, and power generation (Suhr Pierce et al. 2010). The goal of this paper is to critically evaluate snow-based seasonal water supply prediction during drought, to identify persistent sources of errors and opportunities to improve predictions using alternative training protocols during the forecast season.

Increased interannual variability in the classic snowpack–-streamflow relationship is expected to continue during current and future drought years due to recently documented changes in the underlying physical mechanisms. Declines in the mountain snowpack (Barnett et al. 2005; Mote et al. 2005, 2018), resulting from increasing snow-to-rain transitions (Lute et al. 2015) and shifts in the timing of snow ablation (Kapnick and Hall 2012), have caused slower snowmelt rates (Musselman et al. 2017 2021) and earlier snowmelt (Dettinger and Cayan 1995; Stewart et al. 2004) for at least the past five decades. These changes, attributable to widespread changes in temperature and precipitation (Cubasch et al. 2001; Hamlet et al. 2005; Serreze et al. 1999), are expected to continue impacting water supplies across the western United States. Further, persistent dry states partially attributable to climate warming have already manifested during the early years of the twenty-first century (MacDonald et al. 2008; Williams et al. 2020). Overall declines in seasonal streamflow volume have been accompanied by lowered runoff efficiency (Nowak et al. 2012; Woodhouse et al. 2016) and increased winter snowmelt (Pagano et al. 2004). All these factors combined present a major challenge ahead for the WSF forecast skill for current and future drought prediction (He et al. 2016; Livneh and Badger 2020).

WSFs can be broadly classified into three categories: statistical, dynamical, and hybrid. Statistical WSFs include regression-based and data-driven models that rely on empirical relationships. Dynamical WSFs encompass process-based models that represent the underlying physics. Hybrid WSFs consist of multimodel combinations such as coupling of statistical and dynamical techniques. All WSFs ultimately rely on two sources of predictability: initial hydrologic conditions (IHCs) obtained from a range of in situ observations or remote sensing data products like that of snow, meteorological conditions; and gauged streamflow, and seasonal climate forecasts that provide the estimates of seasonal conditions ahead of time. In regions across the West, most predictive information is still derived from knowledge of snowpack conditions (Fleming and Goodbody 2019; Koster et al. 2010; Pagano 2010; Wood et al. 2016) and hence snow water equivalent (SWE), around the date of peak SWE, is considered to be a skillful predictor for WSFs (Pagano et al. 2004). Statistical WSFs have conventionally relied on IHCs that include SWE and accumulated precipitation as well as the occasional use of additional predictors like antecedent streamflow and soil moisture. However, recent use of climate indices (Robertson and Wang 2012) and seasonal climate forecast information (Lehner et al. 2017; Slater and Villarini 2018) have helped to mitigate the impacts of climate nonstationarity on streamflow predictability by accounting for ongoing influences of ocean–atmosphere oscillations. They are typically issued by the National Resources Conservation Service (NRCS) and are well established using linear (Garen 1992) and multivariate regression approaches (Koster et al. 2010; Lehner et al. 2017). Commonly used advanced statistical (or machine learning) WSFs like artificial neural networks (Kişi 2007) or support vector machines (Asefa et al. 2006; Guo et al. 2011) have thus far seen application primarily within research-based contexts (Fleming and Goodbody 2019). Nevertheless, recent demonstrations of improved physical interpretability (Fleming et al. 2021b; McGovern et al. 2019; Reichstein et al. 2019), increasingly better performance (Kratzert et al. 2019; Nearing et al. 2021), and the development of the NRCS next-generation WSF system [multimodel machine-learning metasystem (M4); Fleming and Goodbody 2019; Fleming et al. 2021a], make advanced statistical frameworks a viable contender to contemporary WSFs within the near future. Major strengths of statistical WSFs are data-driven modeling, straightforward interpretability, and low computational requirements (Pagano et al. 2009). However, they pose drawbacks including limitations in observational data availability for certain regions and time periods, lack of explicit physical consideration, and an inability to account for water inputs after the forecast date.

Dynamical and hybrid approaches involve the use of physics-based models (Day and Asce 1985) and rely on both IHCs and seasonal climate forecast for predictive skill (Wood et al. 2016). Both dynamical (Day and Asce 1985; Werner et al. 2004; Wood and Schaake 2008) and hybrid approaches (Robertson et al. 2013; Slater and Villarini 2018) have been developed to address the regression-based limitations posing different degrees of algorithmic complexity and data requirements. Major strengths of these approaches include a continuous generation of plausible future streamflow states and in principle a more physically consistent sensitivity to nonstationary conditions on the basis of model representations of physical process. However, these approaches can present considerable complexity in identifying model parameters and may further necessitate computationally intensive and potentially poorly constrained calibration. In cases where physics-based models perform poorly, embedding machine learning or advanced statistical techniques may allow for better predictions than purely process-driven approaches (Fisher and Koven 2020). Overall, skill from seasonal climate forecast information is currently limited compared to that obtained from IHCs, particularly in snow-dominated settings, such as those presented in his study (Wood et al. 2016).

Regardless of the approach used, the IHCs play a substantial role in the forecast skill of the WSFs (Shukla and Lettenmaier 2011; Wood et al. 2016), particularly across the snow-dominated regions in the West where they provide the majority of predictive information. For example, the NRCS snow-based statistical WSFs have been a widely used tool for streamflow forecast information. They are based on a variety of regression approaches [*Z*-score regression, principal component regression (PCR)] that isolate the contribution of IHCs and minimize the influence of overfitting from predictor's collinearity (Pagano et al. 2009). The dependency of such WSFs on IHCs raises two questions. First is whether using common fixed-date forecasts, for example, initialized on 1 April, provides the maximum predictive skill, and second, is whether overall forecast performance in drought years is comparable to normal, nondrought years. Historically, 1 April has been associated with peak SWE conditions and has been considered to provide maximum predictive information (Pagano et al. 2004). Despite the contemporary forecast skill of 1 April SWE, peak SWE has been projected to occur closer to 1 March for 62% of snow-dominated regions by the end of the century, driven largely by climate warming (Livneh and Badger 2020). In addition, long-term historical trends indicate higher geographical variability in peak SWE around 1 April and a substantial increase in snowmelt before 1 April at 42% of stations across the western United States (Musselman et al. 2021). Hence, reductions in 1 April snowpack conditions during drought would portend lower predictive skill of seasonal streamflow volume. As a result, the addition of ancillary nonsnow predictors like precipitation and soil moisture and an earlier surrogate for peak SWE, such as 1 March SWE, are anticipated to mitigate the reduction in SWE-based predictability in future drought years (Koster et al. 2010; Livneh and Badger 2020; Pagano et al. 2009).

Recent studies (He et al. 2016; Livneh and Badger 2020; Sturtevant and Harpold 2019) have largely attributed reduced predictability in drought years from snowpack to the interannual variability in the snowpack–streamflow relationship (Lehner et al. 2017). Drought years are typically accompanied by below-average snowpack conditions and lowered runoff efficiency. Hence, assessing the reliability of snow-based statistical WSFs on a fixed forecast date or training models on predetermined historical years may be insufficient to capture the full potential predictability in drought years. Instead, evaluation of predictive skill at different forecast dates as well as quantifying the influence of training on different historical years (i.e., climatological stratification) is warranted to tackle potential issues of statistical WSFs. Although climatological stratification is not a complex concept, studies such as McInerney et al. (2021), have shown that climatological stratification (based on flow) improves the reliability of subseasonal forecasts of high and low flows. Nevertheless, to our knowledge, no systematic analysis into the impact of climatological stratification on streamflow predictability has been published, at least across the snow-dominated basins in the western United States, possibly due to data availability for training forecast models (e.g., Llewellyn et al. 2018).

Given the above challenges, we conduct a critical evaluation of the snowpack–streamflow relationship during historical drought years to understand changes in predictive performance as a result of both the forecast date, as well as the historical training years selected. Improvements to WSFs have been documented through key methodological developments. For example, Sturtevant and Harpold (2019) show that systematic overprediction of seasonal streamflow volumes from statistical WSFs in drought years can be partially addressed using a nonlinear transformation of predictor variables. Other studies have reported improvements to statistical forecasts through the addition of nonsnow predictors (He et al. 2016; Lehner et al. 2017; Livneh and Badger 2020), hybrid statistical–dynamical approaches (Robertson and Wang 2012; Slater and Villarini 2018), and the development of modular frameworks (Fleming et al. 2021a). As a point of departure from these developments in statistical WSFs, the novelty of this study is first an assessment of the influence of different historical IHCs in training models to make predictions in drought years and second in investigating the evolution of predictive skill at different forecast dates. Motivated by operational methods used by the NRCS, we use a linear regression approach to model the relationship between SWE and April–July streamflow volume in small headwater catchments, seeking a simple model structure with the least number of parameters. We organize past years' April–July streamflow volumes on the basis of their historical percentiles in order to create different subsets of historical IHCs for training the model. The primary drought forecast experiments are designed akin to an imposed nonstationarity, where the most extreme historical drought years, that is, where the April–July streamflow volume is below the 15th percentile (P_{15}) of the historical record, are withheld from the training period. This is done in order to evaluate the utility of different snowpack–streamflow training approaches to capture "unprecedented drought" conditions. Each forecast experiment evaluates predictive skill throughout the entire forecast season beginning on 1 January, allowing us to quantify the sensitivity of skill to different forecast dates. We also explore these forecast experiments in large Upper Colorado River basin (UCRB) basins using a modified NRCS standard procedure as an independent case study. Finally, we explore the potential for a guided stratification of training years based on antecedent SWE conditions to make predictions in drought years, while exploring the implications of this approach for normal and wet years.

2. Methods

We first introduce the statistical model that predicts streamflow based upon snowpack information in small headwater catchments (section 2a). Percentile thresholds of April–July streamflow are used to create different subsets of training years [section 2a(1)], from which a set of forecast experiments are developed to evaluate the impact of different training years on forecast skill in small headwater catchments [section 2a(2)]. These forecast experiments are also assessed over case study's large basins whose streamflow forecasting procedure is separately detailed in section 2a(3). In section 2b, an "adaptive sampling" application is described, which explores the potential improved forecast skill through a guided stratification of training years based on antecedent SWE conditions. A description of all skill metrics and the statistical test is provided in section 2c, while data sources and screening procedures are detailed in section 2d.

a. *Experimental design*

Given the significant contribution of snowmelt to total runoff in snow-dominated basins (Li et al. 2017), we conduct a series of forecast experiments [section 2a(2)] for selected Snowpack Telemetry (SNOTEL) stations and their corresponding U.S. Geological Survey (USGS) stream gauges (Fig. 1), in which snowpack is exclusively used to predict streamflow in order to isolate snowpack predictive skill directly. We fit a simple linear model with SWE as a predictor

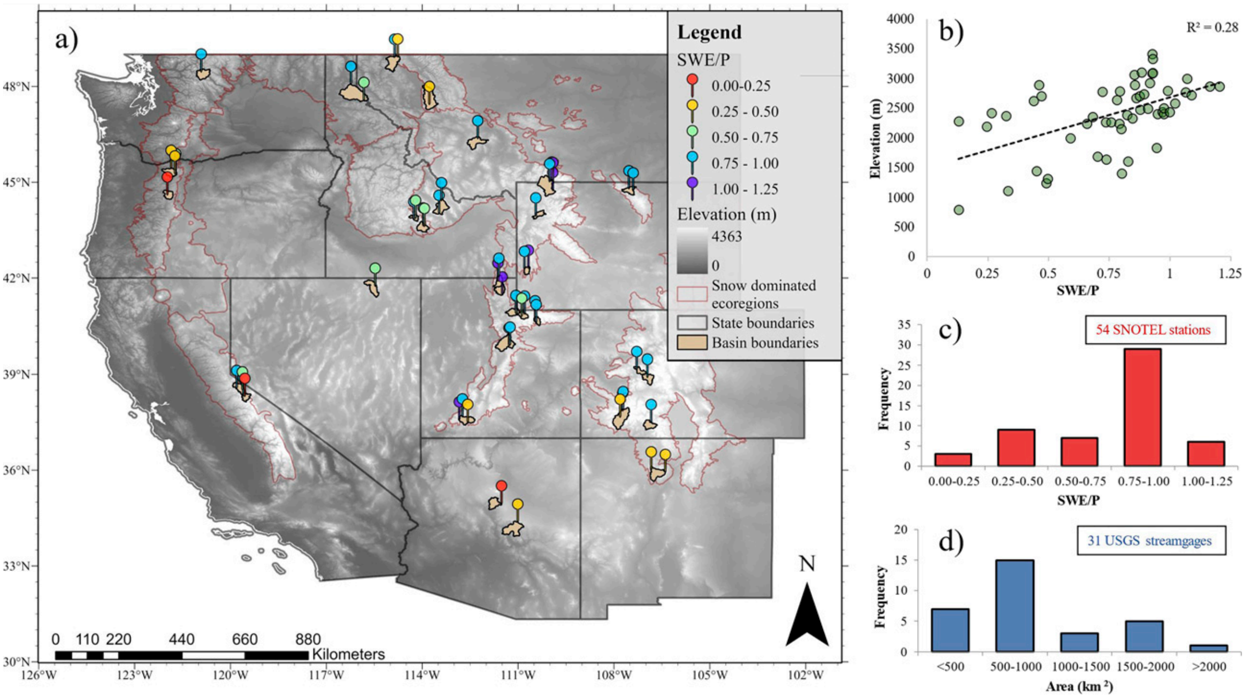

FIG. 1. (a) A map of the study domain, comprising 31 drainage basins and 54 SNOTEL stations across the western United States colored by the ratio of 1 Apr SWE/P, (b) SWE/P plotted against elevation illustrating an overall increase in the fraction of snow with elevation, (c) histogram of the SWE/P, and (d) basin size from selected SNOTEL stations and USGS stream gauges, respectively. A description of the data is provided in section 2d.

and April–July streamflow volume (AMJJ-V) as a predictand and is given in Eq. (1) as

$$Q = a_i \mathrm{SWE}_i + b_i, \quad (1)$$

where Q is the warm season streamflow volume (AMJJ-V), i represents the SWE at a given date (for instance, 1 April), and a and b are the model coefficients. The linear model uses ordinary least squares (OLS) regression rather than the similar approaches (principal component regression or z-score regression) employed by NRCS (Garen 1992) due to the use of a single explanatory variable—SWE, providing deterministic predictions for a given forecast date. We chose a simple linear regression model, in particular, to isolate the predictive value of snowpack and minimize the influence of model parameterization on the forecast errors. Though such a model is easily interpretable and requires minimal computing requirements, it is not ideal when there are data limitations or an emergent physical process that modifies the relationship between predictors and predictand. These cases may necessitate the addition of new observational data as predictors, predictor/predictand transformation, or leveraging information from physically based dynamical models—all of which require careful consideration before operational implementation (Pagano et al. 2009).

1) Flow-based climatological stratification

Transforming meteorological and hydrological conditions such as precipitation, streamflow, soil moisture, reservoir storage, and groundwater levels into percentiles can be a useful, nonparametric way to categorize drought conditions (Steinemann et al. 2015). The U.S. Drought Monitor (USDM) classifies hydrological drought into five major categories using streamflow percentile thresholds, that is, streamflow below these thresholds, including abnormally dry (D0—P_{30}), moderate drought (D1—P_{20}), severe drought (D2—P_{10}), extreme drought (D3—P_5), and exceptional drought (D4—P_2), from the least intense to the most intense (Svoboda et al. 2002). Here, we analyze hydrological drought where the AMJJ-V is below the 15th percentile (P_{15}) of the historical record. We withhold drought years [P_0, P_{15}] from the historical record, that is, years available between 1985 and 2020 water years (WY), of AMJJ-V observations and create a subset of years with the rest [i.e., nondrought years; (P_{15}, P_{100}]] to evaluate the impact of different subsets of training years on the forecast skill during withheld drought years. By withholding drought years, we are effectively assessing predictive skill in unprecedented drought conditions, akin to an imposed nonstationarity.

The historical years are stratified into three categories using percentile thresholds of historical AMJJ-V observations (Fig. 2b): "drought" [P_0, P_{15}]—years withheld for evaluation representing a set of extremely dry years, "below median" (P_{15}, $P_{57.5}$]—years with percentiles lower than the new shifted median (i.e., $P_{57.5\%}$) of the remaining nondrought years, and "above median" ($P_{57.5}$, P_{100}]—years with percentiles above the new shifted median. These subsets were independently derived for each selected basin using their corresponding stream gauge observations.

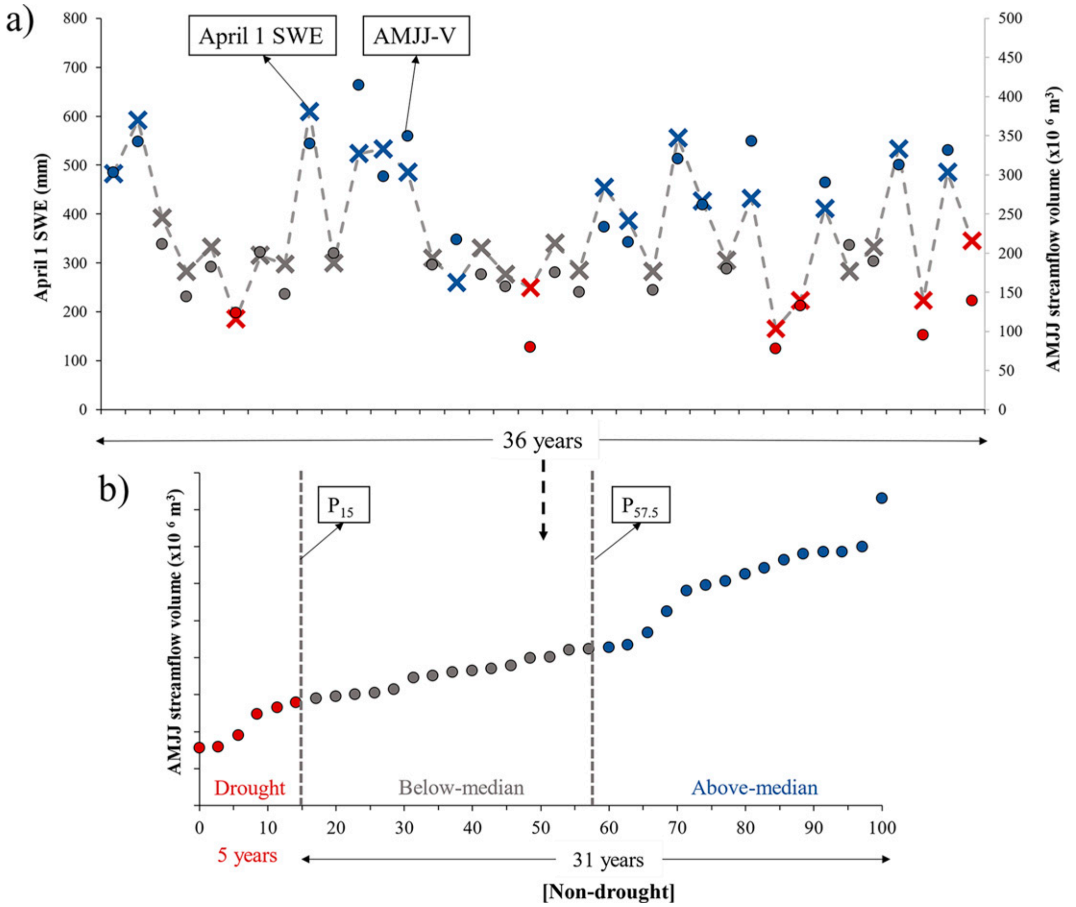

FIG. 2. Example of the experimental design: (a) Time series of 1 Apr SWE (dotted line with "x" markers) and AMJJ-V (solid circles) for 36 historical years. (b) Percentiles based on AMJJ-V are calculated from which three subsets are shown—drought years [P_0, P_{15}]; below-median years (P_{15}, $P_{57.5}$], and above-median years ($P_{57.5}$, P_{100}]. Below-median and above-median are collectively known as nondrought years (P_{15}, P_{100}]. Data are plotted from SNOTEL Butte, CO (380), and USGS East River at Almont, CO (09112500), from 1985 to 2020 water years. Historical data features and screening procedures are described in section 2d.

Figure 3 indicates locally chosen withheld drought years (red filled boxes) in addition to wet [P_{85}, P_{100}] and nonextreme years (P_{15}, P_{85}) for each SWE observation station between 1985 and 2020 WY and primarily represents the spatial variability in historical drought years across the study domain.

2) FORECAST EXPERIMENTS

A set of four forecast experiments were designed to evaluate the impact of different training subsets on the forecast skill and in particular, to evaluate the robustness of WSFs in drought years when trained on different sets of historical years. Four forecast experiments, with different training and evaluation subsets (Fig. 4a), were performed separately for each of the selected 54 SNOTEL observation sites and their corresponding 31 USGS streamflow gauges (full details regarding the observational data and screening procedure is provided in section 2d). We pair SWE at each SNOTEL site with total basin AMJJ-V in order to evaluate the unique relationship that governs snowpack evolution with water supply. In sum, forecast experiments were performed both in a one-on-one fashion as well as using the NRCS approach that averages SWE from all sites within and adjacent to the basin. We perform daily forecasts starting from 1 January through 15 May for each of the experiments using daily SWE and AMJJ-V observations. We choose this time horizon to accommodate the regional differences in the timing of peak SWE (Musselman et al. 2021) and commensurate with the NRCS procedure of issuing forecasts beginning in January (Pagano et al. 2009).

The "conventional" experiment in Fig. 4a follows the practice of training forecast models on long-term historical conditions (usually period of record). Here, the model is trained on the full set of nondrought years and evaluated on withheld drought years. Instead of using the long-term historical conditions predeterminedly, we design a climate-state-based experiment, known as "selective," where the model is trained on below-median years, that is, years exhibiting relatively dry

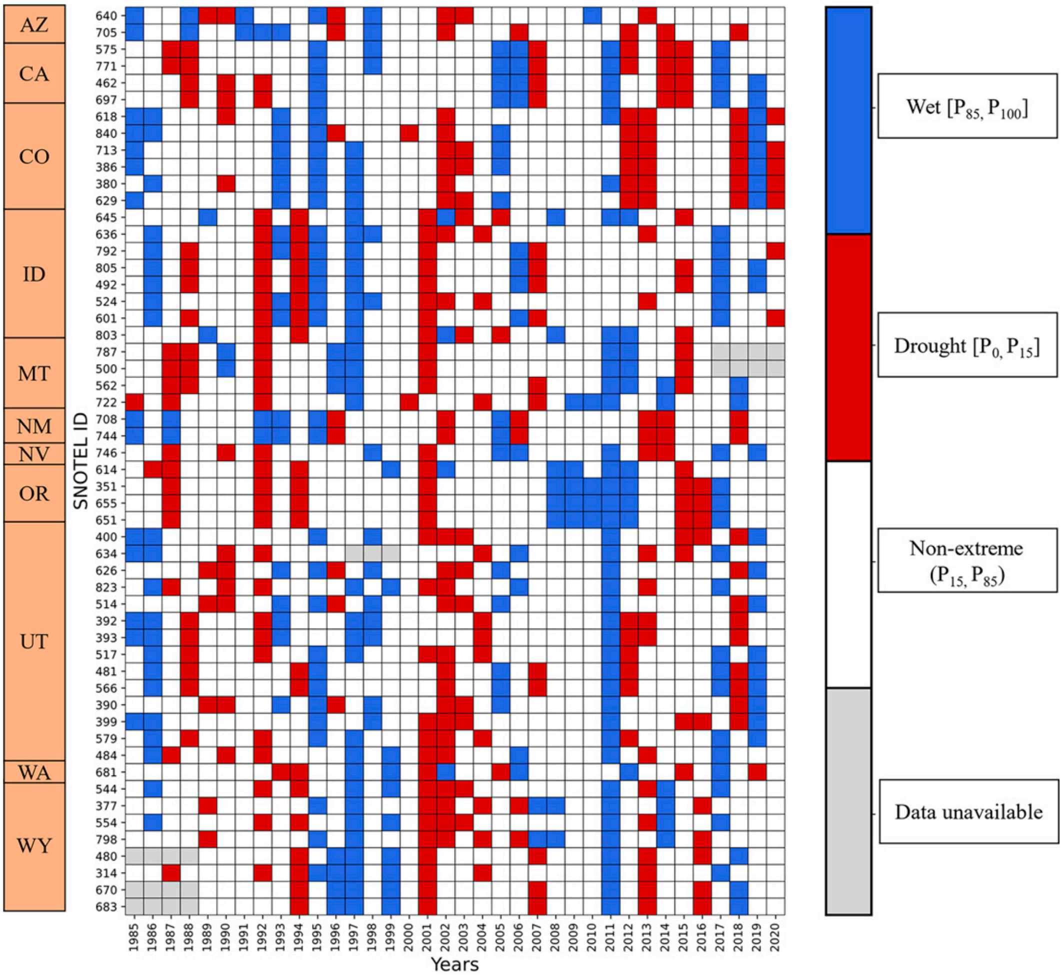

FIG. 3. Annual matrix showing locally chosen drought [P_0, P_{15}], nonextreme (P_{15}, P_{85}), and wet [P_{85}, P_{100}] years for each SNOTEL station. The orange rectangular boxes on the left indicate the state locations of the SNOTEL sites. The gray matrix elements refer to the unavailability of either the SNOTEL SWE or the corresponding stream gauge observations for the marked year.

conditions and evaluated on withheld drought years. To investigate the sensitivity of the "selective" experiment to the range of chosen years, we conduct a separate experiment using four different training subsets: (P_{30}, $P_{62.5}$], (P_{25}, $P_{57.5}$], (P_{20}, $P_{52.5}$], and (P_{15}, $P_{47.5}$], spanning wetter to drier conditions with respect to withheld drought years.

The statistical model, when both trained and evaluated on the same set of years, that is, nondrought years (P_{15}, P_{100}], is expected to reflect the maximum predictive ability of the observations themselves and is referred to as an "overfit" experiment. As a result, it creates a benchmark of forecast skill for all designed experiments. Finally, with the "underfit" experiment, a trade-off scenario is portrayed where the forecast skill in nondrought years is evaluated from the model trained on below-median years. The forecast experiments are illustrated for a representative site along with its corresponding snowpack–streamflow relationship (Fig. 4b). In Fig. 4b, we also illustrate slope in withheld drought years, based on a linear fit between SWE and AMJJ-V. We acknowledge that a linear fit on small sample size (here $n = 6$) is not ideal and may produce biased regression estimates. The sequence of steps associated with the forecast experiments is demonstrated in the top workflow (Fig. 5).

Years in training and evaluation set are chosen independently, that is, we assume a stateless case and therefore are not examining the impact of sequential dependent events, for example, a multiyear drought event on the forecast skill. As a result, forecast skill generated from these experiments can be attributed to the time-independent snowpack–streamflow relationship alone. In a separate experiment, we also compare these forecast experiments by easing the restriction of withheld drought years in training; to represent a de facto scenario assuming that such drought events have occurred in the past. The two training subsets, in this case, include the period of record and actual below-median years [P_0, P_{50}] instead of nondrought and shifted below-median years, respectively.

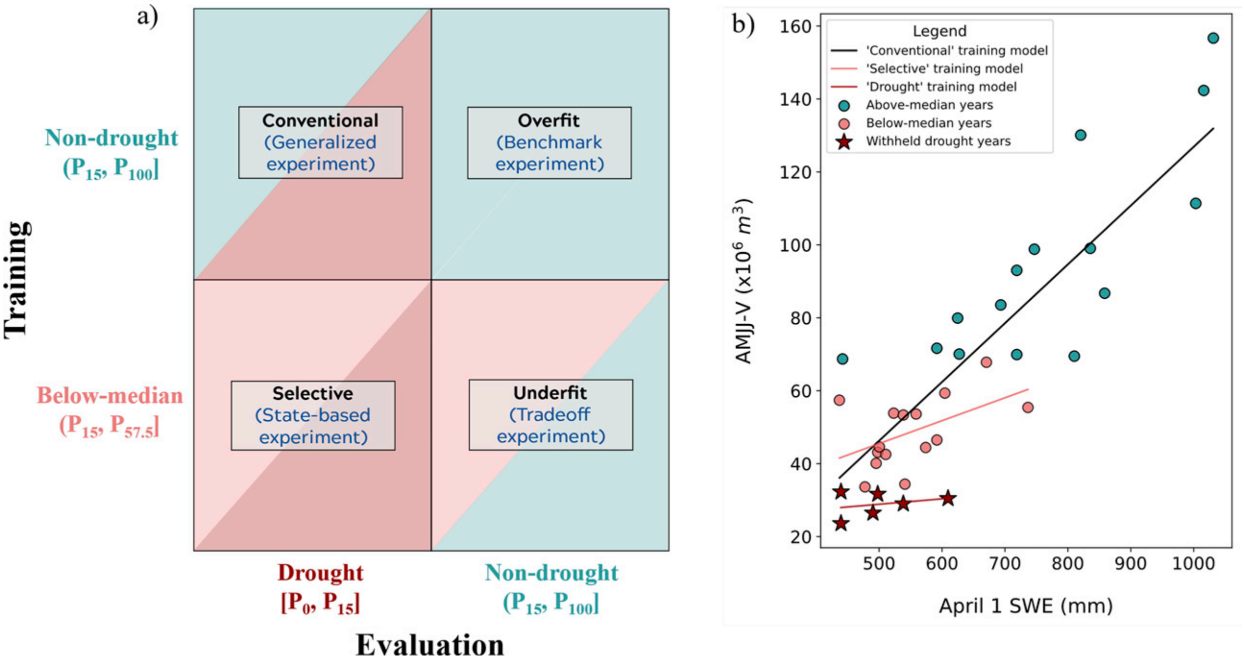

FIG. 4. Design of forecast experiments: (a) Training and evaluation subsets for four forecast experiments where “conventional” and “selective” are evaluated on withheld drought years and trained on nondrought and below-median years, respectively, and “overfit” and “underfit” are evaluated on nondrought years and trained on nondrought and below-median, respectively. (b) Representative site illustrating the snowpack–streamflow relationship showing the training and evaluation subsets, relative to the withheld drought years. Data are plotted from SNOTEL Indian Creek, WY (544), and USGS Hams Fork Below Pole Creek, near Frontier, WY (09223000).

3) CASE STUDY ON NINE LARGE UCRB BASINS: STREAMFLOW FORECASTING PROCEDURE

For greater relevance and to draw more generalizable findings of our work, we perform a case study focusing on nine large UCRB basins where we employ a modified NRCS standard WSF procedure. We compare the forecast skill from the conventional and selective forecast experiments in the withheld drought years by mimicking the operational NRCS forecast procedure of using a PCR. We train PCR on predictors from SNOTEL and naturalized streamflow data from the U.S. Bureau of Reclamation. SNOTEL predictors of SWE and accumulated precipitation are transformed into standardized anomalies (i.e., subtraction of mean and division by standard deviation based on the training years), and AMJJ streamflow volume is seminormalized via a square root transformation (Lehner et al. 2017; Garen 1992). However, a modification to the NRCS procedure is undertaken relating to the process of retaining principal components. While the NRCS procedure (now as NRCS PCR) uses a significance and sign test on regression coefficients to retain the number

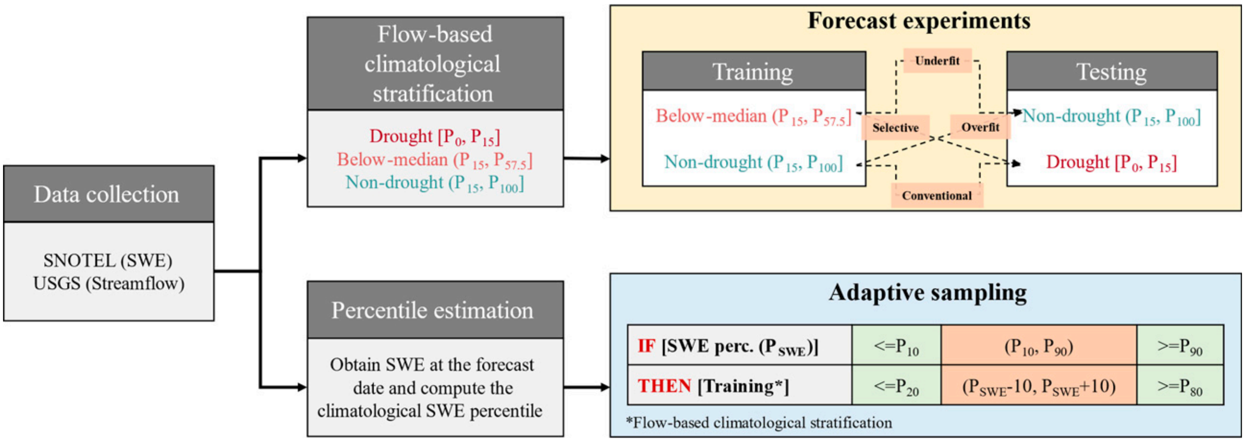
Data collection
SNOTEL (SWE)
USGS (Streamflow)
Flow-based climatological stratification
Drought $[P_0, P_{15}]$
Below-median $(P_{15}, P_{57.5}]$
Non-drought $(P_{15}, P_{100}]$
Forecast experiments
Training
Below-median $(P_{15}, P_{57.5}]$
Non-drought $(P_{15}, P_{100}]$
Underfit
Selective
Overfit
Conventional
Testing
Non-drought $(P_{15}, P_{100}]$
Drought $[P_0, P_{15}]$
Percentile estimation
Obtain SWE at the forecast date and compute the climatological SWE percentile
Adaptive sampling
IF [SWE perc. (P_{SWE})]
<=P_{10}
(P_{10}, P_{90})
>=P_{90}
THEN [Training*]
<=P_{20}
(P_{SWE}-10, P_{SWE}+10)
>=P_{80}
*Flow-based climatological stratification

FIG. 5. Workflow demonstrating the sequence of steps in the (top) forecast experiments and (bottom) adaptive sampling.

of principal components via an iterative process, due to the design of the forecast experiments in our study, a cross-validation approach is used here to retain the principal components (now as CV PCR). Specifically, a tenfold cross validation, that is, a "test" of model on 10 different samples, calculates the model skill score using the mean-squared error, with the addition of the principal component one at a time. The number of principal component/s corresponding to the best model skill score are retained. To evaluate whether the modified method, that is, CV PCR, is consistent with the NRCS PCR, we conduct an additional analysis that compares leave-one-out (or jackknife resampling) errors between the NRCS PCR and CV PCR trained on period of record as well as CV PCR trained on conventional (P_{15}, P_{100}] and selective [P_0, P_{15}] years.

b. *Adaptive sampling—Selection of training years using antecedent SWE conditions*

As an application of the above experiments, we explore the potential for a guided sampling of training years based on antecedent SWE conditions. For a given forecast date, we obtain the SWE conditions on that date and compute the percentile based on the historical SWE record at the calendar date. We create training subsets by selecting years that fall within a range of ±10 percentile points around the computed percentile. A range of ±10 was chosen to maximize the representativeness of SWE states on the sampling of years and satisfy enough data points for training the model. For instance, if the estimated SWE percentile on a given forecast date is 25, then years between the 15th and 35th percentile of AMJJ-V are chosen for training. In the case when the estimated percentile is below 10 or above 90, the years below 20th and above 80th percentile are selected for training. All available years except the evaluation year are included in training the model at a given forecast date. The sequence of steps associated with the adaptive sampling is demonstrated in the bottom workflow (Fig. 5).

c. *Metrics and statistical testing*

Residuals are estimated to determine the model's predictive ability that can be examined through their magnitude and direction. Residuals (e) are expressed as a percentage of the observed median in Eq. (2) as

$$e_i = \frac{(\mathrm{sim}_i - \mathrm{obs}_i)}{\mathrm{median(obs)}}, \tag{2}$$

where sim and obs represent model simulations and observations, respectively, and $i = 1, 2, 3, \ldots, n$, with n being the total number of years in evaluation. We use the normalized root-mean-square error (NRMSE; %) to analyze the predictive skill from the forecast experiments against the corresponding streamflow observations. The normalization of root-mean-square error facilitates comparison across different forecast models and is useful for benchmarking (Hyndman and Koehler 2006). It is expressed as a percentage and shown in Eq. (3) as

$$\mathrm{NRMSE} = \frac{\mathrm{RMSE}}{\overline{\mathrm{obs}}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{sim}_i - \mathrm{obs}_i)^2}}{\overline{\mathrm{obs}}} \times 100(\%), \tag{3}$$

where $\overline{\mathrm{obs}}$ represents mean of observations. A one-sided Wilcoxon signed-rank test is also conducted to determine whether two training models, when evaluated on a similar set of years, have a comparable forecast skill (NRMSE). The nonparametric hypothesis test was chosen over a parametric Student's paired t test as it performs well with nonnormally distributed data. Statistical significance was reported at the 95% confidence level ($\alpha = 0.05$).

In an exploratory analysis, we also assess the relative spread of 1 April SWE and AMJJ-V in historical drought years [P_0, P_{15}] as compared to nondrought years (P_{15}, P_{100}] using the robust relative dispersion metric, the coefficient of median absolute deviation (CMAD). CMAD is resistant to outliers and compares variability reasonably well among different categories of nonnormal distributions (Arachchige et al. 2020). The CMAD here is defined in Eq. (4) and is represented as

$$\mathrm{CMAD} = \frac{\mathrm{med}|x_i - m|}{m}, \tag{4}$$

where "med" denotes the median, m is the median estimate of sample x, and $i = 1, 2, 3, \ldots, n$ with n being the total number of years.

d. *Observational datasets and screening procedure*

Daily SWE observations from the Natural Resource Conservation Service's SNOTEL network and the cumulative seasonal streamflow volume (April–July) estimates from daily USGS National Water Information System (NWIS) data were obtained for SNOTEL sites marked with pins and USGS streamflow gauges corresponding to basins rendered as orange polygons, respectively (Fig. 1a). The water year 1985 is chosen as a starting point as most of the SNOTEL and streamflow observations are continuously available thereafter until 2020. A similar set of years are maintained across each SNOTEL station and corresponding USGS stream gauge to preserve the analysis between SWE and AMJJ-V. The mean annual ratio of 1 April SWE, used here as a proxy for peak SWE (Pagano et al. 2004), to water-year to date cumulative precipitation (SWE/P) is calculated over the water years 1985–2020 (Fig. 1a; continuous precipitation measurements are available at most SNOTEL sites starting from the water year 1985) to ensure and incorporate varying snowpack characteristics across the western United States. A weaker correlation is observed between the SWE/P ratio and elevation at SNOTEL sites, which broadly states that the SWE/P ratio usually increases with elevation (Fig. 1b). It should be noted that a few SNOTEL sites demonstrate inconsistency in the relationship between the snow and precipitation, that is, SWE/$P > 1$, which is due to windy conditions that cause the precipitation gauges to undercatch precipitation and propagate snowdrifts on the measuring snow pillow (Meyer et al. 2012).

For the case study, daily SWE and accumulated precipitation were obtained from SNOTEL, whereas the natural streamflow estimates from the U.S. Bureau of Reclamation (Bureau of Reclamation, accessed February 2022, https://www.usbr.gov/lc/region/g4000/NaturalFlow/). Due to data availability, we constrained our analysis in the case study from 1986 to 2019 WY.

SCREENING PROCEDURE

A diverse set of SWE observation sites and their corresponding drainage basins were selected across the western United States, exhibiting a range of hydroclimatological characteristics and different snow regimes (maritime, continental, and intermountain; Trujillo and Molotch 2014). The following screening procedure was followed to identify basins and snow observations suitable for this analysis:

1) Drainage basin areas were constrained between 350 and 2500 km^2 in size to avoid major over/underrepresentation of basinwide snowpack on streamflow.
2) Drainage basins required at least one SWE station inside the basin boundary or within a 10-km radius for a proximal representation of basinwide snowpack conditions and to serve as a predictor in the statistical model.
3) At least 30 years of SWE and streamflow observations available to support the model training and evaluation.
4) Drainage basins were required to fall within snow-dominated ecoregions (i.e., North American terrestrial level III ecoregions; Barnhart et al. 2016; Wiken et al. 2011) with exceptions to a few basins in Nevada, Arizona, and New Mexico that receive less snowfall in general (Fig. 1a). The basins in these ecoregions have appreciable snow accumulation and they generate snowmelt-driven runoff for downstream communities (Bales et al. 2006).
5) A requirement of minimal anthropogenic influence on streamflow observations from upstream reservoirs, impoundments, and other man-made structures in order for observations to represent a clear connection between snowmelt and streamflow. The identification of such basins was performed by analyzing the geospatial attributes from USGS Geospatial Attributes of Gauges for Evaluating Streamflow (GAGES II; Falcone 2011; Falcone et al. 2010) and Hydro-Climatic Data Network (HCDN; Slack and Landwehr 1992) datasets, which otherwise also recognizes the gauges providing natural streamflow observations.

For the case study, nine large UCRB basins with areas greater than 4000 km^2 (up to 21 000 km^2) were identified based on their availability in U.S. Bureau of Reclamation records and being present in the GAGES II dataset. These basins are usually regulated with reservoirs or interbasin transfers, and estimation of natural flows is performed by using observed streamflow data and removing the human impacts such as effects of irrigation withdrawals or reservoir operations (Bureau of Reclamation, accessed February 2022, https://www.usbr.gov/lc/region/g4000/NaturalFlow/). SNOTEL stations, inside the basin boundary or within a 10-km radius, with continuous data availability of SWE and accumulated precipitation for at least 30 years were selected for consistency.

3. Results

a. *Comparison of forecast skill on 1 April*

The model residuals when trained on below-median (selective) and nondrought (conventional) years are shown for all SNOTEL sites in Fig. 6. Both models show overprediction in drought years. However, consistent with our expectation, the model overprediction is less (smaller residuals, Fig. 6b) with training on below-median years as compared to nondrought years (Fig. 6a). This is evident from NRMSE shown for all SNOTEL sites where overall mean NRMSE dropped, for sites greater than SWE/P of 0.5, by 10% for below-median years (Fig. 6b). This is a consequence of differences in training approaches where, in general, the model slopes are relatively lower for below-median years and similar to the slope in withheld drought years (drought slope) as compared to nondrought years (Figs. 4b and 6c). We observe a general pattern of decreasing model residuals with an increasing SWE/P in both cases, likely due to a greater influence of snowpack on the relationship between snowpack and streamflow.

With nondrought years in training (Fig. 6a), the conventional forecasts show a high degree of variation in residuals across the zero residual line, signaling neither consistent overprediction nor underprediction of AMJJ-V. On the contrary, smaller magnitude and more consistently negative residuals are obtained with the selective forecasts, indicating a systematic overprediction of AMJJ-V. Due to lower SWE values in drought years, high residual errors (>100%) are also observed at a few SNOTEL sites for both training subsets. The regression statistics, including slope, intercept, R^2, and residual standard error, are reported in supplemental Table S1 in the online supplemental material for all SNOTEL sites.

The impact of different training subsets on 1 April forecast skill during drought and nondrought years is examined further and shown in Fig. 7. Similar to the above-described behavior of model residuals, higher forecast skill is obtained in drought years when the model is trained on below-median years (selective), relative to nondrought years (conventional) (Fig. 7a). A consistent gain in skill is observed across all categories of the SWE/P ratio, with a maximum of 20% overall for the SWE/P 0.50–0.75 category. Roughly 74% of locations show better overall performance relative to nondrought training years (Fig. 7b) due to improved fitting of model slopes and lower residuals. Contrary to forecast skill in drought years, we observe the opposite skill pattern in nondrought years (Figs. 7c,d), indicating a trade-off, reduced skill when training on below-median years (underfit) relative to nondrought years (overfit). The drier set of training years lack sampling of nondrought years, and therefore the model cannot reliably capture the relationship between snowpack and streamflow, resulting in high bias. Spatially, streamflow forecasts are considerably more skillful in maritime and intermountain regions (California, Montana, and Idaho) than the continental regions (Colorado and Utah) with below-median years, as shown in Fig. 7b. We remind the reader that the case described above is overly conservative since it assumes that drought years have never occurred before and are not included in the training. However, in a separate experiment, we also find that by including the withheld drought years in training, the gains in forecast skill with below-median years are comparable, albeit slightly better than the above case (Fig. S1).

We further investigate the potential for alternative training subsets to improve skill in drought years. Figure 8a shows the

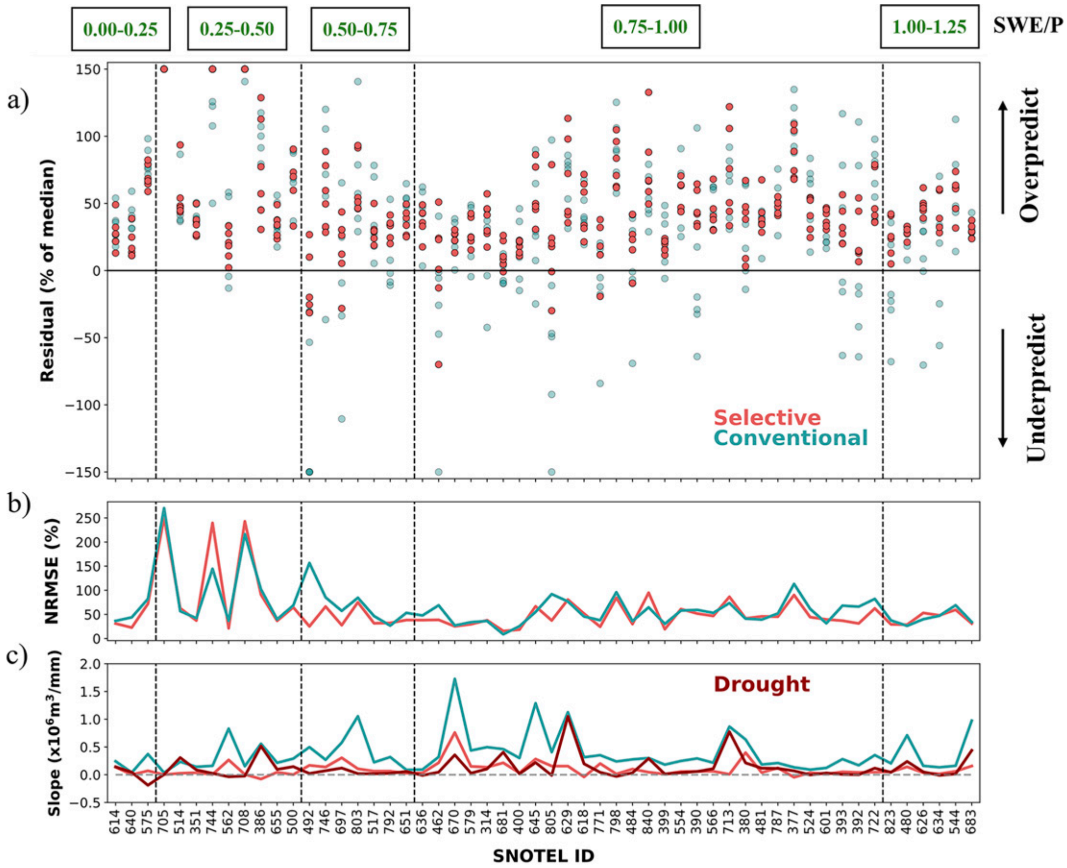

FIG. 6. (a) Model residuals and (b) NRMSE (%) shown for all SNOTEL sites for selective and conventional forecast experiments in withheld drought years, and (c) training model slopes from Conventional and Selective forecast experiments compared to the slope in withheld drought years. Residuals in (a) are expressed as a median percentage of the observed AMJJ-V from withheld drought years. All model slopes in (c) are estimated based on a linear fit between SWE and AMJJ-V.

change in NRMSE for different training subsets relative to nondrought training years across the study domain, with the biggest gains for the driest (P_{15}, $P_{47.5}$] and losses for the least dry (P_{30}, $P_{62.5}$] training subset, respectively. The two driest training subsets (P_{15}, $P_{47.5}$] and (P_{20}, $P_{52.5}$] show significantly better skill (p value $\leq$ 0.05) than nondrought training years (P_{15}, P_{100}] based on a one-sided Wilcoxon signed-rank test. Furthermore, roughly 82% of locations showed better overall performance for the driest training subset relative to nondrought years (not shown). We also assess the change in forecast skill across the SWE/P ratio categories and similarly observe consistent gains and lowest uncertainty for the driest training subset (Fig. 8b).

b. Comparison of forecast skill across the forecast season

Given the interest in water supply predictions throughout the forecasting season (January–May), we assess the impact of different training subsets on the daily forecast skill for each forecast experiment. This comparison is shown for 29 stations with SWE/P ranging from 0.75 to 1.00, representing the largest group of SNOTEL stations and those with high contributions of snowmelt to AMJJ-V. Forecast skill is evaluated for drought (Fig. 9a) and nondrought (Fig. 9b) years for a continuous set of forecast dates spanning from 1 January to 15 May. As shown in Fig. 9a, significant error reductions ranging up to 40% are obtained early in the season (January–February) for below-median years (selective) as compared to nondrought years (conventional). On the contrary, poor performance is observed for below-median years (underfit) relative to nondrought years (overfit) resulting from the lack of information in the context of nondrought years (Fig. 9b). We also identify the calendar dates corresponding to the lowest median NRMSE and find better overall performance after 1 April for all forecast experiments. This is because these stations are mostly in colder regions like Colorado, Utah, Montana, and Wyoming that, on average, receive snow until mid- to late April and tend to provide robust skill around peak SWE. Similar comparisons are also performed for two other SWE/P ratio categories (0.50–0.75; 1.00–1.25) in drought years and are included in the supplemental material (Fig. S2), showing similar, consistent gains in forecast skill with below-median years. Due to reduced snowmelt contribution to runoff, higher uncertainty and poor

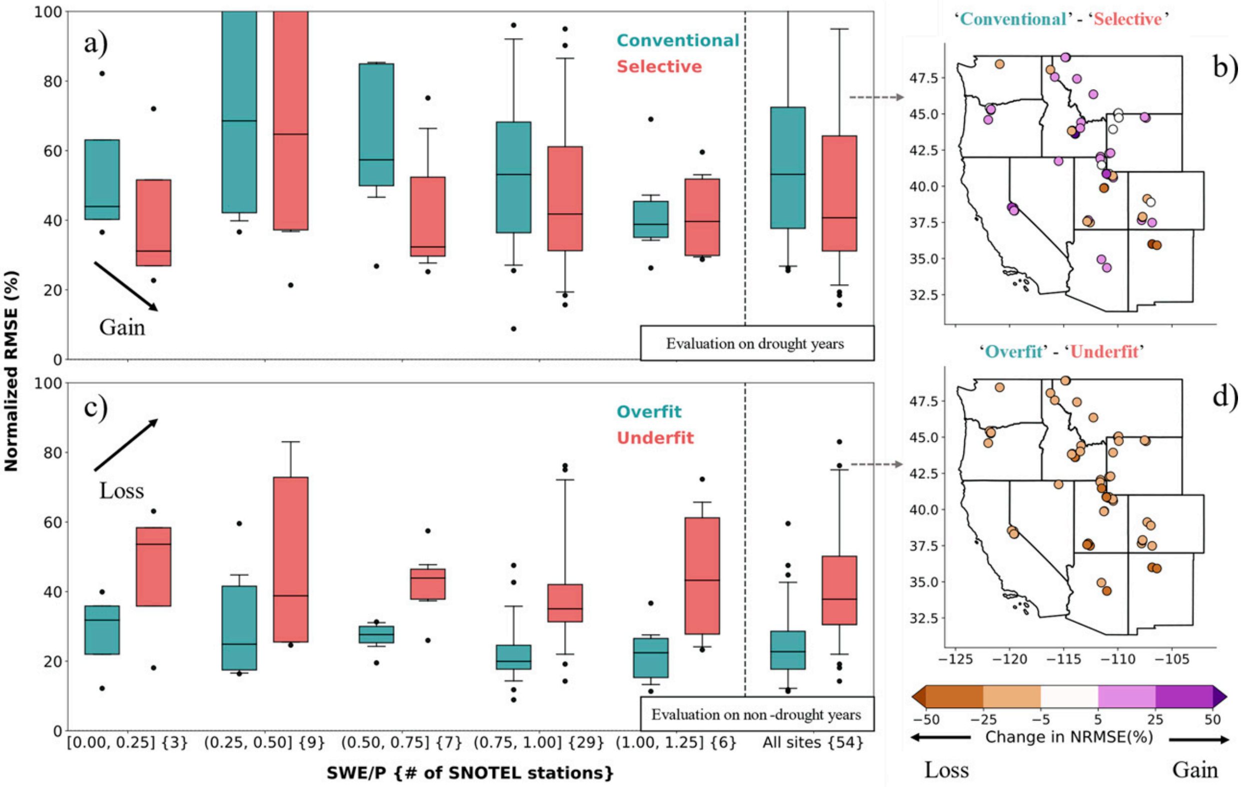

FIG. 7. (a) Forecast skill (NRMSE) evaluated in drought years from the conventional and selective forecast experiments and (b) forecast skill evaluated in nondrought years from the overfit and underfit forecast experiments over the range of SWE/P. (c) Change in NRMSE (%) between the conventional and selective forecast experiments and (d) change in NRMSE between the overfit and underfit forecast experiments across the selected SNOTEL stations. The boxplots in (a) and (c) represent a 90% confidence interval and the curly braces (on the x axis) indicate the number of SNOTEL stations in each SWE/P ratio category.

performance is observed across the forecast season for low SWE/P categories (<0.5). The use of snow as a sole predictor in these cases is likely to become problematic, particularly in low snow and drought years, hence we focus our presentation on results for SWE/$P > 0.5$ categories.

c. *Case study: Comparison of forecast skill in large basins*

We compare the forecast skill from the conventional and selective forecasts, using a modified NRCS's PCR procedure (CV PCR), for nine large UCRB basins to understand the degree of influence of snowpack–streamflow relationship on streamflow generation, particularly in drought years. Prior to our implementation of CV PCR-based forecast experiments, we compare the leave-one-out errors from NRCS PCR and CV PCR and observe similar performance when each are trained on the period of record (Fig. S5). We also find similar performance when training CV PCR on nondrought years (conventional). However, when training on below-median years (selective), large leave-one-out errors at longer lead times (i.e., in January and February) are observed, perhaps attributable to smaller sample sizes (i.e., [P_{15}, $P_{57.5}$] years) and in turn, a larger impact of outliers (Fig. S5). Figure 10a shows the model residuals in withheld drought years for the conventional and selective PCR-based forecasts across different lead times. Commensurate with our earlier findings, we see overprediction in drought years (Fig. 10a, upper subplots) and generally smaller model residuals with selective forecast as compared to conventional forecasts for most basins and across most lead times (see, the NRMSE estimates in Fig. 10a, lower subplots). The performance of conventional and selective forecasts in withheld drought years can be largely explained by the similarity of model slopes, that is, the slope between AMJJ streamflow and SWE, with respect to the slope in the withheld drought years (Fig. 10b). This underscores the importance of the snowpack–streamflow relationship even across larger basins that can aid in improving the understanding of snow-based streamflow predictability.

d. *Improved forecast skill in drought years with adaptive sampling*

We evaluate an adaptive sampling application that dynamically selects training years based on the SWE percentile at every forecast date. We compare the adaptively sampled forecast skill against two alternative training subsets, one using no assumption of a climate state, that is, uses the period of record, excluding the forecast year, and one that trains a dry climate state using below-median years. As shown in Fig. 11a,

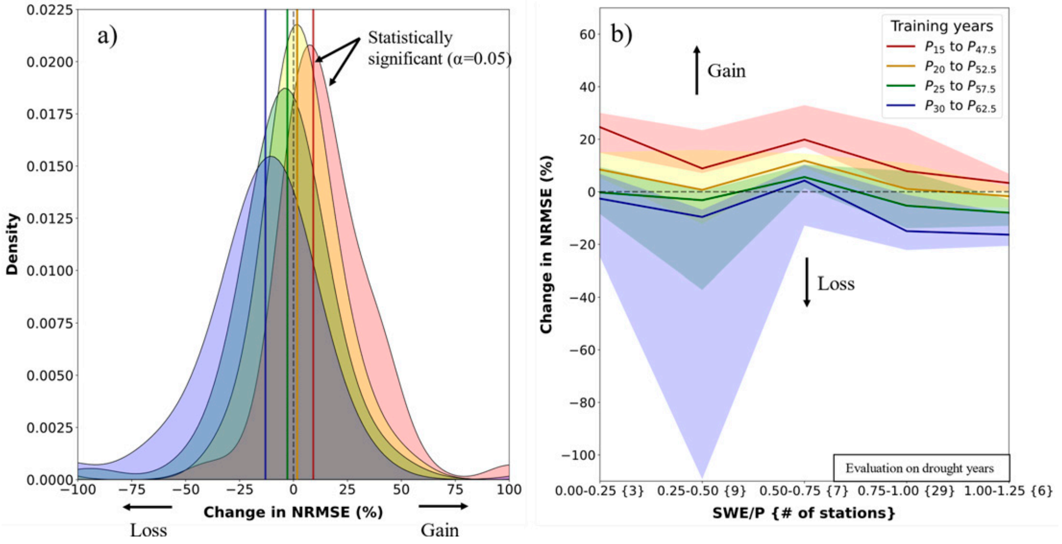

FIG. 8. (a) Change in NRMSE (%) evaluated in drought years across the entire study domain between four different sets of training years and nondrought years (P_{15}, P_{100}] and (b) as in (a), but binned by SWE/P. The median is plotted as solid lines and the interquartile range as a color ribbon. The curly braces in (b) indicate the number of SNOTEL stations in each SWE/P category.

below-median and adaptively sampled years show skillful forecast in drought years when compared to a model trained on the period of record for stations with SWE/P ranging from 0.75 to 1. Consistent error reductions of up to 40%, particularly early in the season, are observed for both, with the largest in below-median years. This is because training on below-median years is geared solely toward drought, whereas, in the case of adaptive sampling, the years are dynamically selected based on antecedent SWE conditions. However, this drought assumption faces considerable uncertainty year-to-year and at longer lead times (Hao et al. 2018), illustrated in Fig. 11c where an incorrect assumption of drought in wet years [P_{85}, P_{100}] can lead to significant forecast errors throughout the forecast season. This is not an issue with adaptively sampled years that rely on antecedent SWE conditions for its assumption of the climate state. Despite moderate error reductions of up to 20% earlier in the season, the skill from adaptively sampled years improves throughout the forecast season in drought years and indeed slightly outperforms the below-median years later in the season (Fig. 11a). With adaptive sampling, a trade-off is seen in "normal years" (Fig. 11b) likely due to training the model on a narrower range of years—spanning only 20 percentile points—relative to training the model on the period of record, which spans nearly 100 percentile points.

This skill improvement of adaptive sampling in drought and wet years is attributable to the evolving relationships and moderate narrowing of SWE and AMJJ-V conditions throughout the forecast season. An example of forecast skill and the time-evolving relationships is shown in Figs. 12a and 12d for drought and Figs. 12c and 12f for wet years at one SNOTEL station. Drawbacks in adaptive sampling can be seen in normal years [$P_{42.5}$, $P_{57.5}$] (Fig. 12b) where it underperforms, in particular, early in the forecast season when the spread among SWE conditions is greatest, becoming narrower by 1 April (Fig. 12e).

4. Discussion

A retrospective analysis was conducted to investigate the snowpack–streamflow relationship and its impact on water supply forecast skill under imposed nonstationary scenarios. This work was motivated by reduced snow-based streamflow predictability in drought years owing to the change in snowpack conditions and lowered runoff efficiency. This analysis into historic forecast skill and training approaches sought to quantify the reliability of snow-based streamflow predictability in the most sensitive management periods, that is, during drought.

Streamflow was overpredicted during drought years, but we found smaller residuals when the model was trained on below-median years as compared to all nondrought years (Fig. 6). Model residuals from training on nondrought years pose high variability across the zero residual line and is the manifestation of the increased 1 April SWE variability in drought years. The distribution of 1 April SWE indicated higher variability in drought years relative to nondrought years, as evident from the CMAD measures (Fig. S3). This is particularly important for cooler continental regions across the western United States where snowfall accumulation variability has been projected to increase toward the end of the twenty-first century (Lute et al. 2015).

Smaller model slopes (shown for a representative site in Fig. 4b) were consistently seen when training the forecast model on below-median years, leading to consistent negative residuals. In these cases, less snow meltwater was reaching the stream gauge, instead contributing more to soil moisture

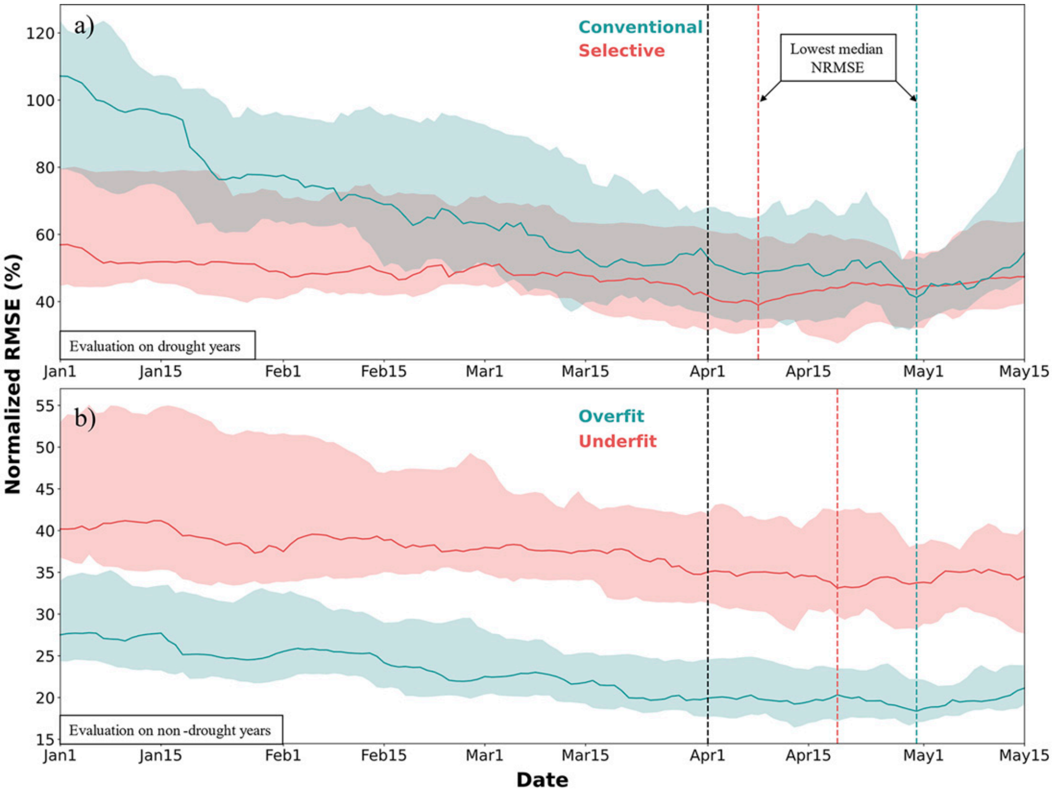

FIG. 9. Forecast skill (NRMSE) during (a) drought and (b) nondrought years across stations with SWE/P ranging from 0.75 to 1.00 from the four forecast experiments. The color ribbons represent the interquartile range with a black line denoting 1 Apr. The colored lines (red and blue) indicate the calendar date corresponding to the lowest median NRMSE for the four forecast experiments (conventional—29 Apr; selective—7 Apr; overfit—18 Apr; underfit—29 Apr).

recharge and evapotranspiration losses to the atmosphere. This lowered runoff efficiency (e.g., Livneh and Badger 2020; Nowak et al. 2012; Woodhouse et al. 2016) means that a model with a lower slope would provide better predictions in drought years due to similarity in slopes between training and evaluation years. However, drawbacks with below-median years can occur, in particular at sites with lower SWE/P in drought years (Fig. 6). Importantly, predictions during extreme drought years, that is, when SWE = 0, solely rely on the model intercepts [see Eq. (1)]. In the case of flatter slopes produced from training on either below-median or nondrought years, these model intercepts sometimes exceed the median of observed streamflow from drought years. This leads to high residual errors, even exceeding 100%, particularly for locations with low SWE/P and where the frequency of zero peak SWE is projected to become increasingly common toward the end of the twenty-first century (Lute et al. 2015; Livneh and Badger 2020). Similar behavior is observed for model residuals at basin-scale that uses the NRCS approach of averaging SWE from SNOTEL sites within and adjacent to the basin (Fig. S4a). This is evident from the NRMSE shown for all basins where overall mean NRMSE dropped by 4% for below-median years (Fig. S4b). The regression statistics, including slope, intercept, R^2, and residual standard error, are reported in supplemental Table S2 for all basins.

Consistent with the above, we observed improvements in seasonal forecast skill derived from 1 April SWE in drought years when training on below-median years. We found that the seasonal forecast skill improved overall at 74% of selected SNOTEL sites with below-median years as compared to nondrought years (Fig. 7). An improvement in skill is further shown with an even drier training subset (P_{15}, $P_{47.5}$] where 82% of SNOTEL sites perform better (Fig. 8). Overall, these results confirm that forecast skill in drought years can be mitigated by selectively training on a subset of years with drier conditions as compared to using nondrought years. The implications of below-median years in training are examined further across the forecast season, where the biggest improvements are seen early in the forecast season (January–February), becoming more comparable later in the season (March–April) relative to training on nondrought years (Fig. 9). This feature could be useful for agricultural, municipal, and industrial sectors that rely on the early season forecast for water transfers and availability estimates. Best predictions are seen after 1 April from all forecast experiments across the stations in colder regions (high SWE/P), hinting toward the potential drawbacks of using 1 April as a proxy to peak SWE (Fig. 9).

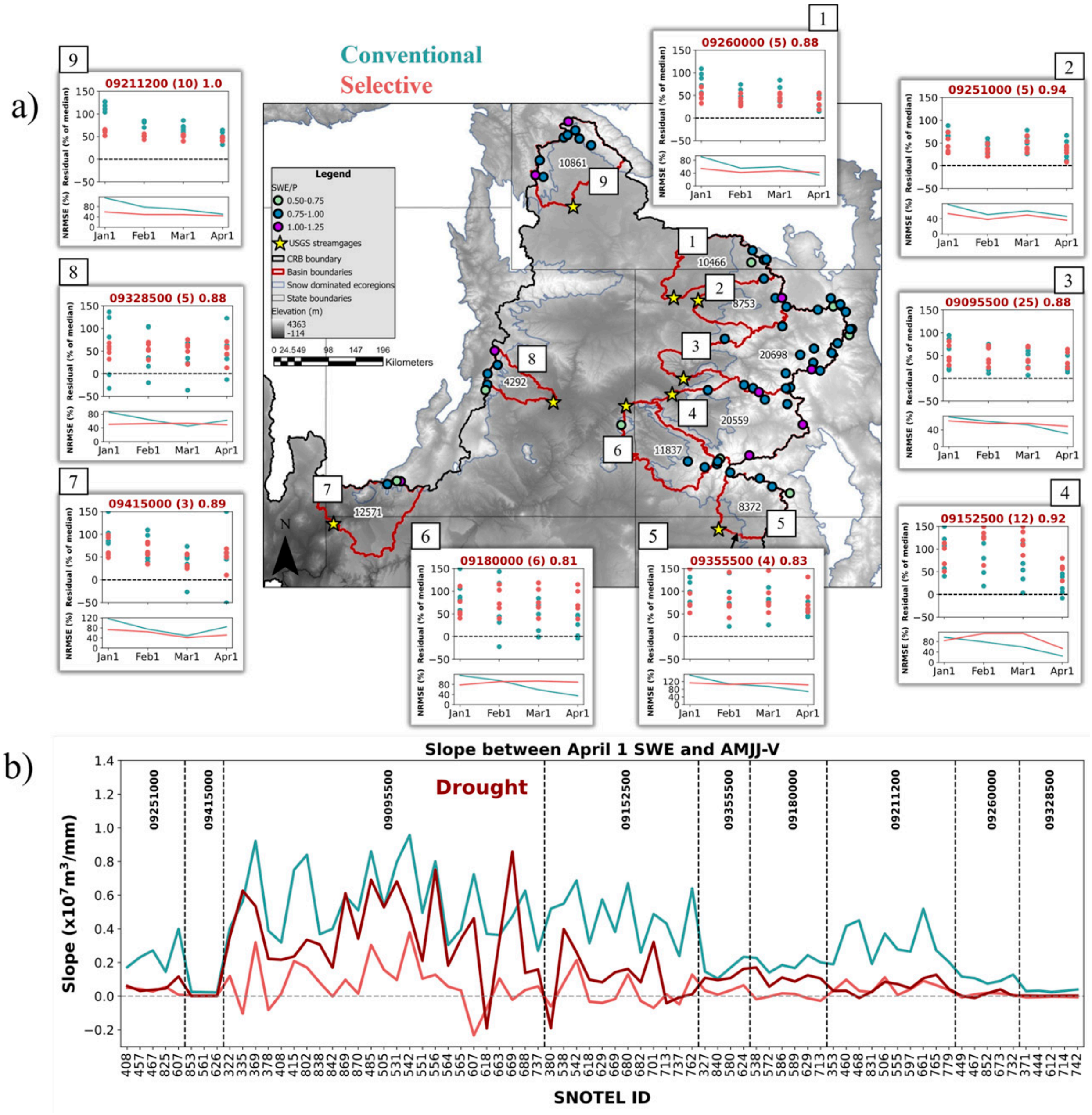

FIG. 10. (a) Model residuals in withheld drought years for the nine large UCRB basins from selective and conventional forecasts. (b) Training model slopes from conventional and selective forecast experiments compared to slopes in withheld drought years. Residuals in (a) are expressed as a median percentage of the observed AMJJ-V from withheld drought years. All model slopes in (b) are estimated based on a linear fit between SWE and AMJJ-V. The halo text in the spatial map within each basin represents the drainage area in units of km^2.

However, with reductions in future snow, the utility of an earlier date like 1 March has been evaluated and shown to perform better toward the end of the century than 1 April (Livneh and Badger 2020).

This forecast experiments in small headwater catchments carries several key limitations. Perhaps most notable is the use of snow as the sole predictor and relying on a simple linear regression approach. We fit a linear model between SWE and AMJJ-V due to its easy interpretation and associated retrospective performance, but such a model clearly neglects the representation of many critical surface processes. Presumably, using additional nonsnow predictors (Koster et al. 2010; Lehner et al. 2017) and more sophisticated forecasting techniques (Sharma and Machiwal 2021) could boost the skill levels achieved. Another limitation is the use of a one-to-one SWE-AMJJ-V relationship throughout the study that captures unique

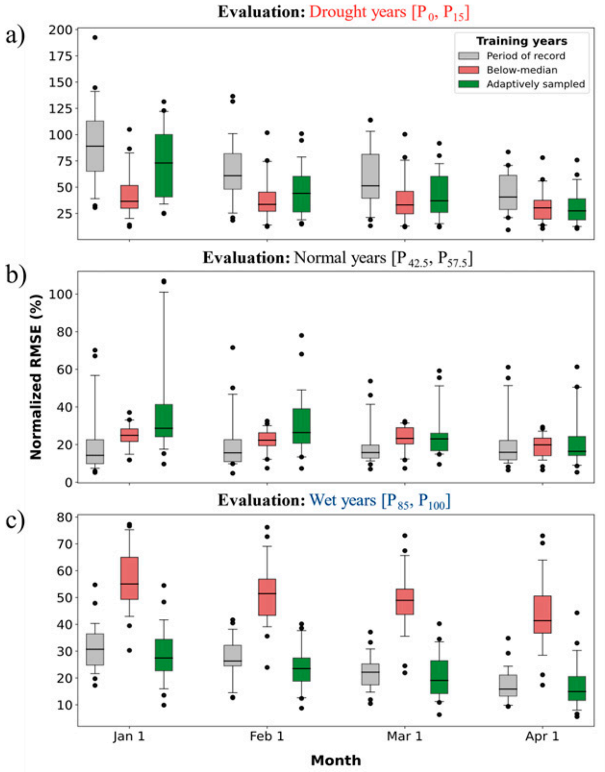

FIG. 11. Forecast skill on the first day of the month for three different training subsets across stations with SWE/P ranging from 0.75 to 1.00 in (a) drought years [P_0, P_{15}], (b) normal years [$P_{42.5}$, $P_{57.5}$], and (c) wet years [P_{85}, P_{100}]. The three training subsets include the period of record, below-median years, and adaptively sampled years. The boxplots represent a 90% confidence interval. Note: the vertical axis range differs for each panel.

relationships between snowpack evolution and water supply. To evaluate the impact of using one-to-one relationships, we repeated our analysis following the NRCS's approach that combines SWE from all sites within and adjacent to the basin and generally observed a similar skill behavior. Despite this, using single or multiple SNOTEL stations still lacks the spatial representativeness of snow conditions across the entire basin. SNOTEL placement, often within local areas of relatively higher snow accumulation regions (Broxton et al. 2019), may not serve as the best proxy for basinwide snowpack conditions overall. We constrained our analysis to those stations with at least 30 years of SWE and AMJJ-V observations, but we acknowledge the limitations in our relatively short historical period.

We attempt to resolve some of the above limitations by incorporating an approach similar in complexity to the NRCS forecasting approach in a separate case study. The impact of different training approaches on forecast performance can be largely reconciled by the characteristics of the snowpack–streamflow relationship (Figs. 6 and 7). However, this relationship does not directly account for impacts like longer lag times, spatial heterogeneity, anthropogenic disturbances, as well as meteorological factors (temperature, wind, humidity, etc.) and physical characteristics (land use, soil type, vegetation, etc.) on streamflow generation in the large basins. Through using larger basins and a different regression approach in our case study (similar to NRCS's PCR procedure), we confirm that the performance of conventional and selective experiments is closely associated with similarity of SWE–streamflow slopes between training and evaluation years (Fig. 10). These slopes are reflective of changing runoff efficiencies between drought and nondrought years.

Nevertheless, an important caveat with these improvements in drought years is they rely on a priori knowledge of a year being in drought or not, which would not be available in a true forecast. Although there have been developments in drought prediction techniques, the anticipation of drought in any forecast year still poses challenges, especially for longer lead times (~3–6 months), due to the inherent unpredictable variability in the atmosphere as well as complex interactions between natural and anthropogenic factors that combine to limit anticipation of future droughts (Hao et al. 2018). In this context, we proposed an adaptive sampling application that dynamically selects training years based on antecedent SWE conditions. We evaluated forecast skill using adaptively sampled training sets relative to training on the entire period of record or using only below-median years. Both the adaptively sampled and below-median training subsets perform better than the period of record in drought and wet years attributable to synchronous relationships between SWE and AMJJ-V (Fig. 11). We believe our exposition into adaptive sampling to be novel mainly in its climatological stratification using initial hydrologic conditions (i.e., antecedent SWE) and its application within a statistical framework. There have been applications analogous to "adaptive sampling" in the streamflow forecasting literature. For example, conditioning the climatology in an ensemble streamflow prediction (ESP) framework with either precipitation or climate indices (Hamlet and Lettenmaier 1999; Werner et al. 2004) or via the selection of hydrologic model parameters based on the climate state (Hay et al. 2009). Regardless, flow-based climatological stratification dependent on the initial hydrologic state within a statistical framework has not been explored yet in a publication to our knowledge. Limitations of adaptive sampling are highlighted in the case of normal years due primarily to the wide spread in SWE conditions relative to AMJJ-V, particularly for forecasts issued early in the forecast season, that is, January and February (Fig. 12), perhaps attributable to training on narrower range of years. The adaptive sampling application is built on a simple model structure and a single predictor that guides a climate state in a given forecast year. Exploring the value of this application with ancillary predictive information from nonsnow predictors like soil moisture and climate indices could provide future opportunities for improved predictions from statistical WSFs. Overall, this work demonstrated that better streamflow predictions with alternate model fitting protocols may offer a useful perspective for decision-makers to consider in snow-based forecasting approaches.

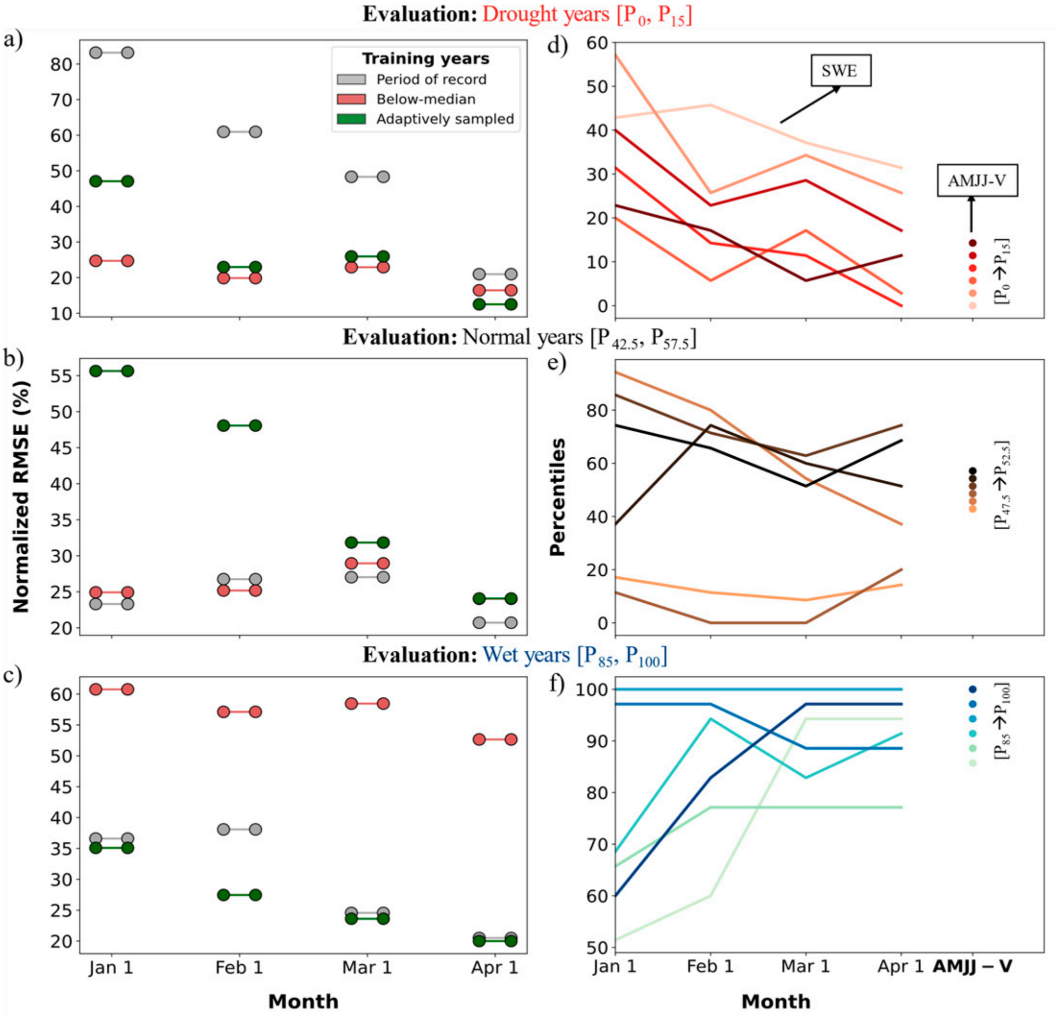

FIG. 12. (a)–(c) Forecast skill (NRMSE) on the first day of each month and (d)–(f) associated SWE (lines) and AMJJ-V (solid circles) percentiles for drought years [P_0, P_{15}], normal years [$P_{47.5}$, $P_{57.5}$], and wet years [P_{85}, P_{100}], respectively. Representation of forecast skill and SWE-AMJJ-V relationship is based on single SNOTEL station 601 (Lost-wood Divide, ID) and its corresponding USGS stream gauge 13120000 (NF Big Lost River at Wild Horse Nr Chilly, ID). Note the vertical axis ranges differ by the panel.

5. Conclusions

We analyzed the skill of seasonal streamflow volume predictions in historical drought years across the western United States and evaluated the impact of different training years on drought forecast skill via designed forecast experiments in small headwater catchments as well as in nine large UCRB basins. The bulk of our analysis withheld severe drought years from the training period, as a way to evaluate the prediction of "unprecedented drought," through a kind of imposed nonstationarity. Our analysis showed that predictability in withheld drought years could be improved by excluding wet years (or above-median years) from the training period. For example, in small headwater catchments, the exclusion of wet years from training period led to forecasts issued on 1 April that showed an overall decrease of 10% in model residuals relative to those forecasts trained on all historical years. This type of improvement was seen in roughly 74% of locations, mostly in colder maritime and intercontinental regions. The best predictions were generally obtained in mid- to late April for the majority of stations, in particular for colder regions. Through our case study over large UCRB basins, we further confirm the importance of the fundamental snowpack–streamflow relationship on streamflow predictability using training protocols more consistent with operations.

We also developed and presented an adaptive sampling application that used the percentile of antecedent SWE conditions on each day of the forecast season to select a set of training years. The adaptively sampled training years produced more skillful forecasts throughout the forecast season in drought years as compared to training on the period of record that poses no assumption of a climate state. Improvements in forecast skill of up to 20% were seen, particularly in drought and extremely wet years, due to the strong coupling between SWE and AMJJ-V conditions earlier in the forecast season. However, these variables were not as tightly coupled when conditions were near the median. The result was that adaptively sampled forecasts performed poorer than those trained on the period of record during "normal years," suggesting that the span of 20 percentile points in adaptive sampling training being too narrow to reflect the snowpack–streamflow relationship during near-median conditions. Overall, the alternate training protocols presented here have the potential to improve the reliability of snow-based forecasting approaches, providing opportunities for addressing the challenges during drought years where water supply information is critical.

Acknowledgments. We acknowledge funding support from the NOAA Grant NA20OAR4310420 Identifying Alternatives to Snow-based Streamflow Predictions to Advance Future Drought Predictability and NSF Grant BCS 2009922 Water-Mediated Coupling of Natural-Human Systems: Drought and Water Allocation across Spatial Scales.

Data availability statement. All data products used in the analysis are publicly available. A total of 54 SNOTEL stations and 31 drainage basins are selected following screening criteria that ensure minimal upstream regulation and continuous data availability for at least 30 years. In addition, nine large UCRB basins and their corresponding 75 SNOTEL sites are selected for the case study. Snowpack observations (SWE) are obtained from the NRCS SNOTEL (https://www.wcc.nrcs.usda.gov/snow/), and the seasonal streamflow volumes are obtained from the U.S. Geological Survey streamflow gages (https://waterdata.usgs.gov/nwis/rt).

REFERENCES

Abatzoglou, J. T., R. Barbero, J. W. Wolf, and Z. A. Holden, 2014: Tracking interannual streamflow variability with drought indices in the U.S. Pacific Northwest. *J. Hydrometeor.*, **15**, 1900–1912, https://doi.org/10.1175/JHM-D-13-0167.1.

Arachchige, C. N. P. G., L. A. Prendergast, and R. G. Staudte, 2020: Robust analogs to the coefficient of variation. *J. Appl. Stat.*, **49**, 268–290, https://doi.org/10.1080/02664763.2020.1808599.

Asefa, T., M. Kemblowski, M. McKee, and A. Khalil, 2006: Multi-time scale stream flow predictions: The support vector machines approach. *J. Hydrol.*, **318**, 7–16, https://doi.org/10.1016/j.jhydrol.2005.06.001.

Bales, R. C., N. P. Molotch, T. H. Painter, M. D. Dettinger, R. Rice, and J. Dozier, 2006: Mountain hydrology of the western United States. *Water Resour. Res.*, **42**, W08432, https://doi.org/10.1029/2005WR004387.

Barnett, T. P., J. C. Adam, and D. P. Lettenmaier, 2005: Potential impacts of a warming climate on water availability in snow-dominated regions. *Nature*, **438**, 303–309, https://doi.org/10.1038/nature04141.

Barnhart, T. B., N. P. Molotch, B. Livneh, A. A. Harpold, J. F. Knowles, and D. Schneider, 2016: Snowmelt rate dictates streamflow. *Geophys. Res. Lett.*, **43**, 8006–8016, https://doi.org/10.1002/2016GL069690.

Broxton, P. D., W. J. D. Van Leeuwen, and J. A. Biederman, 2019: Improving snow water equivalent maps with machine learning of snow survey and lidar measurements. *Water Resour. Res.*, **55**, 3739–3757, https://doi.org/10.1029/2018WR024146.

Cubasch, U., and Coauthors, 2001: Projections of future climate change. *Climate Change 2001: The Scientific Basis*, Cambridge University Press, 525–582, https://www.ipcc.ch/site/assets/uploads/2018/03/TAR-09.pdf.

Daly, S. F., R. Davis, E. Ochs, and T. Pangburn, 2000: An approach to spatially distributed snow modelling of the Sacramento and San Joaquin basins, California. *Hydrol. Processes*, **14**, 3257–3271, https://doi.org/10.1002/1099-1085(20001230)14:18<3257::AID-HYP199>3.0.CO;2-Z.

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

Dettinger, M. D., and D. R. Cayan, 1995: Large-scale atmospheric forcing of recent trends toward early snowmelt runoff in California. *J. Climate*, **8**, 606–623, https://doi.org/10.1175/1520-0442(1995)008<0606:LSAFOR>2.0.CO;2.

Doesken, N. J., and A. Judson, 1996: *The Snow Booklet: A Guide to the Science, Climatology, and Measurement of Snow in the United States.* Department of Atmospheric Science, Colorado State University, 92 pp., https://climate.colostate.edu/pdfs/snowbook.pdf.

Falcone, J. A., 2011: GAGES-II: Geospatial attributes of gages for evaluating streamflow. USGS, accessed 15 April 2021, https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml.

——, D. M. Carlisle, D. M. Wolock, and M. R. Meador, 2010: GAGES: A stream gage database for evaluating natural and altered flow conditions in the conterminous United States. *Ecology*, **91**, 621–621, https://doi.org/10.1890/09-0889.1.

Fisher, R. A., and C. D. Koven, 2020: Perspectives on the future of land surface models and the challenges of representing complex terrestrial systems. *J. Adv. Model. Earth Syst.*, **12**, e2018MS001453, https://doi.org/10.1029/2018MS001453.

Fleming, S. W., and A. G. Goodbody, 2019: A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. *IEEE Access*, **7**, 119 943–119 964, https://doi.org/10.1109/ACCESS.2019.2936989.

——, D. C. Garen, A. G. Goodbody, C. S. McCarthy, and L. C. Landers, 2021a: Assessing the new natural resources conservation service water supply forecast model for the American west: A challenging test of explainable, automated, ensemble artificial intelligence. *J. Hydrol.*, **602**, 126782, https://doi.org/10.1016/j.jhydrol.2021.126782.

——, V. V. Vesselinov, and A. G. Goodbody, 2021b: Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach. *J. Hydrol.*, **597**, 126327, https://doi.org/10.1016/j.jhydrol.2021.126327.

Garen, D. C., 1992: Improved techniques in regression-based streamflow volume forecasting. *J. Water Resour. Plann. Manage.*, **118**, 654–670, https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654).

Guo, J., J. Zhou, H. Qin, Q. Zou, and Q. Li, 2011: Monthly streamflow forecasting based on improved support vector machine model. *Expert Syst. Appl.*, **38**, 13 073–13 081, https://doi.org/10.1016/j.eswa.2011.04.114.

Hamlet, A. F., and D. P. Lettenmaier, 1999: Columbia River streamflow forecasting based on ENSO and PDO climate signals. *J. Water Resour. Plann. Manage.*, **125**, 333–341, https://doi.org/10.1061/(ASCE)0733-9496(1999)125:6(333).

——, P. W. Mote, M. P. Clark, and D. P. Lettenmaier, 2005: Effects of temperature and precipitation variability on snowpack trends in the western United States. *J. Climate*, **18**, 4545–4561, https://doi.org/10.1175/JCLI3538.1.

Hao, Z., V. P. Singh, and Y. Xia, 2018: Seasonal drought prediction: Advances, challenges, and future prospects. *Rev. Geophys.*, **56**, 108–141, https://doi.org/10.1002/2016RG000549.

Hay, L. E., G. J. McCabe, M. P. Clark, and J. C. Risley, 2009: Reducing streamflow forecast uncertainty: Application and qualitative assessment of the upper Klamath River basin, Oregon. *J. Amer. Water Resour. Assoc.*, **45**, 580–596, https://doi.org/10.1111/j.1752-1688.2009.00307.x.

He, M., M. Russo, and M. Anderson, 2016: Predictability of seasonal streamflow in a changing climate in the Sierra Nevada. *Climate*, **4**, 57, https://doi.org/10.3390/cli4040057.

Hyndman, R. J., and A. B. Koehler, 2006: Another look at measures of forecast accuracy. *Int. J. Forecasting*, **22**, 679–688, https://doi.org/10.1016/j.ijforecast.2006.03.001.

Kapnick, S., and A. Hall, 2012: Causes of recent changes in western North American snowpack. *Climate Dyn.*, **38**, 1885–1899, https://doi.org/10.1007/s00382-011-1089-y.

Kişi, Ö., 2007: Streamflow forecasting using different artificial neural network algorithms. *J. Hydrol. Eng.*, **12**, 532–539, https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(532).

Koster, R. D., S. P. P. Mahanama, B. Livneh, D. P. Lettenmaier, and R. H. Reichle, 2010: Skill in streamflow forecasts derived from large-scale estimates of soil moisture and snow. *Nat. Geosci.*, **3**, 613–616, https://doi.org/10.1038/ngeo944.

Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing, 2019: Toward improved predictions in ungauged basins: Exploiting the power of machine learning. *Water Resour. Res.*, **55**, 11 344–11 354, https://doi.org/10.1029/2019WR026065.

Lehner, F., A. W. Wood, D. Llewellyn, D. B. Blatchford, A. G. Goodbody, and F. Pappenberger, 2017: Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the U.S. Southwest. *Geophys. Res. Lett.*, **44**, 12 208–12 217, https://doi.org/10.1002/2017GL076043.

Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier, 2017: How much runoff originates as snow in the western United States, and how will that change in the future? *Geophys. Res. Lett.*, **44**, 6163–6172, https://doi.org/10.1002/2017GL073551.

Livneh, B., and A. M. Badger, 2020: Drought less predictable under declining future snowpack. *Nat. Climate Change*, **10**, 452–458, https://doi.org/10.1038/s41558-020-0754-8.

Llewellyn, D., A. Wood, and F. Lehner, 2018: Runoff efficiency and seasonal streamflow predictability in the U.S. Southwest. Bureau of Reclamation Final Rep. ST-2015-8730-01, 63 pp., https://www.usbr.gov/research/projects/download_product.cfm?id=2760.

Lute, A. C., J. T. Abatzoglou, and K. C. Hegewisch, 2015: Projected changes in snowfall extremes and interannual variability of snowfall in the western United States. *Water Resour. Res.*, **51**, 960–972, https://doi.org/10.1002/2014WR016267.

MacDonald, G. M., and Coauthors, 2008: Climate warming and 21st-century drought in southwestern North America. *Eos, Trans. Amer. Geophys. Union*, **89**, 82–82, https://doi.org/10.1029/2008EO090003.

McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer, and T. Smith, 2019: Making the black box more transparent: Understanding the physical implications of machine learning. *Bull. Amer. Meteor. Soc.*, **100**, 2175–2199, https://doi.org/10.1175/BAMS-D-18-0195.1.

McInerney, D., M. Thyer, D. Kavetski, R. Laugesen, F. Woldemeskel, N. Tuteja, and G. Kuczera, 2021: Improving the reliability of sub-seasonal forecasts of high and low flows by using a flow-dependent nonparametric model. *Water Resour. Res.*, **57**, e2020WR029317, https://doi.org/10.1029/2020WR029317.

Meyer, J. D. D., J. Jin, and S.-Y. Wang, 2012: Systematic patterns of the inconsistency between snow water equivalent and accumulated precipitation as reported by the Snowpack Telemetry network. *J. Hydrometeor.*, **13**, 1970–1976, https://doi.org/10.1175/JHM-D-12-066.1.

Mote, P. W., A. F. Hamlet, M. P. Clark, and D. P. Lettenmaier, 2005: Declining mountain snowpack in western North America. *Bull. Amer. Meteor. Soc.*, **86**, 39–50, https://doi.org/10.1175/BAMS-86-1-39.

——, S. Li, D. P. Lettenmaier, M. Xiao, and R. Engel, 2018: Dramatic declines in snowpack in the western US. *npj Climate Atmos. Sci.*, **1**, 2, https://doi.org/10.1038/s41612-018-0012-1.

Musselman, K. N., M. P. Clark, C. Liu, K. Ikeda, and R. Rasmussen, 2017: Slower snowmelt in a warmer world. *Nat. Climate Change*, **7**, 214–219, https://doi.org/10.1038/nclimate3225.

——, N. Addor, J. A. Vano, and N. P. Molotch, 2021: Winter melt trends portend widespread declines in snow water resources. *Nat. Climate Change*, **11**, 418–424, https://doi.org/10.1038/s41558-021-01014-9.

Nearing, G. S., F. Kratzert, A. K. Sampson, C. S. Pelissier, D. Klotz, J. M. Frame, C. Prieto, and H. V. Gupta, 2021: What role does hydrological science play in the age of machine learning? *Water Resour. Res.*, **57**, e2020WR028091, https://doi.org/10.1029/2020WR028091.

Nowak, K., M. Hoerling, B. Rajagopalan, and E. Zagona, 2012: Colorado River basin hydroclimatic variability. *J. Climate*, **25**, 4389–4403, https://doi.org/10.1175/JCLI-D-11-00406.1.

Pagano, T., and D. Garen, 2005: A recent increase in western U.S. streamflow variability and persistence. *J. Hydrometeor.*, **6**, 173–179, https://doi.org/10.1175/JHM410.1.

——, ——, and S. Sorooshian, 2004: Evaluation of official western U.S. seasonal water supply outlooks, 1922–2002. *J. Hydrometeor.*, **5**, 896–909, https://doi.org/10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2.

Pagano, T. C., 2010: Soils, snow and streamflow. *Nat. Geosci.*, **3**, 591–592, https://doi.org/10.1038/ngeo948.

——, D. C. Garen, T. R. Perkins, and P. A. Pasteris, 2009: Daily updating of operational statistical seasonal water supply forecasts for the western U.S. *J. Amer. Water Resour. Assoc.*, **45**, 767–778, https://doi.org/10.1111/j.1752-1688.2009.00321.x.

Palmer, P. L., 1988: The SCS snow survey water supply forecasting program: Current operations and future directions. *Proc. 56th Annual Western Snow Conf.*, Kalispell, MT, Western Snow Conference, 43–51.

Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat, 2019: Deep learning and process understanding for data-driven Earth system science. *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

Robertson, D. E., and Q. J. Wang, 2012: A Bayesian approach to predictor selection for seasonal streamflow forecasting. *J.*

Hydrometeor., **13**, 155–171, https://doi.org/10.1175/JHM-D-10-05009.1.

——, P. Pokhrel, and Q. J. Wang, 2013: Improving statistical forecasts of seasonal streamflows using hydrological model output. *Hydrol. Earth Syst. Sci.*, **17**, 579–593, https://doi.org/10.5194/hess-17-579-2013.

Serreze, M. C., M. P. Clark, R. L. Armstrong, D. A. McGinnis, and R. S. Pulwarty, 1999: Characteristics of the western United States snowpack from snowpack telemetry (SNOTEL) data. *Water Resour. Res.*, **35**, 2145–2160, https://doi.org/10.1029/1999WR900090.

Sharma, P., and D. Machiwal, 2021: Streamflow forecasting: Overview of advances in data-driven techniques. *Advances in Streamflow Forecasting From Traditional to Modern Approaches*, Elsevier, 1–50.

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.

Slack, J. R., and J. M. Landwehr, 1992: Hydro-Climatic Data Network (HCDN): A U.S. Geological Survey streamflow data set for the United States for the study of climate variations, 1874–1988. USGS Open-File Rep. 92-129, 193 pp., http://pubs.usgs.gov/of/1992/ofr92-129/.

Slater, L. J., and G. Villarini, 2018: Enhancing the predictability of seasonal streamflow with a statistical-dynamical approach. *Geophys. Res. Lett.*, **45**, 6504–6513, https://doi.org/10.1029/2018GL077945.

Steinemann, A., S. F. Iacobellis, and D. R. Cayan, 2015: Developing and evaluating drought indicators for decision-making. *J. Hydrometeor.*, **16**, 1793–1803, https://doi.org/10.1175/JHM-D-14-0234.1.

Stewart, I. T., D. R. Cayan, and M. D. Dettinger, 2004: Changes in snowmelt runoff timing in western North America under a 'business as usual' climate change scenario. *Climatic Change*, **62**, 217–232, https://doi.org/10.1023/B:CLIM.0000013702.22656.e8.

Sturtevant, J. T., and A. A. Harpold, 2019: Forecasting the effects of snow drought on streamflow volumes in the western U.S. *Proc. 87th Annual Western Snow Conf.*, Reno, NV, Western Snow Conference, 1–4.

Suhr Pierce, J. A., and Coauthors, 2010: A measure of snow: Case studies of the snow survey and water supply forecasting program. USDA National Resources Conservation Service, 111 pp., https://www.nrcs.usda.gov/wps/wcm/connect/wcc/e2323413-1532-457c-bdc3-2e76eafbdd47/MeasureofSnowFullReport.pdf?MOD=AJPERES&CVID=nHe-XYF&CVID=nHe-XYF.

Svoboda, M., and Coauthors, 2002: The Drought Monitor. *Bull. Amer. Meteor. Soc.*, **83**, 1181–1190, https://doi.org/10.1175/1520-0477-83.8.1181.

Trujillo, E., and N. P. Molotch, 2014: Snowpack regimes of the western United States. *Water Resour. Res.*, **50**, 5611–5623, https://doi.org/10.1002/2013WR014753.

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

Wiken, E. D., F. J. Nava, and G. Griffith, 2011: *North American Terrestrial Ecoregions—Level III.* Commission for Environmental Cooperation, 149 pp.

Williams, A. P., and Coauthors, 2020: Large contribution from anthropogenic warming to an emerging North American megadrought. *Science*, **368**, 314–318, https://doi.org/10.1126/science.aaz9600.

Wood, A. W., and J. C. Schaake, 2008: Correcting errors in streamflow forecast ensemble mean and spread. *J. Hydrometeor.*, **9**, 132–148, https://doi.org/10.1175/2007JHM862.1.

——, T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016: Quantifying streamflow forecast skill elasticity to initial condition and climate prediction skill. *J. Hydrometeor.*, **17**, 651–668, https://doi.org/10.1175/JHM-D-14-0213.1.

Woodhouse, C. A., G. T. Pederson, K. Morino, S. A. McAfee, and G. J. McCabe, 2016: Increasing influence of air temperature on upper Colorado River streamflow. *Geophys. Res. Lett.*, **43**, 2174–2181, https://doi.org/10.1002/2015GL067613.