Investigating the role of snow water equivalent on streamflow predictability during drought

Parthkumar A. Modi,[a] Eric E. Small,[b] Joseph Kasprzyk,[a] Ben Livneh,[a,c]

[a] *Department of Civil, Environmental and Architectural Engineering, University of Colorado Boulder, Boulder, Colorado, USA*

[b] *Department of Geological Sciences, Boulder, Colorado, USA*

[c] *Cooperative Institute for Research in Environmental Science (CIRES), University of Colorado Boulder, Boulder, Colorado, USA*

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

ABSTRACT

Snowpack provides the majority of predictive information for water supply forecasts (WSFs) in snow-dominated basins across the western US. 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 gages 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, i.e., 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, i.e., 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 NRCS standard procedure in nine large 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 and wet 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.

1. Introduction

In mountainous regions of the western US, 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 US 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 support municipal, agricultural, industrial water supply planning, trade, and power generation (NRCS 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 US. Further, persistent dry states partially attributable to climate warming have

already manifested during the early years of the 21st 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 which represent the underlying physics. Hybrid WSFs consist of multi-model 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 gaged 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 National Resources Conservation Services (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 (M4 — multi-model machine-learning metasystem; 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 1985) and rely on both IHCs and seasonal climate forecast for predictive skill (Wood et al. 2016). Both dynamical (Day 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 non-stationary 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 April 1, provides the maximum

predictive skill, and second, is whether overall forecast performance in drought years is comparable to normal, non-drought years. Historically, April 1 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 April 1 SWE, peak SWE has been projected to occur closer to March 1st 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 April 1 and a substantial increase in snowmelt before April 1 at 42% of stations across the western US (Musselman et al. 2021). Hence, reductions in April 1 snowpack conditions during drought would portend lower predictive skill of seasonal streamflow volume. As a result, the addition of ancillary non-snow predictors like precipitation and soil moisture and an earlier surrogate for peak SWE, such as March 1 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 sub-seasonal 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 US, 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 non-linear transformation of predictor variables. Other studies have reported improvements to statistical forecasts through the addition of non-snow 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 snow water equivalent (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 non-stationarity, where the most extreme historical drought years, i.e., 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 January 1, allowing us to quantify the sensitivity of skill to different forecast dates. We also explore these forecast experiments in large UCRB (Upper Colorado River Basin) 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 2.1). Percentile thresholds of April-July streamflow are used to create different subsets of training years (Sec. 2.1.1), from which a set

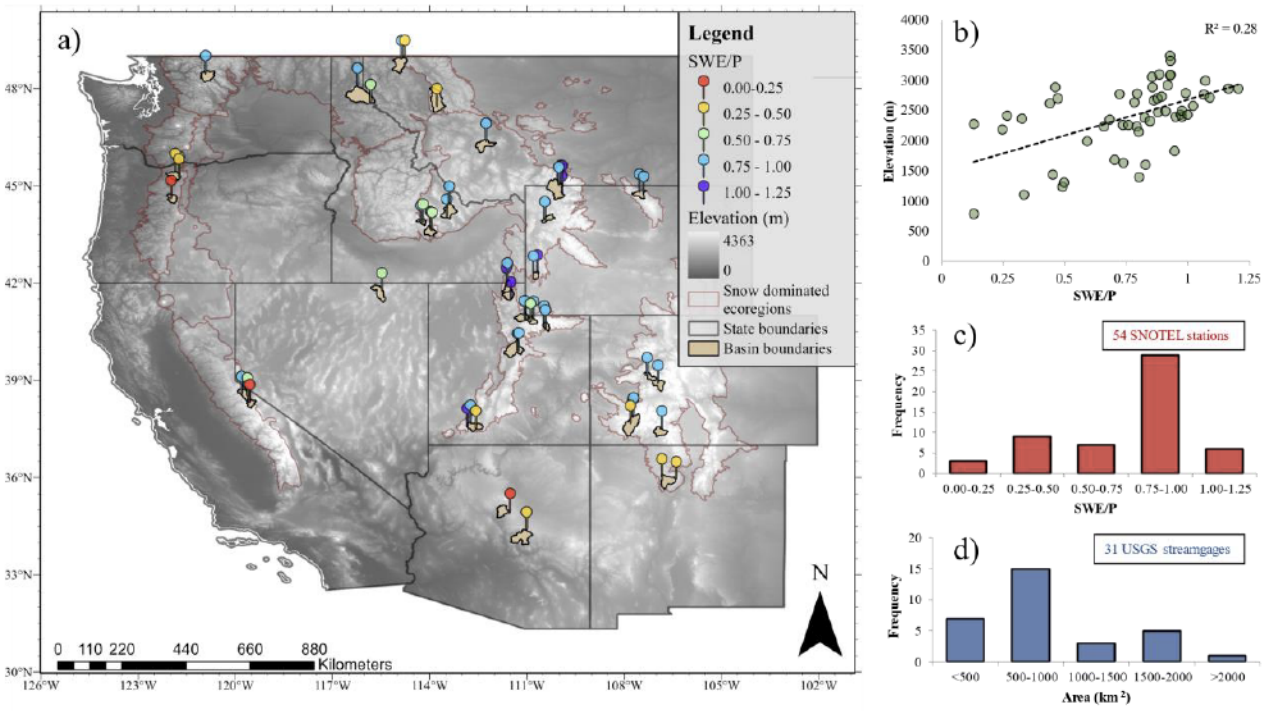
of forecast experiments are developed to evaluate the impact of different training years on forecast skill in small headwater catchments (Sec. 2.1.2). These forecast experiments are also assessed over case study's large basins whose streamflow forecasting procedure is separately detailed in section 2.1.3. In section 2.2, 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 2.3, while data sources and screening procedures are detailed in section 2.4.

2.1 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 (Sec. 2.1.2) for selected SNOTEL stations and their corresponding USGS stream gages (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 and April-July streamflow volume (AMJJ-V) as a predictand and is given in Eq.1 as:

$$Q = a_i SWE_i + b_i \quad (1)$$

Where Q is the warm season streamflow volume (AMJJ-V), i represents the SWE at a given date (for instance, April 1), 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 the National Resources Conservation Service [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 parametrization 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).

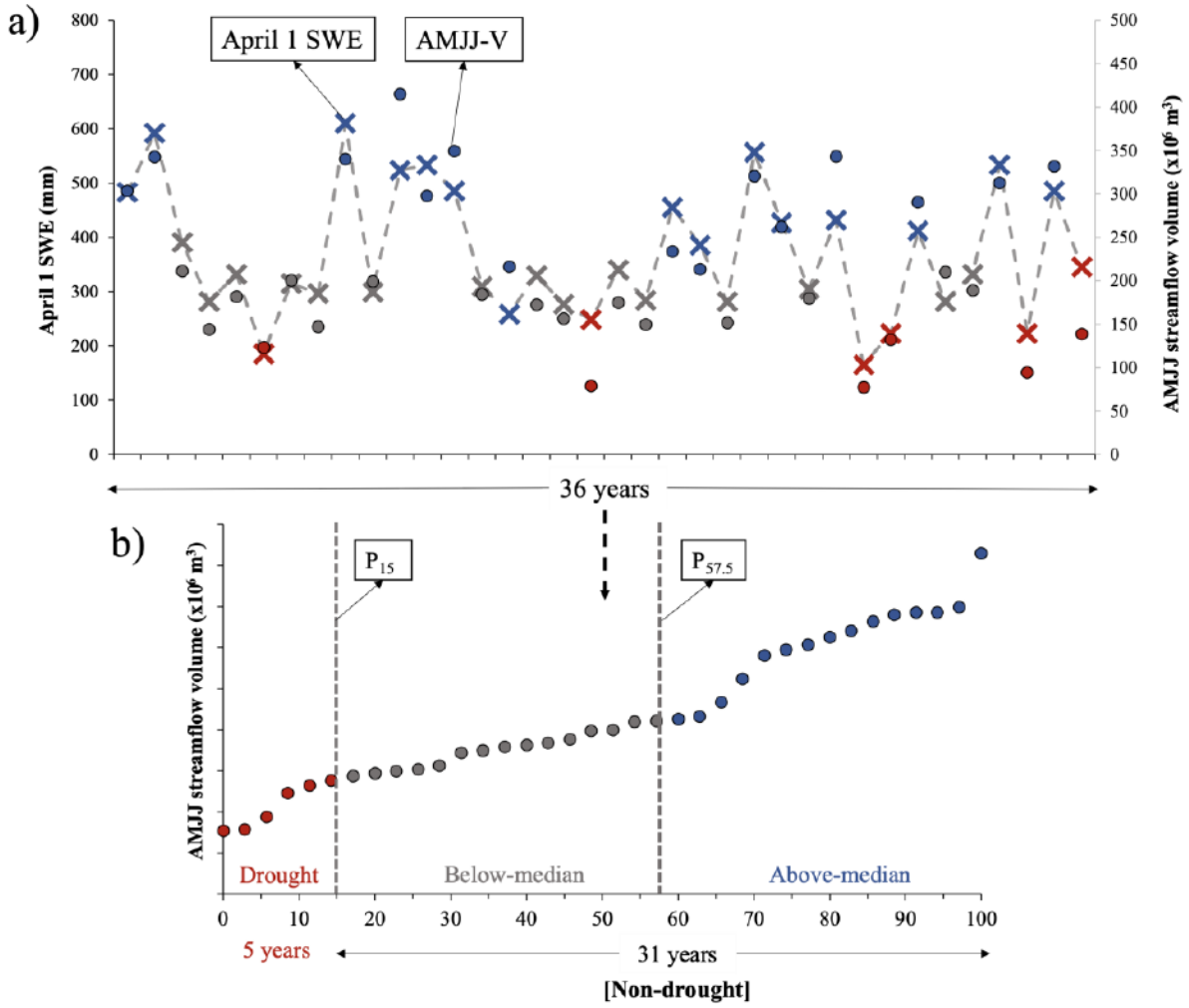


232

233 **Fig. 1.** (a) A map of the study domain, comprising 31 drainage basins and 54 SNOTEL
234 stations across the western US colored by the ratio of April 1 SWE to water-year to date
235 cumulative precipitation ratio (SWE/P), (b) SWE/P plotted against elevation illustrating an
236 overall increase in the fraction of snow with elevation (c) Histogram of the SWE/P and (d)
237 basin size from selected SNOTEL stations and USGS stream gages respectively. A
238 description for the data is provided in Section 2.4.

239 2.1.1 FLOW-BASED CLIMATOLOGICAL STRATIFICATION

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



254

255 **Fig. 2.** Example of the experimental design: (a) Time-series of April 1 SWE (dotted line with
256 "x" markers) and AMJJ streamflow volume (AMJJ-V; solid circles) for 36 historical years.
257 (b) Percentiles based on AMJJ-V are calculated from which three subsets are shown –
258 drought years [P_0, P_{15}]; below-median years (P_{15}, $P_{57.5}$], and above-median years ($P_{57.5}$, P_{100}].
259 Below-median and above-median are collectively known as non-drought years (P_{15}, P_{100}].
260 Data are plotted from SNOTEL Butte, CO (380) and USGS East River at Almont, CO
261 (09112500) from 1985-2020 water years. Historical data features and screening procedures
262 are described in section 2.4.

263 The historical years are stratified into three categories using percentile thresholds of
264 historical AMJJ-V observations (Fig. 2b): "Drought" [P_0, P_{15}] – years withheld for evaluation
265 representing a set of extremely dry years, "Below-median" (P_{15}, $P_{57.5}$] – years with
266 percentiles lower than the new shifted median (i.e., $P_{57.5\%}$) of the remaining non-drought
267 years, and "Above-median" ($P_{57.5}$, P_{100}] – years with percentiles above the new shifted
268 median. These subsets were independently derived for each selected basin using their
269 corresponding stream gage observations. Fig. 3 indicates locally chosen withheld drought

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

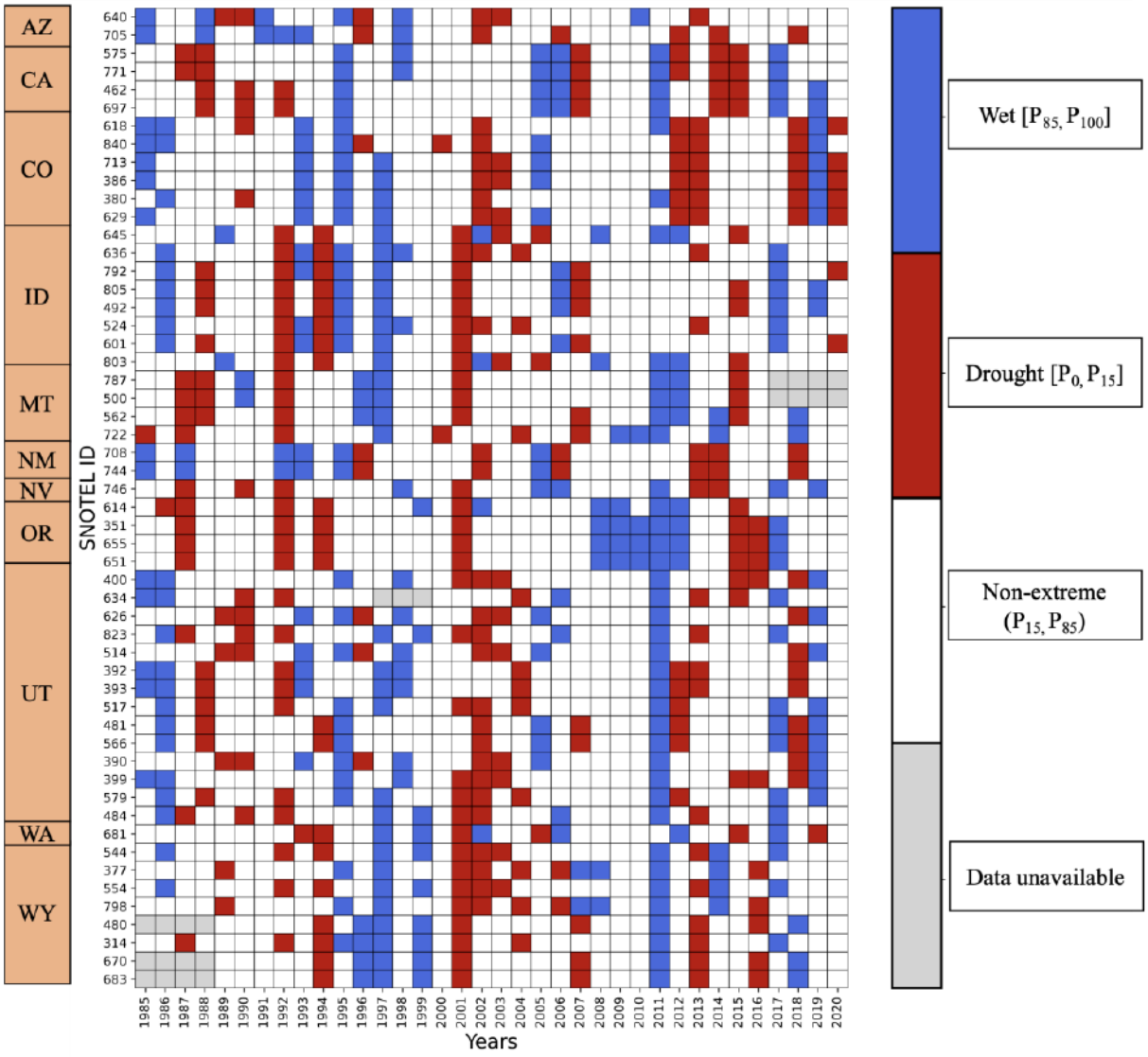


Fig. 3. Annual matrix showing locally chosen drought [P_0, P_{15}], non-extreme (P_{15}, P_{85}), and wet [P_{85}, P_{100}] years for each SNOTEL station. The orange rectangular boxes on the left indicate the state locations of the SNOTEL sites. The grey matrix elements refer to the unavailability of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

2.1.2 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 gages (full details regarding the observational data and screening procedure is provided in section 2.4). 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 January 1 through May 15 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).

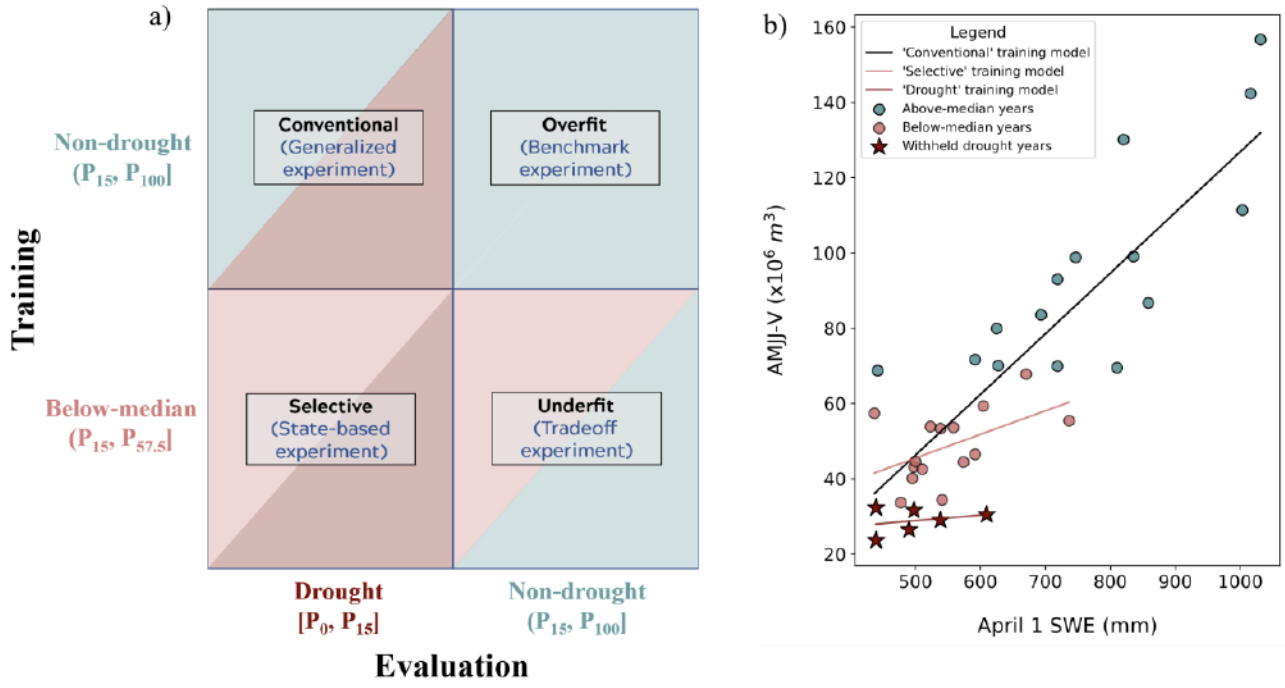


Fig. 4. Design of forecast experiments: (a) Training and evaluation subsets for four forecast experiments where 'Conventional' and 'Selective' are evaluated on withheld drought years and trained on non-drought and below-median years respectively and 'Overfit' and 'Underfit' are evaluated on non-drought years and trained on non-drought 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).

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 non-drought years and evaluated on withheld drought years. Instead of using

306 the long-term historical conditions predeterminedly, we design a climate state-based
307 experiment, known as 'Selective', where the model is trained on below-median years, i.e.,
308 years exhibiting relatively dry conditions and evaluated on withheld drought years. To
309 investigate the sensitivity of the 'Selective' experiment to the range of chosen years, we
310 conduct a separate experiment using four different training subsets: (P_{30}, $P_{62.5}$], (P_{25}, $P_{57.5}$],
311 (P_{20}, $P_{52.5}$], and (P_{15}, $P_{47.5}$], spanning wetter to drier conditions with respect to withheld
312 drought years.

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

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

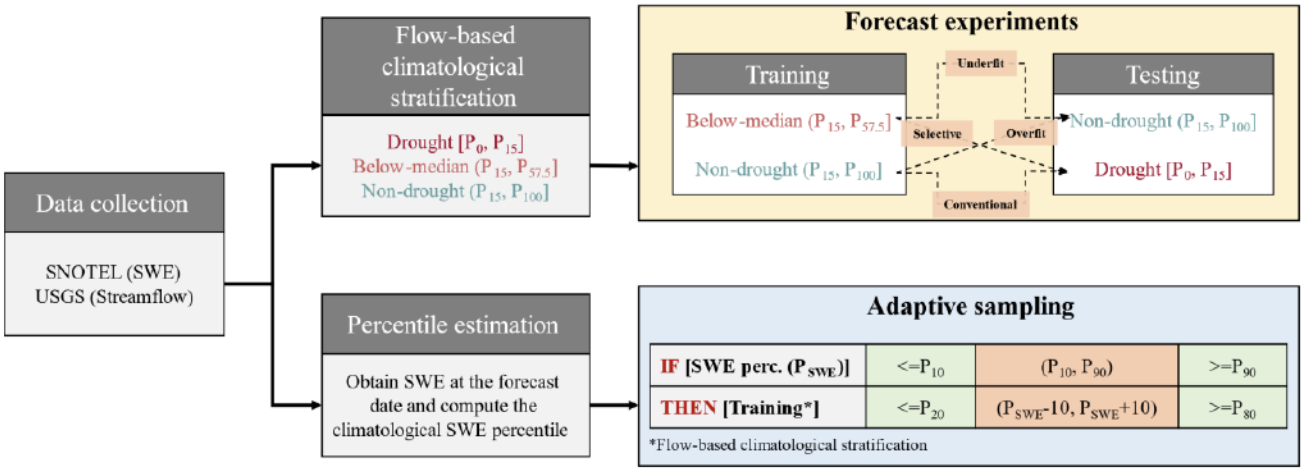


Fig 5: Workflow demonstrating the sequence of steps in the forecast experiments (top) and adaptive sampling (bottom).

2.1.3 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 (Upper Colorado River Basin) 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 Principal Component Regression (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 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 10-fold cross-validation, i.e., a 'test' of model on ten 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, i.e., 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.

2.2 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).

2.3 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:

$$e_i = \frac{\left(sim_i - obs_i\right)}{median(obs)} \quad (2)$$

Where sim and obs represent model simulations and observations, respectively, and i=1, 2, 3, …, n, with n being the total number of years in evaluation. We use the Normalized Root-Mean-Square Error (NRMSE, in %) 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:

$$NRMSE = \frac{RMSE}{\overline{obs}} = \frac{\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(sim_i - obs_i\right)^2}}{\overline{obs}} \times 100\ (\%) \quad (3)$$

Where $\overline{obs}$ 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 non-parametric hypothesis test was chosen over a parametric Student's paired t-test as it performs well with non-normally distributed data. Statistical significance was reported at the 95% confidence level (α=0.05).

In an exploratory analysis, we also assess the relative spread of April 1 SWE and AMJJ-V in historical drought years [P_0, P_{15}] as compared to non-drought 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 non-normal distributions (Arachchige et al. 2020). The CMAD here is defined in Eq. (4) and is represented as:

$$CMAD = \frac{med\left|x_i - m\right|}{m} \quad (4)$$

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

2.4 Observational datasets and screening procedure

Daily SWE observations from the Natural Resource Conservation Service's SNOTEL (SNOwpack TELemetry) network and the cumulative seasonal streamflow volume (Apr-Jul) estimates from daily U.S. Geological Survey's National Water Information System (USGS NWIS) data were obtained for SNOTEL sites marked with pins and USGS streamflow gages 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 gage to preserve the analysis between SWE and AMJJ-V. The mean annual ratio of April 1 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 US. 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, i.e., SWE/P>1, which is due to windy conditions that cause the precipitation gages 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 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 1986-2019 WY.

2.4.1 Screening procedure

A diverse set of SWE observation sites and their corresponding drainage basins were selected across the western US, exhibiting a range of hydro-climatological characteristics and different snow regimes (maritime, continental and intermountain; Trujillo & 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 km^2 to 2500 km^2 in size to avoid major over/under-representation of basin-wide 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 basin-wide 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 Gages for Evaluating Streamflow (GAGES II; Falcone, 2011; Falcone et al., 2010) and Hydro-Climatic Data Network (HCDN; Slack & Landwehr, 1992b, 1992a) datasets, which otherwise also recognizes the gages providing natural streamflow observations.

For the case study, nine large UCRB basins with areas greater than 4000 km^2 (up to 21000 km^2) were identified based on their availability in Bureau of Reclamation records and being present in the GAGES II dataset. These basins are usually regulated with reservoirs or inter-basin 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

3.1 Comparison of forecast skill on April 1

The model residuals when trained on below-median ('Selective') and non-drought ('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 non-drought 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 non-drought years (Fig. 4b & 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 non-drought 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 Supplementary Table S1 for all SNOTEL sites.

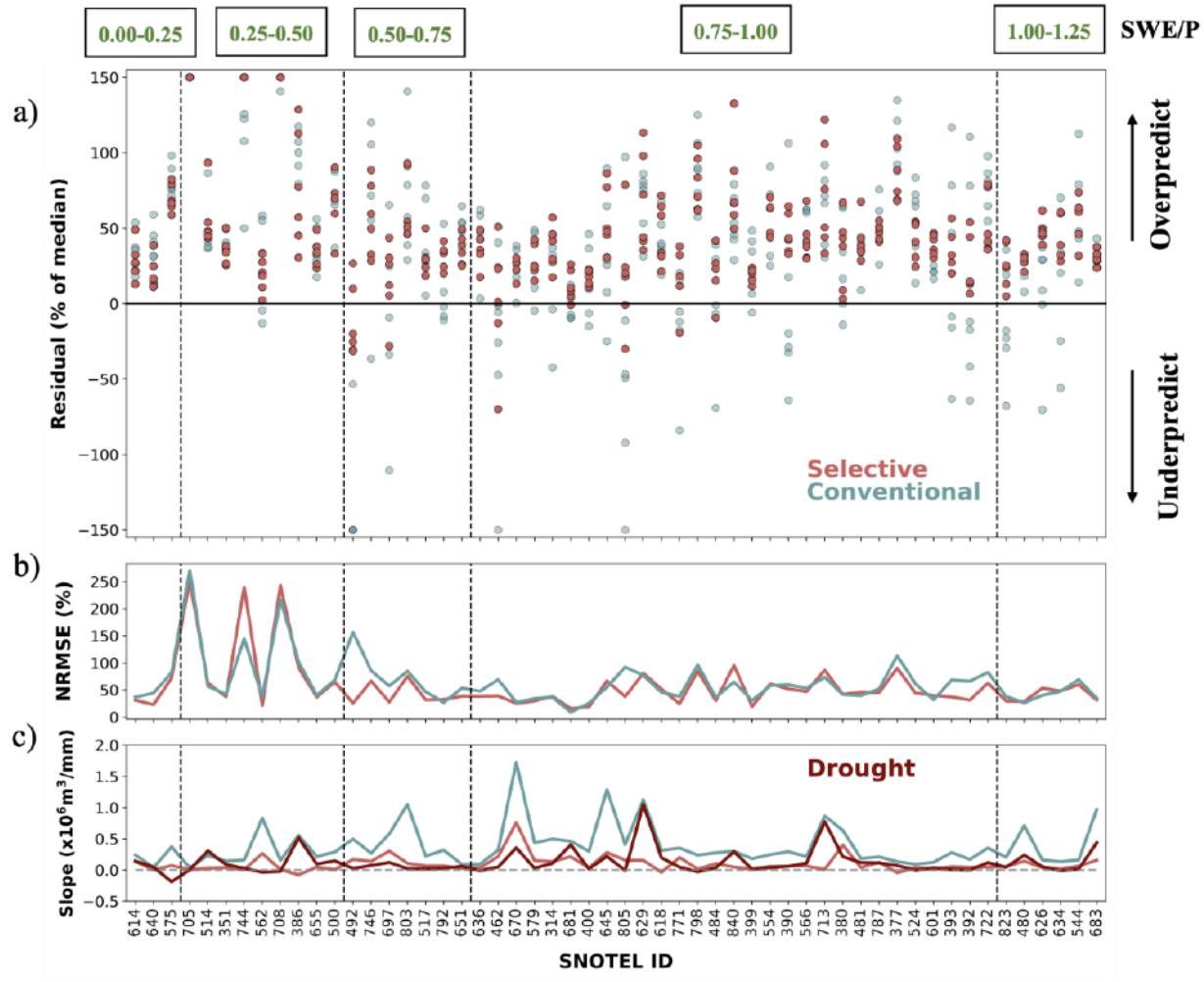
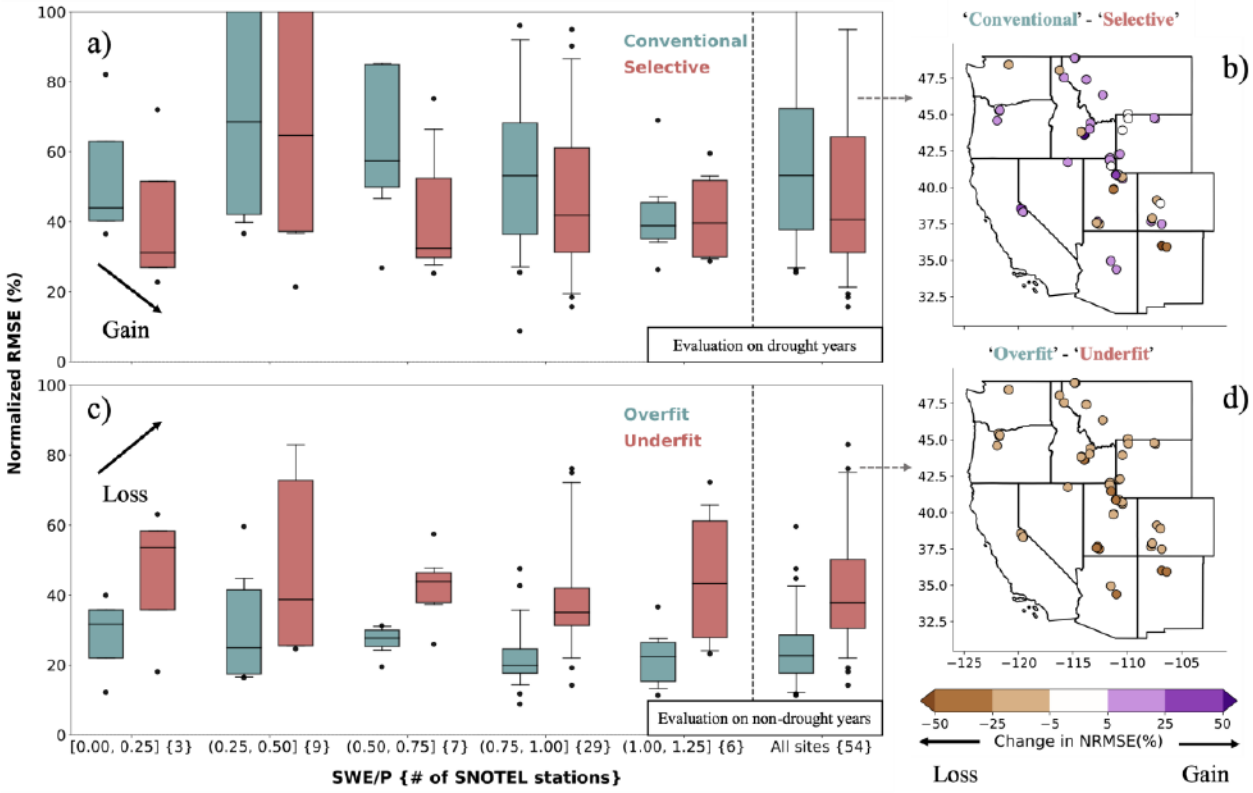


Fig. 6. (a) Model residuals and (b) NRMSE (%) shown for all SNOTEL sites for 'Selective' and 'Conventional' forecast experiments in withheld drought years, and (c) training model slopes from 'Conventional' and 'Selective' forecast experiments compared to the slope in withheld drought years. Residuals 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.

The impact of different training subsets on April 1 forecast skill during drought and non-drought 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 non-drought 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 non-drought 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 non-drought years (Fig. 7c&d), indicating a tradeoff, reduced skill when training on below-median years ('Underfit') relative to non-drought years ('Overfit'). The drier set of training years lack sampling of non-drought 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).



509

510 **Fig. 7.** (a) Forecast skill (NRMSE) evaluated in drought years from the 'Conventional' and
511 'Selective' forecast experiments and (b) Forecast skill evaluated in non-drought years from
512 the 'Overfit' and 'Underfit' forecast experiments over the range of SWE/P. (c) Change in
513 NRMSE (%) between the 'Conventional' and 'Selective' forecast experiments and (d)
514 Change in NRMSE between the 'Overfit' and 'Underfit' forecast experiments across the
515 selected SNOTEL stations. The boxplots (a) and (c) represent a 90% confidence interval and
516 the curly braces (on the x-axis) indicate the number of SNOTEL stations in each SWE/P ratio
517 category.

518 We further investigate the potential for alternative training subsets to improve skill in
519 drought years. Fig. 8a shows the change in NRMSE for different training subsets relative to
520 non-drought training years across the study domain, with the biggest gains for the driest (P_{15},
521 $P_{47.5}$] and losses for the least dry (P_{30}, $P_{62.5}$] training subset, respectively. The two driest
522 training subsets (P_{15}, $P_{47.5}$] and (P_{20}, $P_{52.5}$] show significantly better skill (p-value<=0.05) than
523 non-drought training years (P_{15}, P_{100}] based on a one-sided Wilcoxon signed-rank test.
524 Furthermore, roughly 82% of locations showed better overall performance for the driest
525 training subset relative to non-drought years (not shown). We also assess the change in
526 forecast skill across the SWE/P ratio categories and similarly observe consistent gains and
527 lowest uncertainty for the driest training subset (Fig. 8b).

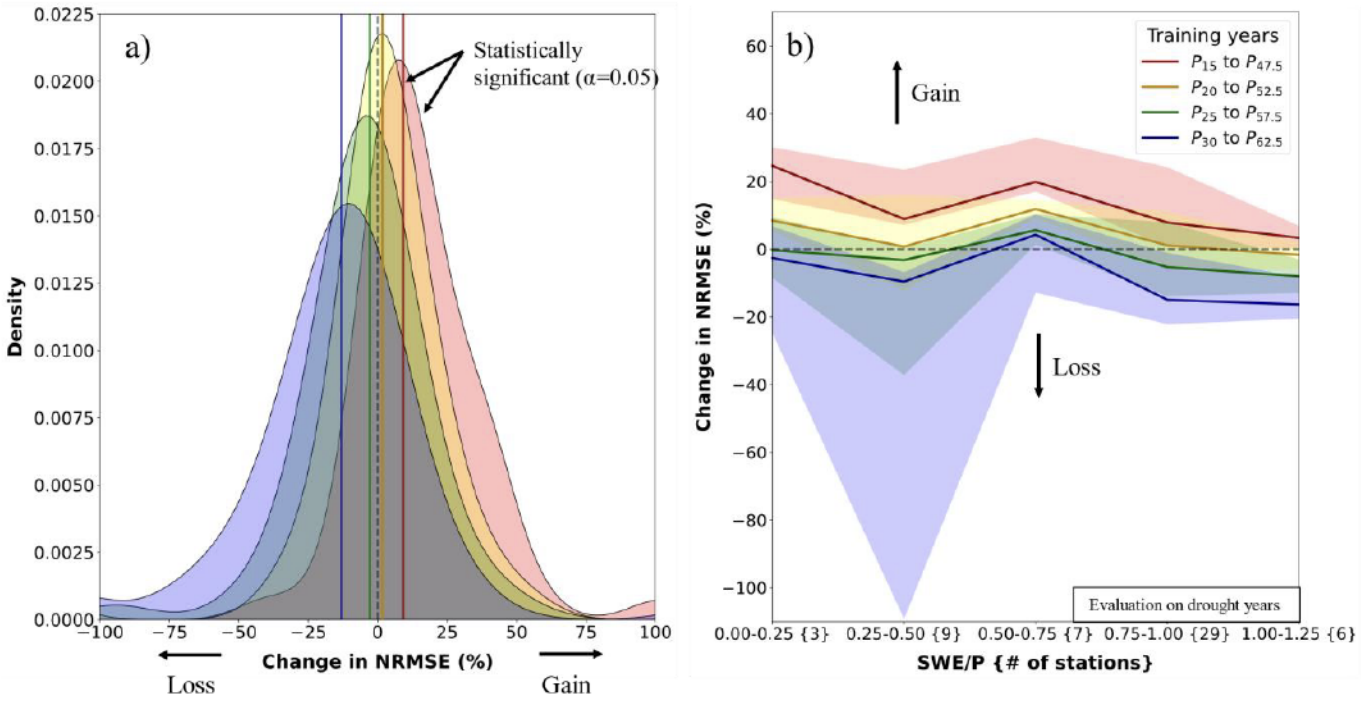


Fig. 8. (a) Change in NRMSE (%) evaluated in drought years across the entire study domain between four different sets of training years and non-drought years (P_{15}, P_{100}] and (b) The same change in NRMSE (%) as (a) but binned by SWE/P. The median is plotted as solid lines and the interquartile range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

3.2 Comparison of forecast skill across the forecast season

Given the interest in water supply predictions throughout the forecasting season (Jan-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 non-drought (Fig. 9b) years for a continuous set of forecast dates spanning January 1 to May 15. As shown in Fig. 9a, significant error reductions ranging up to 40% are obtained early in the season (Jan-Feb) for below-median years ('Selective') as compared to non-drought years ('Conventional'). On the contrary, poor performance is observed for below-median years ('Underfit') relative to non-drought years ('Overfit') resulting from the lack of information in the context of non-drought years (Fig. 9b). We also identify the calendar dates corresponding to the lowest median NRMSE and find better overall performance after April 1 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 Supplementary 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 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.

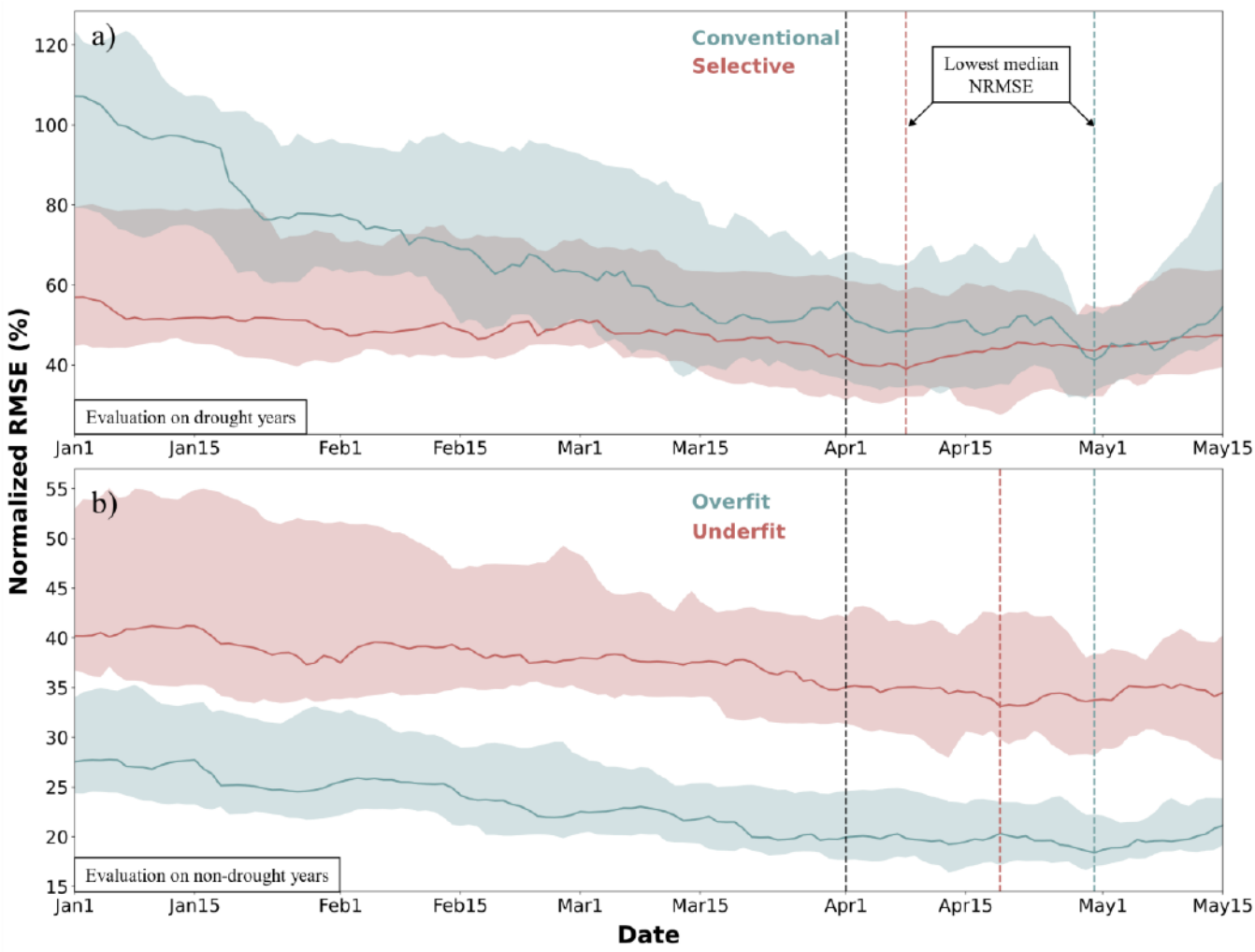


Fig. 9. Forecast skill (NRMSE) during (a) drought and (b) non-drought years across stations with SWE/P ranging from 0.75 to 1.00 from the four forecast experiments. The color ribbons represent the interquartile range with a black line denoting April 1. The colored lines (red & blue) indicate the calendar date corresponding to the lowest median NRMSE for the four forecast experiments ('Conventional' – 29th April; 'Selective' – 7th April; 'Overfit' – 18th April; 'Underfit' – 29th April).

3.3 Case study: Comparison of forecast skill in large basins

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

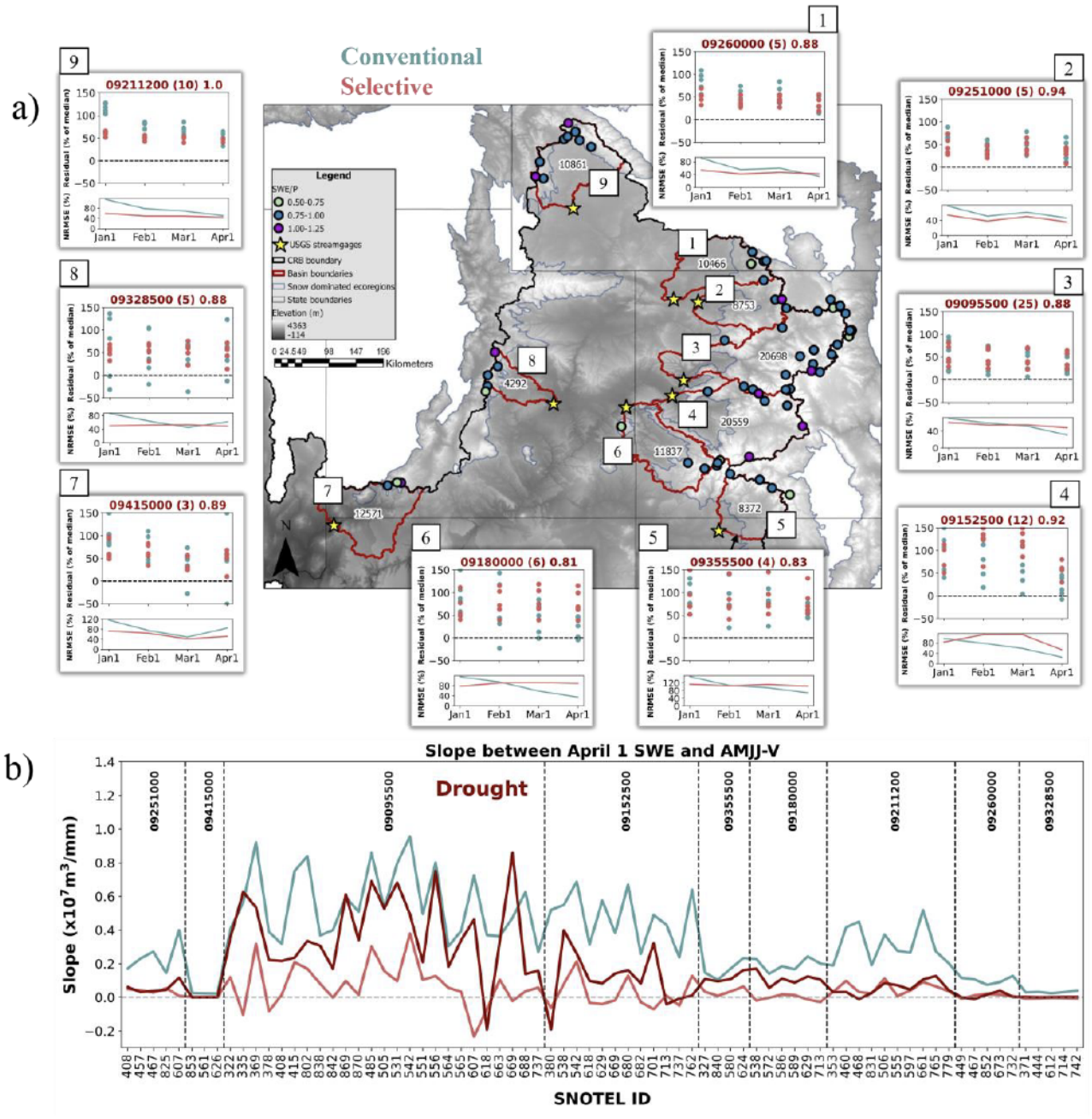
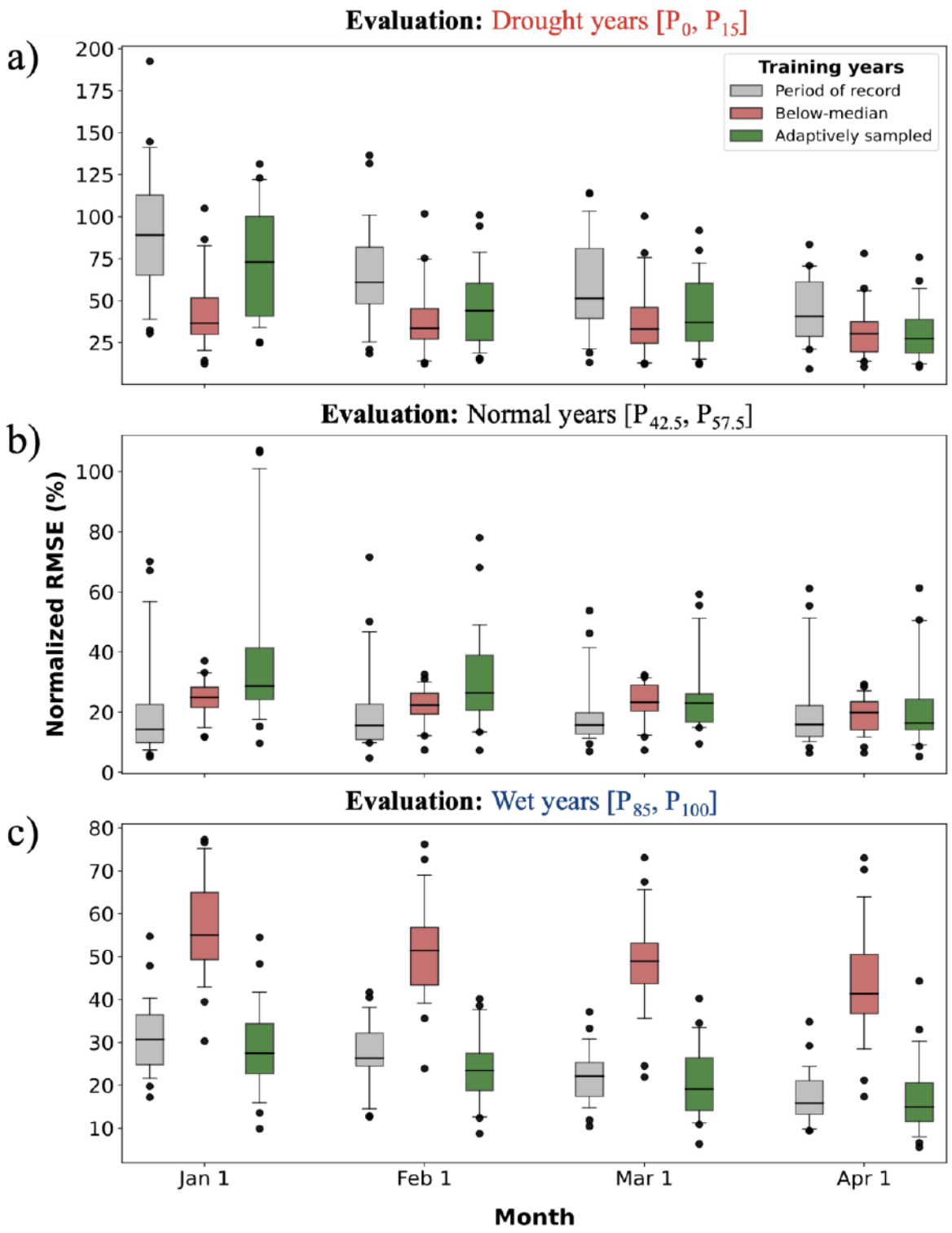


Fig. 10. (a) Model residuals in withheld drought years for the nine large UCRB basins from 'Selective' and 'Conventional' forecasts. (b) Training model slopes from 'Conventional' and 'Selective' forecast experiments compared to slopes in withheld drought years. Residuals in (a) are expressed as a median percentage of the observed AMJJ-V from withheld drought years. All model slopes in (c) 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.

3.4 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 i.e., 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, 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 towards 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 tradeoff 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.



615

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

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 Fig. 12a&d for drought and Fig. 12c&f 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 April 1 (Fig. 12e).

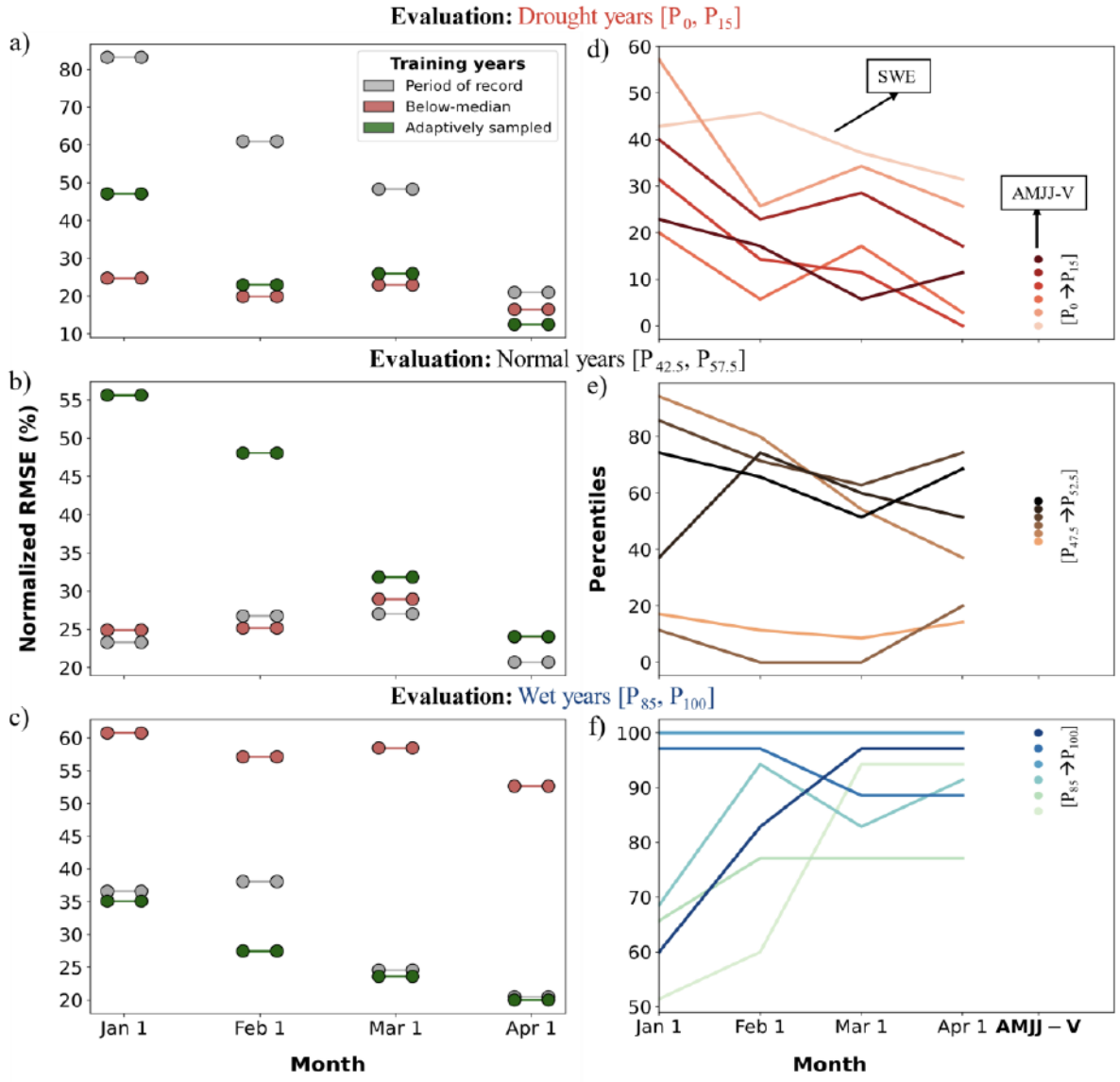


Fig. 12. (a)-(c) Forecast skill (NRMSE) on the first day of each month and (d)-(f) associated SWE (lines) and AMJJ-V (solid circles) percentiles for drought years [P_0, P_{15}], normal years [$P_{47.5}$, $P_{57.5}$], and wet years [P_{85}, P_{100}], respectively. Representation of forecast skill and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

4. Discussion

A retrospective analysis was conducted to investigate the snowpack-streamflow relationship and its impact on water supply forecast skill under imposed non-stationary 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, i.e., 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 non-drought years (Fig. 6). Model residuals from training on non-drought years pose high variability across the zero residual line and is the manifestation of the increased April 1 SWE variability in drought years. The distribution of April 1 SWE indicated higher variability in drought years relative to non-drought years, as evident from the CMAD measures (Fig. S3). This is particularly important for cooler continental regions across the western US where snowfall accumulation variability has been projected to increase towards the end of the 21st 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 snowmelt water was reaching the stream gage, instead contributing more to soil moisture 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, i.e., 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 non-drought 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 towards the end of the 21st century (Lute et al., 2015; Livneh and Badger, 2020). Similar behavior is

667 observed for model residuals at basin-scale that uses the NRCS approach of averaging SWE
668 from SNOTEL sites within and adjacent to the basin (Fig. S4a). This is evident from the
669 NRMSE shown for all basins where overall mean NRMSE dropped by 4% for below-median
670 years (Fig. S4b). The regression statistics, including slope, intercept, $R^{2,}$ and residual standard
671 error, are reported in Supplementary Table S2 for all basins.

672 Consistent with the above, we observed improvements in seasonal forecast skill
673 derived from April 1 SWE in drought years when training on below-median years. We found
674 that the seasonal forecast skill improved overall at 74% of selected SNOTEL sites with
675 below-median years as compared to non-drought years (Fig. 7). An improvement in skill is
676 further shown with an even drier training subset (P_{15}, $P_{47.5}$] where 82% of SNOTEL sites
677 perform better (Fig. 8). Overall, these results confirm that forecast skill in drought years can
678 be mitigated by selectively training on a subset of years with drier conditions as compared to
679 using non-drought years. The implications of below-median years in training are examined
680 further across the forecast season, where the biggest improvements are seen early in the
681 forecast season (Jan-Feb), becoming more comparable later in the season (Mar-Apr) relative
682 to training on non-drought years (Fig. 9). This feature could be useful for agricultural,
683 municipal, and industrial sectors that rely on the early season forecast for water transfers and
684 availability estimates. Best predictions are seen after April 1 from all forecast experiments
685 across the stations in colder regions (high SWE/P), hinting towards the potential drawbacks
686 of using April 1 as a proxy to peak SWE (Fig. 9). However, with reductions in future snow,
687 the utility of an earlier date like March 1 has been evaluated and shown to perform better
688 towards the end of the century than April 1 (Livneh and Badger 2020).

689 This forecast experiments in small headwater catchments carries several key
690 limitations. Perhaps most notable is the use of snow as the sole predictor and relying on a
691 simple linear regression approach. We fit a linear model between SWE and AMJJ-V due to
692 its easy interpretation and associated retrospective performance, but such a model clearly
693 neglects the representation of many critical surface processes. Presumably, using additional
694 non-snow predictors (Koster et al. 2010; Lehner et al. 2017) and more sophisticated
695 forecasting techniques (Sharma and Machiwal 2021) could boost the skill levels achieved.
696 Another limitation is the use of a one-to-one SWE-AMJJ-V relationship throughout the study
697 that captures unique relationships between snowpack evolution and water supply. To evaluate
698 the impact of using one-to-one relationships, we repeated our analysis following the NRCS's

699 approach that combines SWE from all sites within and adjacent to the basin and generally
700 observed a similar skill behavior. Despite this, using a single or multiple SNOTEL stations
701 still lacks the spatial representativeness of snow conditions across the entire basin. SNOTEL
702 placement, often within local areas of relatively higher snow accumulation regions (Broxton
703 et al. 2019), may not serve as the best proxy for basin-wide snowpack conditions overall. We
704 constrained our analysis to those stations with at least 30 years of SWE and AMJJ-V
705 observations, but we acknowledge the limitations in our relatively short historical period.

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

718 Nevertheless, an important caveat with these improvements in drought years is they rely
719 on a priori knowledge of a year being in drought or not, which would not be available in a
720 true forecast. Although there have been developments in drought prediction techniques, the
721 anticipation of drought in any forecast year still poses challenges, especially for longer lead
722 times (~3-6 months), due to the inherent unpredictable variability in the atmosphere as well
723 as complex interactions between natural and anthropogenic factors that combine to limit
724 anticipation of future droughts (Hao et al., 2018). In this context, we proposed an 'adaptive
725 sampling' application that dynamically selects training years based on antecedent SWE
726 conditions. We evaluated forecast skill using adaptively sampled training sets relative to
727 training on the entire period of record or using only below-median years. Both the adaptively-
728 sampled and below-median training subsets perform better than the period of record in
729 drought and wet years attributable to synchronous relationships between SWE and AMJJ-V
730 (Fig. 11). We believe our exposition into 'adaptive sampling' to be novel mainly in its

731 climatological stratification using initial hydrologic conditions (i.e., antecedent SWE) and its
732 application within a statistical framework. There have been applications analogous to
733 "adaptive sampling" in the streamflow forecasting literature. For example, conditioning the
734 climatology in an Ensemble Streamflow Prediction (ESP) framework with either precipitation
735 or climate indices (Hamlet and Lettenmaier 1999; Werner et al. 2004) or via the selection of
736 hydrologic model parameters based on the climate state (Hay et al. 2009). Regardless, flow-
737 based climatological stratification dependent on the initial hydrologic state within a statistical
738 framework has not been explored yet in a publication to our knowledge. Limitations of
739 adaptive sampling are highlighted in the case of normal years due primarily to the wide
740 spread in SWE conditions relative to AMJJ-V, particularly for forecasts issued early in the
741 forecast season, i.e., January and February (Fig. 12), perhaps attributable to training on
742 narrower range of years. The adaptive sampling application is built on a simple model
743 structure and a single predictor that guides a climate state in a given forecast year. Exploring
744 the value of this application with ancillary predictive information from non-snow predictors
745 like soil moisture and climate indices could provide future opportunities for improved
746 predictions from statistical WSFs. Overall, this work demonstrated that better streamflow
747 predictions with alternate model fitting protocols may offer a useful perspective for decision
748 makers to consider in snow-based forecasting approaches.

749 ## 5. Conclusions

750 We analyzed the skill of seasonal streamflow volume predictions in historical drought
751 years across the western US and evaluated the impact of different training years on drought
752 forecast skill via designed forecast experiments in small headwater catchments as well as in
753 nine large UCRB basins. The bulk of our analysis withheld severe drought years from the
754 training period, as a way to evaluate the prediction of 'unprecedented drought', through a
755 kind of imposed non-stationarity. Our analysis showed that predictability in withheld drought
756 years could be improved by excluding wet years (or above-median years) from the training
757 period. For example, in small headwater catchments, the exclusion of wet years from training
758 period led to forecasts issued on April 1 that showed an overall decrease of 10% in model
759 residuals relative to those forecasts trained on all historical years. This type of improvement
760 was seen in roughly 74% of locations, mostly in colder maritime and intercontinental regions.
761 The best predictions were generally obtained in mid to late April for the majority of stations,
762 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 did 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 SNOw TELemetry (SNOTEL) (https://www.wcc.nrcs.usda.gov/snow/), and the seasonal streamflow volumes are obtained from the US Geological Survey streamflow gages (https://waterdata.usgs.gov/nwis/rt).

794

795 REFERENCES

796 Abatzoglou, J. T., R. Barbero, J. W. Wolf, and Z. A. Holden, 2014: Tracking Interannual
797 Streamflow Variability with Drought Indices in the U.S. Pacific Northwest. *Journal of*
798 *Hydrometeorology*, **15**, 1900–1912, https://doi.org/10.1175/JHM-D-13-0167.1.

799 Arachchige, C. N. P. G., L. A. Prendergast, and R. G. Staudte, 2020: Robust analogs to the
800 coefficient of variation. *Journal of Applied Statistics*, 1–23,
801 https://doi.org/10.1080/02664763.2020.1808599.

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

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

808 Barnhart, T. B., N. P. Molotch, B. Livneh, A. A. Harpold, J. F. Knowles, and D. Schneider,
809 2016: Snowmelt rate dictates streamflow. *Geophys. Res. Lett.*, **43**, 8006–8016,
810 https://doi.org/10.1002/2016GL069690.

811 Broxton, P. D., W. J. D. Leeuwen, and J. A. Biederman, 2019: Improving Snow Water
812 Equivalent Maps With Machine Learning of Snow Survey and Lidar Measurements.
813 *Water Resour. Res.*, **55**, 3739–3757, https://doi.org/10.1029/2018WR024146.

814 Cubasch, U., and Coauthors, 2001: *Projections of Future Climate Change*.

815 Daly, S. F., R. Davis, E. Ochs, and T. Pangburn, 2000: An approach to spatially distributed
816 snow modelling of the Sacramento and San Joaquin basins, California. *Hydrological*
817 *Processes*, **14**, 3257–3271, https://doi.org/10.1002/1099-
818 1085(20001230)14:18<3257::AID-HYP199>3.0.CO;2-Z.

819 Day, G. N., 1985: Extended Streamflow Forecasting Using NWSRFS. *Journal of Water*
820 *Resources Planning and Management*, **111**, 157–170,
821 https://doi.org/10.1061/(ASCE)0733-9496(1985)111:2(157).

822 Dettinger, M. D., and D. R. Cayan, 1995: Large-Scale Atmospheric Forcing of Recent Trends
823 toward Early Snowmelt Runoff in California. *Journal of Climate*, **8**, 606–623,
824 https://doi.org/10.1175/1520-0442(1995)008<0606:LSAFOR>2.0.CO;2.

825 Doesken, N., and A. Judson, 1996: The Snow Booklet: A Guide to the Science, Climatology,
826 and Measurement of Snow in the United States, Dep. of Atmos. *Sci., Colorado State*
827 *Univ., Fort Collins, CO*, 5.

828 Falcone, J. A., 2011: GAGES-II: geospatial attributes of gages for evaluating streamflow.
829 https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml (Accessed
830 April 15, 2021).

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

834 Fisher, R. A., and C. D. Koven, 2020: Perspectives on the Future of Land Surface Models
835 and the Challenges of Representing Complex Terrestrial Systems. *J. Adv. Model.*
836 *Earth Syst.*, **12**, https://doi.org/10.1029/2018MS001453.

837 Fleming, S. W., and A. G. Goodbody, 2019: A Machine Learning Metasystem for Robust
838 Probabilistic Nonlinear Regression-Based Forecasting of Seasonal Water Availability
839 in the US West. *IEEE Access*, **7**, 119943–119964,
840 https://doi.org/10.1109/ACCESS.2019.2936989.

841 ——, D. C. Garen, A. G. Goodbody, C. S. McCarthy, and L. C. Landers, 2021a: Assessing
842 the new Natural Resources Conservation Service water supply forecast model for the
843 American West: A challenging test of explainable, automated, ensemble artificial
844 intelligence. *Journal of Hydrology*, **602**, 126782,
845 https://doi.org/10.1016/j.jhydrol.2021.126782.

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

850 Garen, D. C., 1992: Improved Techniques in Regression-Based Streamflow Volume
851 Forecasting. *Journal of Water Resources Planning and Management*, **118**, 654–670,
852 https://doi.org/10.1061/(ASCE)0733-9496(1992)118:6(654).

853 Guo, J., J. Zhou, H. Qin, Q. Zou, and Q. Li, 2011: Monthly streamflow forecasting based on
854 improved support vector machine model. *Expert Systems with Applications*, **38**,
855 13073–13081, https://doi.org/10.1016/j.eswa.2011.04.114.

856 Hamlet, A. F., and D. P. Lettenmaier, 1999: Columbia River Streamflow Forecasting Based
857 on ENSO and PDO Climate Signals. *Journal of Water Resources Planning and*
858 *Management*, **125**, 333–341, https://doi.org/10.1061/(ASCE)0733-
859 9496(1999)125:6(333).

860 ——, P. W. Mote, M. P. Clark, and D. P. Lettenmaier, 2005: Effects of Temperature and
861 Precipitation Variability on Snowpack Trends in the Western United States. *Journal*
862 *of Climate*, **18**, 4545–4561, https://doi.org/10.1175/JCLI3538.1.

863 Hay, L. E., G. J. McCabe, M. P. Clark, and J. C. Risley, 2009: Reducing Streamflow Forecast
864 Uncertainty: Application and Qualitative Assessment of the Upper Klamath River
865 Basin, Oregon. *JAWRA Journal of the American Water Resources Association*, **45**,
866 580–596, https://doi.org/10.1111/j.1752-1688.2009.00307.x.

867 He, M., M. Russo, and M. Anderson, 2016: Predictability of Seasonal Streamflow in a
868 Changing Climate in the Sierra Nevada. *Climate*, **4**, 57,
869 https://doi.org/10.3390/cli4040057.

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

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

875 Kişi, Ö., 2007: Streamflow Forecasting Using Different Artificial Neural Network
876 Algorithms. *Journal of Hydrologic Engineering*, **12**, 532–539,
877 https://doi.org/10.1061/(ASCE)1084-0699(2007)12:5(532).

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

881 Kratzert, F., D. Klotz, M. Herrnegger, A. K. Sampson, S. Hochreiter, and G. S. Nearing,
882 2019: Toward Improved Predictions in Ungauged Basins: Exploiting the Power of
883 Machine Learning. *Water Resour. Res.*, **55**, 11344–11354,
884 https://doi.org/10.1029/2019WR026065.

885 Lehner, F., A. W. Wood, D. Llewellyn, D. B. Blatchford, A. G. Goodbody, and F.
886 Pappenberger, 2017: Mitigating the Impacts of Climate Nonstationarity on Seasonal
887 Streamflow Predictability in the U.S. Southwest. *Geophys. Res. Lett.*, **44**,
888 https://doi.org/10.1002/2017GL076043.

889 Li, D., M. L. Wrzesien, M. Durand, J. Adam, and D. P. Lettenmaier, 2017: How much runoff
890 originates as snow in the western United States, and how will that change in the
891 future?: Western U.S. Snowmelt-Derived Runoff. *Geophys. Res. Lett.*, **44**, 6163–
892 6172, https://doi.org/10.1002/2017GL073551.

893 Livneh, B., and A. M. Badger, 2020: Drought less predictable under declining future
894 snowpack. *Nat. Clim. Chang.*, **10**, 452–458, https://doi.org/10.1038/s41558-020-
895 0754-8.

896 Llewellyn, D., A. Wood, and F. Lehner, 2018: Runoff Efficiency and Seasonal Streamflow
897 Predictability in the U.S. Southwest. *Bureau of Reclamation*, **ST-2015-8730-01**, 63.

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

901 MacDonald, G. M., and Coauthors, 2008: Climate Warming and 21st-Century Drought in
902 Southwestern North America. *Eos Trans. AGU*, **89**, 82–82,
903 https://doi.org/10.1029/2008EO090003.

904 McGovern, A., R. Lagerquist, D. John Gagne, G. E. Jergensen, K. L. Elmore, C. R. Homeyer,
905 and T. Smith, 2019: Making the Black Box More Transparent: Understanding the
906 Physical Implications of Machine Learning. *Bulletin of the American Meteorological*
907 *Society*, **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 Resources Research*, **57**, 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. *Journal of Hydrometeorology*, **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 Clim Atmos Sci*, **1**, 1–6, 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. *Nature 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. Clim. Chang.*, **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 Res*, **57**, https://doi.org/10.1029/2020WR028091.

Nowak, K., M. Hoerling, B. Rajagopalan, and E. Zagona, 2012: Colorado River Basin Hydroclimatic Variability. *Journal of Climate*, **25**, 4389–4403, https://doi.org/10.1175/JCLI-D-11-00406.1.

NRCS, 2010: *A Measure of Snow: Case Studies of the Snow Survey and Water Supply Forecasting Program*.

Pagano, T., and D. Garen, 2005: A Recent Increase in Western U.S. Streamflow Variability and Persistence. *Journal of Hydrometeorology*, **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. *Journal of Hydrometeorology*, **5**, 896–909, https://doi.org/10.1175/1525-7541(2004)005<0896:EOOWUS>2.0.CO;2.

——, D. Garen, T. R. Perkins, and P. A. Pasteris, 2009: Daily Updating of Operational Statistical Seasonal Water Supply Forecasts for the western U.S.1. *JAWRA Journal of the American Water Resources Association*, **45**, 767–778, https://doi.org/10.1111/j.1752-1688.2009.00321.x.

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

948 Palmer, P., 1988: The SCS snow survey water supply forecasting program: Current
949 operations and future directions, In Proc. *Western Snow Conf.*, 43–51.

950 Reichstein, M., G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais, and Prabhat,
951 2019: Deep learning and process understanding for data-driven Earth system science.
952 *Nature*, **566**, 195–204, https://doi.org/10.1038/s41586-019-0912-1.

953 Robertson, D. E., and Q. J. Wang, 2012: A Bayesian Approach to Predictor Selection for
954 Seasonal Streamflow Forecasting. *Journal of Hydrometeorology*, **13**, 155–171,
955 https://doi.org/10.1175/JHM-D-10-05009.1.

956 Robertson, D. E., P. Pokhrel, and Q. J. Wang, 2013: Improving statistical forecasts of
957 seasonal streamflows using hydrological model output. *Hydrol. Earth Syst. Sci.*, **17**,
958 579–593, https://doi.org/10.5194/hess-17-579-2013.

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

963 Sharma, P., and D. Machiwal, 2021: Streamflow forecasting. *Advances in Streamflow*
964 *Forecasting*, Elsevier, 1–50.

965 Shukla, S., and D. P. Lettenmaier, 2011: Seasonal hydrologic prediction in the United States:
966 understanding the role of initial hydrologic conditions and seasonal climate forecast
967 skill. *Hydrol. Earth Syst. Sci.*, **15**, 3529–3538, https://doi.org/10.5194/hess-15-3529-
968 2011.

969 Slack, J. R., and J. M. Landwehr, 1992a: *Hydro-climatic data network (HCDN); a US*
970 *Geological Survey streamflow data set for the United States for the study of climate*
971 *variations, 1874-1988*. US Geological Survey,.

972 Slack, J. R., and J. M. Landwehr, 1992b: Hydro-climatic data network: a US Geological
973 Survey streamflow data set for the United States for the study of climate variations,
974 1874–1988. USGS Open-File Report 92-129. *US Geological Survey*,.

975 Slater, L. J., and G. Villarini, 2018: Enhancing the Predictability of Seasonal Streamflow
976 With a Statistical-Dynamical Approach. *Geophys. Res. Lett.*, **45**, 6504–6513,
977 https://doi.org/10.1029/2018GL077945.

978 Steinemann, A., S. F. Iacobellis, and D. R. Cayan, 2015: Developing and Evaluating Drought
979 Indicators for Decision-Making. *Journal of Hydrometeorology*, **16**, 1793–1803,
980 https://doi.org/10.1175/JHM-D-14-0234.1.

981 Stewart, I. T., D. R. Cayan, and M. D. Dettinger, 2004: Changes in Snowmelt Runoff Timing
982 in Western North America under a `Business as Usual' Climate Change Scenario.
983 *Climatic Change*, **62**, 217–232,
984 https://doi.org/10.1023/B:CLIM.0000013702.22656.e8.

985 Sturtevant, J. T., and A. A. Harpold, 2019: Forecasting the effects of snow drought on
986 streamflow volumes in the Western U.S. Western Snow Conference, 4.

987 Svoboda, M., and Coauthors, 2002: THE DROUGHT MONITOR. *Bull. Amer. Meteor. Soc.*,
988 **83**, 1181–1190, https://doi.org/10.1175/1520-0477-83.8.1181.

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

991 Werner, K., D. Brandon, M. Clark, and S. Gangopadhyay, 2004: Climate Index Weighting
992 Schemes for NWS ESP-Based Seasonal Volume Forecasts. *Journal of*
993 *Hydrometeorology*, **5**, 1076–1090, https://doi.org/10.1175/JHM-381.1.

994 Wiken, E. D., F. J. Nava, and G. Griffith, 2011: North American terrestrial ecoregions—level
995 III. *Commission for Environmental Cooperation, Montreal, Canada*, **149**.

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

999 Wood, A. W., and J. C. Schaake, 2008: Correcting Errors in Streamflow Forecast Ensemble
1000 Mean and Spread. *Journal of Hydrometeorology*, **9**, 132–148,
1001 https://doi.org/10.1175/2007JHM862.1.

1002 ——, T. Hopson, A. Newman, L. Brekke, J. Arnold, and M. Clark, 2016: Quantifying
1003 Streamflow Forecast Skill Elasticity to Initial Condition and Climate Prediction Skill.
1004 *Journal of Hydrometeorology*, **17**, 651–668, https://doi.org/10.1175/JHM-D-14-
1005 0213.1.

1006 Woodhouse, C. A., G. T. Pederson, K. Morino, S. A. McAfee, and G. J. McCabe, 2016:
1007 Increasing influence of air temperature on upper Colorado River streamflow:
1008 Temperature and Colorado Streamflow. *Geophys. Res. Lett.*, **43**, 2174–2181,
1009 https://doi.org/10.1002/2015GL067613.

1010