Forecasting Seasonal Water Supply in the Western US Basins Using Partial Least Square Regression

Atabek Umirbekov[1,2], Changxing Dong[1]
("iamo-team')

[1] Leibniz Institute of Agricultural Development in Transition Economies (IAMO), Theodor-Lieser-Str. 2, 06120 Halle (Saale), Germany
[2] Geography Department, Humboldt-Universität zu Berlin, Unter den Linden 6, 10099 Berlin, Germany

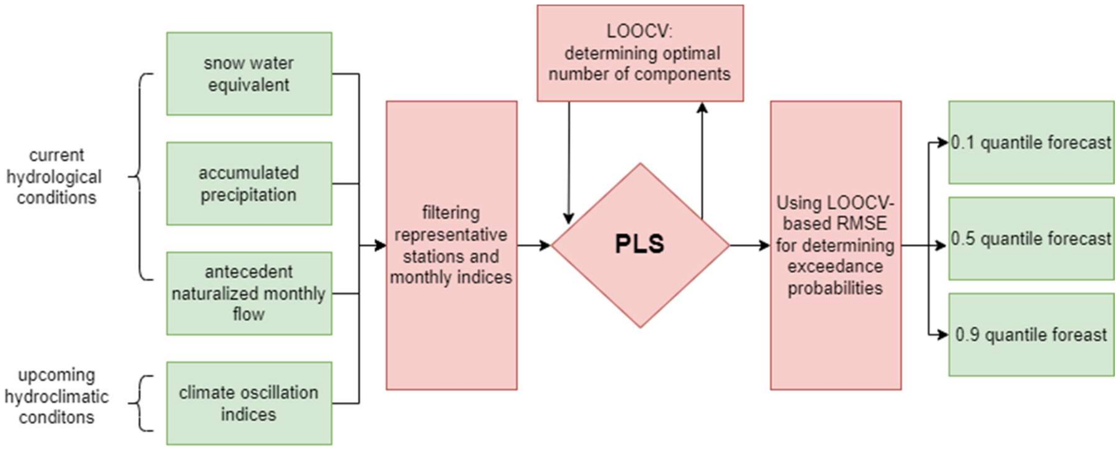
Abstract

Herewith we describe our solution for seasonal hydrological forecasting in the Western US which involves pre-selection of representative predictors, using Partial Least Squares regression (PLS) as a forecast model, extensive cross-validation to determine the optimal number of PLS components, and determining exceedance probabilities using cross-validated Root Mean Squared Error (RMSE) of the final PLS model. The approach relies on estimates of snow water equivalent (SWE) and accumulated precipitation from SNOTEL network, and antecedent monthly flows as input variables. These main predictors were augmented by winter state of Southern Oscillation Index, Pacific Decadal Oscillation, and Pacific-North American pattern, that complement forecasts at early issue dates. The solution involves training a PLS model for each issue date per each river, using a predefined range of PLS components and evaluated using Leave-One-Out Cross-Validation (LOOCV). PLS is chosen for its ability to handle collinearity among predictor variables, resilience to noise and applicability to small data, making it robust in situations where traditional methods may struggle. Compared to other tested data-driven techniques, PLS exhibited better accuracy and consistency in terms of uncertainty propagation across the forecast issue dates. 0.1 and 0.9 quantile forecasts were estimated by using PLS predictions adjusted by RMSE resulted from LOOCV multiplied by coefficient, assuming normal distribution of errors. The study's limitations include its non-accountability for non-stationarity of the system, especially effect of temperature along the training period.

Technical Approach

Algorithm and Architecture Selection

Our solution is based on i) pre-selection of representative predictors, that reflect current and upcoming hydroclimatic conditions, ii) use of Partial Least Squares regression (PLS) as a forecast model, iii) extensive cross validation to determine optimal number of PLS components, iv) determination of exceedance probabilities using cross-validated Root Mean Squared Error (RMSE) of the final PLS model. Figure 1 depicts the overall workflow of our modelling framework and its primary components, which are described in greater detail in the following sections.



current hydrological conditions
upcoming hydroclimatic conditons
snow water equivalent
accumulated precipitation
antecedent naturalized monthly flow
climate oscillation indices
filtering representative stations and monthly indices
LOOCV: determining optimal number of components
PLS
Using LOOCV-based RMSE for determining exceedance probabilities
0.1 quantile forecast
0.5 quantile forecast
0.9 quantile foreast

Figure 1. Workflow of our approach. Elements filled with green represent input or output data, elements filled with red correspond to data pre-(post-)processing and modelling

Our solution generally follows a common approach to seasonal hydrological forecasting which typically builds on two major elements: i) initial hydrological conditions, and, ii) future climate variability (WMO, 2021). For initial hydrological conditions our approach relies on estimates of snow water equivalent (SWE), accumulated precipitation, and antecedent monthly flows. As predictors of upcoming seasonal climate we use indices of some large-scale climate oscillations that are known to condition hydroclimatic variability on seasonal scales. The SWE and monthly naturalized flow exhibit strong temporal relationship with seasonal volume that is gradually increasing. As the correlation reaches its maximum by spring and then gradually decays, we therefore retain their values by specific issue-dates as complementary predictors for all later issue-dates until the end of forecasting season (explained more in detail in the following section).

These predictors served as input variables for the forecast model, for which our approach utilizes Partial Least Squares (PLS), a statistical method similar to Principal Component Analysis (PCA), which is employed by some US agencies for seasonal water supply forecasts. PLS, like PCA, is a multivariate analysis technique aiming at dimensionality reduction and feature extraction. However, their objectives differ, as PCA focuses on maximizing data variance, while PLS seeks to maximize covariance between independent and dependent variables. This makes PLS particularly advantageous in predictive modelling scenarios, such as regression and classification tasks. Notably, our choice of PLS is driven by its ability to handle collinearity among predictor variables, rendering it robust in situations where traditional methods may struggle. Despite its strengths, PLS is not without limitations, including sensitivity to outliers and a potential risk of overfitting in small sample sizes.

PLS involves the construction of latent variables in the predictor space, known as the predictor components. These components are linear combinations of the original predictor variables and are derived iteratively to maximize the covariance between the predictors and the response variables. The goal is to capture the essential information in the predictor variables that is relevant to explaining the variability in the response variables. The number of components is an important parameter in PLS modelling and cross-validation is often employed to determine the optimal number of components. We systematically evaluated model performance on different subsets of the data, using leave one out cross-validation (LOOCV) to select the most appropriate number of components, which also ensured a balance between model complexity and predictive accuracy. This procedure is repeated for each issue date per river basin.

The PLS model with optimal number of components is used to generate final prediction, which corresponds to median 0.5 quantile of water supply forecast. We then use root mean squared error (RMSE) to determine 0.1 and 0.9 exceedance probabilities of the prediction by each issue date.

We have tested several other data-driven techniques to compare their performance with that of PLS. These include Generalized Linear Regression with stepwise feature selection, Random Forest, Support Vector Regression and Gaussian Process regression with both linear and radial-basis kernels, as well as ensemble meta-learner model architectures that stacked predictions of the mentioned machine-learning techniques (not shown here). In all cases PLS exhibited better accuracy in terms of LOOCV-ed RMSE and R-squared coefficient. In some cases, depending on a basin or issue date, some of these models could bear slightly better predictions, but when generalizing across all basins and issue dates PLS still shows superior performance. PLS also showed a better consistency in terms of uncertainty, with prediction errors being expectedly high at early issue dates and gradually decreasing towards the end of the forecasting season. In contrast, some of the tested machine learning models exhibited relatively chaotic changes of the NRMSE across issue dates from January to April. Interestingly, for some forecasts for issue dates in July, a simple linear regression using only naturalized flows over preceding three months (June, May, April) showed slightly better performance than PLS. Despite those episodic outperformances of different machine learning models tested, we decided to stick to using PLS.

Another observation is that linear models generally showed better performance than tree-based models or models with non-linear kernels. We assume that this may be reasoned by the following non-exclusive factors: 1) linearity of the underlying system, 2.) selection of predictors based on rather linear metric (Pearson's correlation), 3.) relatively smaller number of both predictors and observations that make some of the mentioned non-linear machine learning models less efficient.

Data Sources and Feature Engineering

Table 1 below lists the predictors used in our forecasting approach. The seasonal terrestrial storage accumulated during winter time in the western basins of the US is considered the main source of predictability for river discharge during the summer season. Therefore, our primary predictors comprise of Snow Water Equivalent (SWE) and accumulated precipitation from the Snow Telemetry (SNOTEL) network. Another significant predictor we use is the antecedent (naturalized) river flow. In addition to these main predictors, we also retain SWE and monthly flow values resulting from forecast issue dates when they exhibit the maximum correlation with seasonal volume for all forecasts at later dates. We refer to these as 'retained predictors. All values were scaled and normalized by standard deviation, prior to use in the model.

Table 1. Predictors used in the PLS model

Predictor	Source	Description
Main predictors		
snow water equivalent (in-situ)	SNOTEL	estimates from up to 4 SNOTEL stations per each basin
snow water equivalent (reanalysis)	SWANN	Basin averaged SWE derived from gridded SWANN data
accumulated precipitation	SNOTEL	estimates from up to 4 SNOTEL stations per each basin
naturalized monthly flow	USGS	monthly river flow over the last three preceding months
SOI index	NCEI	3-month averaged index
PNA index	CPC	3-month averaged index
PDO index	NCEI	3-month averaged index
Supplementary predictors		
SWE at issue date of maximum correlation to seasonal volume	SNOTEL	derived by analyzing timeseries of SWE from selected SNOTEL stations and seasonal volume
naturalized flow at month of maximum correlation to seasonal volume	USGS	derived by analyzing timeseries of monthly naturalized flow and seasonal volume

To utilize data from the most representative SNOTEL stations, we selected up to four stations for each river basin using certain criteria. This selection was based on Pearson's correlation analysis between the SWE observations at all forecast issue dates and seasonal river volume from 1985 to 2022. For two to three stations, we determined the selection by identifying those with the highest median correlation coefficient across all forecast issue dates. It's noteworthy that, for many stations, the maximum correlation between SWE and seasonal volume occurred in mid-April and was sustained until the end of June, as depicted in Figure 2.

To further enhance the representativeness, one to two additional SNOTEL stations were selected using a similar approach but focusing on correlations for issue dates until May. This ensured a balanced diversity of SNOTEL stations for each basin, incorporating also stations where SWE exhibited higher correlation at early lead times. The diversity in correlation patterns was often associated with the elevation differences among stations, with higher elevated SNOTEL stations demonstrating stronger predictive power during earlier issue dates.

The relationship between snowpack and seasonal flow volume exhibits a robust temporal pattern that aligns with its accumulation and melt phases. Typically, the correlation is lower at the earliest issue dates, gradually increasing over time, reaching its peak by late spring, and then declining by summer. To maintain consistency in predictive performance of the forecast model, we further identified one station (out of the selected four) and one specific issue date that demonstrated the highest correlation with seasonal flow volume. The SWE value for this station at that particular issue date is retained as an additional predictor variable for all later issue dates.

During certain years characterized by dry autumn and presumably higher winter temperatures, some SNOTEL stations, particularly those located in the southwestern part of the domain (e.g., California), registered zero SWE values. While this may suggest a higher likelihood of lower-than-normal volume during the upcoming season, these zero values led to underestimation by the PLS model. Another contributing factor is the total meltout of the snowpack during late issue dates (e.g., June and July), further impacting the forecast model's performance. To address these challenges, we augmented the SWE variable with accumulated precipitation observations obtained from SNOTEL stations. Instead of utilizing precipitation records from the same SNOTEL stations selected for SWE, we conducted a similar analysis to identify stations where accumulated precipitation (from preceding October) exhibited a higher correlation with seasonal flow volume.

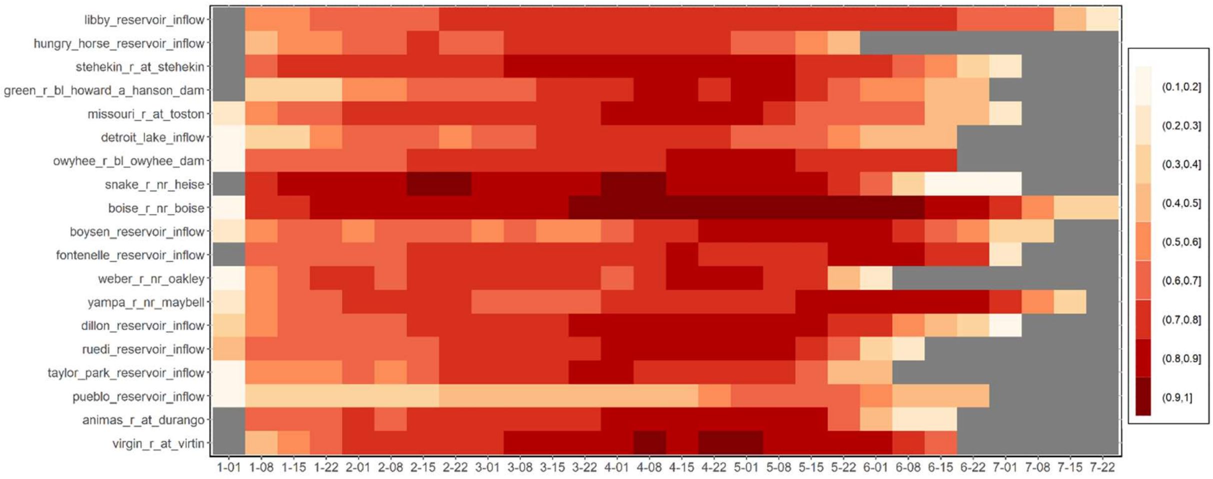


Figure 2.Correlations between selected station's SWE at different forecast issue dates and seasonal volume. Grey areas indicate either no or negative correlations.

Antecedent monthly flow is another important predictor in our approach, displaying high temporal autocorrelation with seasonal flow volume. The relationship between month flow in earlier months is typically low, but late spring months show the highest correlation (Figure 3). Instead of using only the antecedent flow for the previous month, we introduce monthly flows for the preceding three months as separate predictors. For example, by the June 1st issue date, predictors include 'flow_5,' 'flow_4,' and 'flow_3' variables corresponding to flow values in May, April, and March, respectively. Recognizing that monthly flow in some river basins establishes a strong relationship with seasonal flow volume at earlier months, we retain flow values at these months as additional predictors in a similar manner we did it for SWE. For the three Californian basins with missing monthly naturalized flow data, we utilized monthly naturalized streamflow from NRCS data. We selected nearby catchments where aggregated monthly streamflow exhibited the highest correlation with seasonal volume in the target basins.

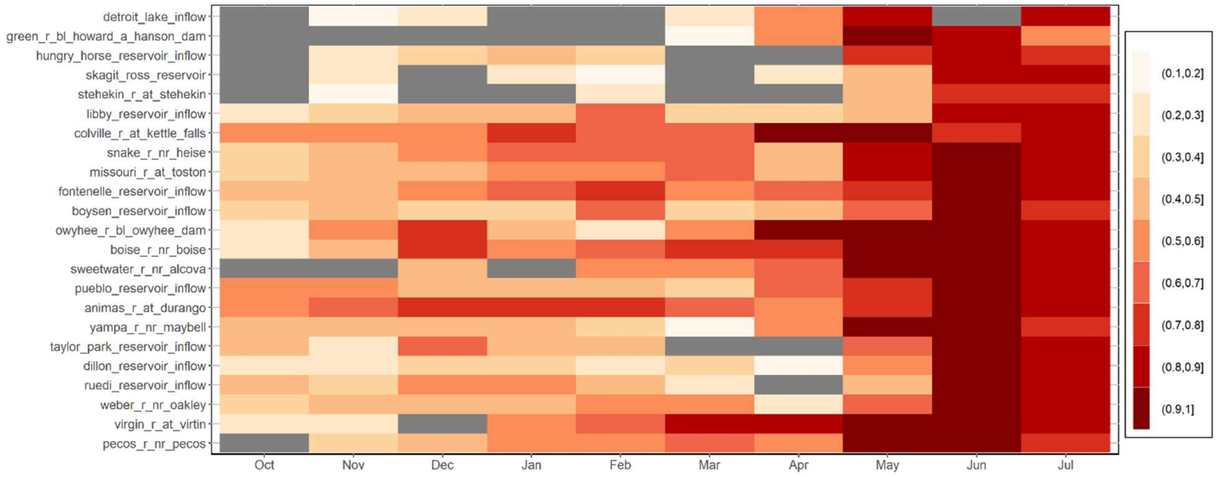


Figure 3. Correlations between river monthly naturalized flow at different months with seasonal volume. Grey areas indicate either no or below 0.1 correlations.

SWE, accumulated precipitation and antecedent monthly flow in our approach are to capture current hydrological conditions. These variables are however of lower predictive power at the earliest issue dates, since they do not sufficiently capture hydrologic trends of the upcoming season. Therefore, we augmented predictors with climate oscillations that serve as precursors of hydroclimatic conditions during targeted season.

One of the major climate oscillations is El Nino-Southern Oscillation (ENSO) which is known to have noticeable impact on hydroclimate variability across the globe, including US (Patricola et al., 2020; Rice & Emanuel, 2017). There were several approved indices that reflect ENSO phenomena, all showing similar teleconnection patterns with seasonal flow volume in targeted basins. We decided to use Southern Oscillation Index (SOI), that uses difference in air pressure, because it emerges relatively earlier sign of the ongoing ENSO phase compared to those ENSO indices based on sea temperature. ENSO impacts during certain phases can be enhanced by the Pacific Decadal Oscillation (PDO) (Tootle et al., 2005). Another climate oscillation we decided to use as predictor is Pacific-North American pattern (PNA), which is known to impact regional temperature and precipitation variability affecting snowpack formation in the western US (Abatzoglou, 2011; Leathers et al., 1991).

To mainstream use of these climate indices as predictors it's necessary to determine their temporal lead-lag times at which they exhibit strongest relationship with interannual seasonal flow variability. As the climate indices show noisy patterns on monthly scale, we first aggregate them to rolling three-month averages. In the next step we assessed correlations between indices at different lead month and seasonal flow volume in each target river basin. Figure 4a show for reference resulted correlations of SOI with seasonal flow volume across the river basins. The hydrologic response of the rivers flow to ENSO signal, as well for PDO and PNA, exhibit robust regional differences: Rivers located in the southwestern region show a divergent response from those in the west-northern part of the domain.

In the final step, we estimated mean absolute correlation coefficient across all river basins (Figure 4b, c, and d) to determine the lead months at which the climate indices have the strongest and widespread effect across the study domain. Our findings suggest that generally winter state of SOI, PDO and PNA have a larger effect of interannual hydrological variability. Therefore, corresponding values of the climate indices (e.g. 3-month SOI index cantered on December) enter as additional predictors for the PLS and remain constant with subsequent issue dates. Per each river basin we use only those climate indices that surpass 0.1 correlation coefficient calculated for the historical period. In addition, climate indices are used as predictors for issue dates until June, after which we assume they effect is negligible and seasonal volume is predominantly driven by accumulated water storage reflected in our solution by antecedent monthly flows and precipitation.

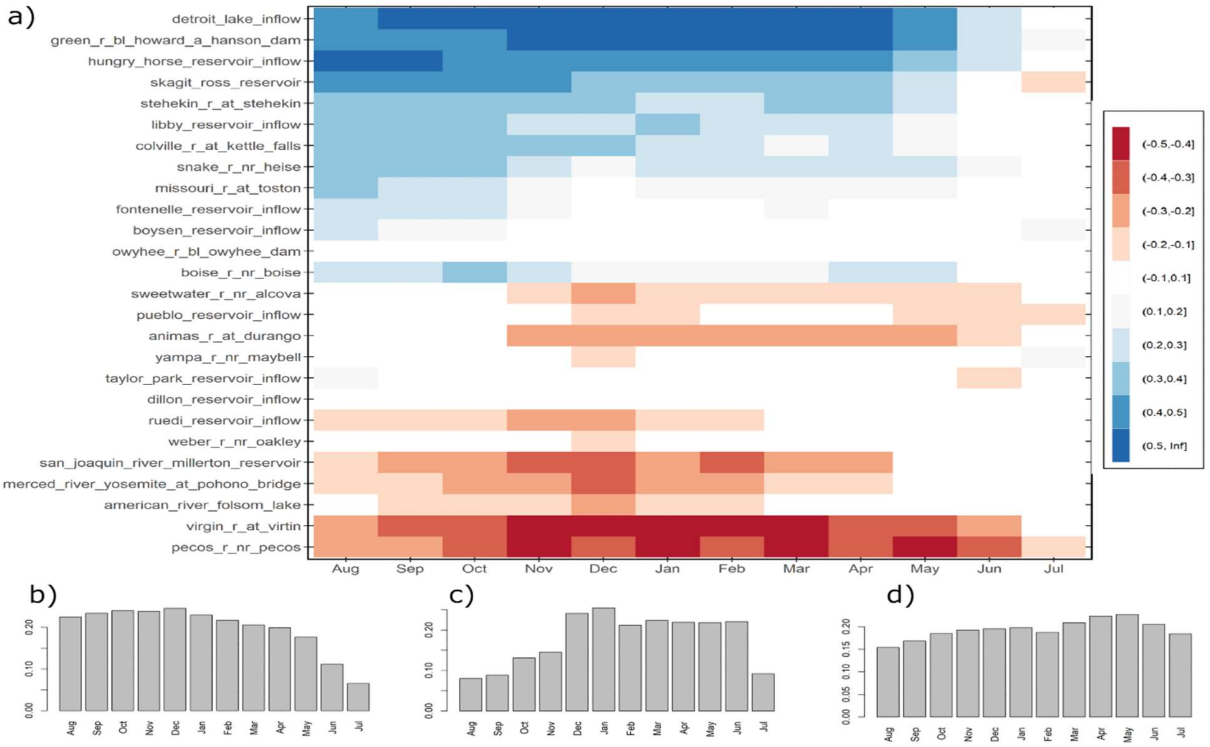


Figure 4. Climate teleconnections of the target river basins. a) correlation between 3-month averaged SOI index with seasonal river flow at different lead months, b), c.) and d.) are mean absolute correlation coefficient across all basins' seasonal flow and SOI, PNA and PDO indices at different index lead month

All noted data for models training was chosen for period starting from 1985 to 2022, which resulted approximately to 30 observations per each basin and issue date. We found this period optimal in terms of the following aspects: not all sites had required historical observations at earlier years; we wanted to minimize effects of land cover/use changes (which our model framework doesn't consider), which may have altered flow formation; and because of non-stationarity of hydrological systems that emerge over longer periods.

One of the major limitations of our approach is its non-accountability for non-stationarity of the system, which especially refers to gradual growth of temperature. During the past decade alone, many parts of the western US witnessed record seasonal temperatures, which likely caused higher evaporation rates and earlier peaking of discharge and eventually reducing seasonal flow volume. Some evidence suggests that this issue could be overcome by incorporating temperature forecasts into water supply predictions (Lehner et al., 2017). However, we haven't incorporated seasonal temperature forecast as additional predictor, because we found that most hindcast data (ECMWF, CPC) is only available starting from 1990s which would have reduced our training sample to less than 20 years. We haven't explored other approved datasets either due to their limited length of observations or complexities with their retrieval and preprocessing. As an example, we consider the possibility that SWE estimates (basin-averaged) from the Snow Data Assimilation System (SNODAS) could potentially serve as good alternatives to or complement SWE estimates from the SNOTEL network. However, SNODAS data is only available from the beginning of the 2000s, and with a ten-year test sample we had to predict, this would leave us with only 12 years of record, which we regard as insufficient for training. Complexities with data retrieval were primarily linked to the fact that our solution has been developed in R, which lacked dedicated API packages for the seamless download of some other approved data.

Uncertainty Quantification

The final PLS model generates forecasts for each issue date corresponding to the 0.5 quantile. To determine the 0.1 and 0.9 forecast quantiles, we used the 0.5 quantile prediction, adjusted by the RMSE resulting from LOOCV multiplied by a 1.28 coefficient. While this simplistic approach assumes a normal distribution of errors, which may not always be the case, it in overall accurately reflected the exceedance probabilities.

In a few instances, the 0.5 and 0.1 quantile forecasts exhibited values below zero. These occurrences were primarily observed during early forecast issue dates, where lower 0.5 forecast values and higher RMSE led to negative 0.1 forecast values. To address this, we constrained the negative 0.1 quantile forecasts by setting their values to 0 and adjusted those few negative 0.5 quantile forecasts to match the historical minimum values of seasonal flow volume.

Training and Evaluation Process

The overall dataset used for training covered period from 1985 to 2023

We trained the models by testing a predefined range of PLS component numbers, ranging from 2 to 4. We evaluated the model with each component using Leave-One-Out Cross-Validation (LOOCV). For assessing the performance of the model, we utilized the Root Mean Square Error (RMSE) resulting from LOOCV. The choice of LOOCV was primarily driven by the relatively short period of available data records, especially for SNOTEL, in each river basin. For certain rivers where data was available, we also evaluated models by assessing their ability to simulate seasonal streamflow volume before 1985. However, even in these cases, the number of available hold-out years was still small to derive robust conclusions. Therefore, our final solution selects the most optimal model parameter (number of Partial Least Squares components) using solely LOOCV on the training data.

Discussion of Performance

Figure 5 below shows performance of the models across rivers and forecast issue dates on the training data. Overall, the performance of our approach follows the typical dynamics observed in hydro-climatic forecasting, characterized by higher errors at the early forecast issue dates and a gradual improvement in accuracy as the lead times decrease. This pattern is consistent with the inherent challenges of predicting complex hydrological systems, where uncertainties are more pronounced in long-range forecasts and diminish as the forecasting horizon shortens

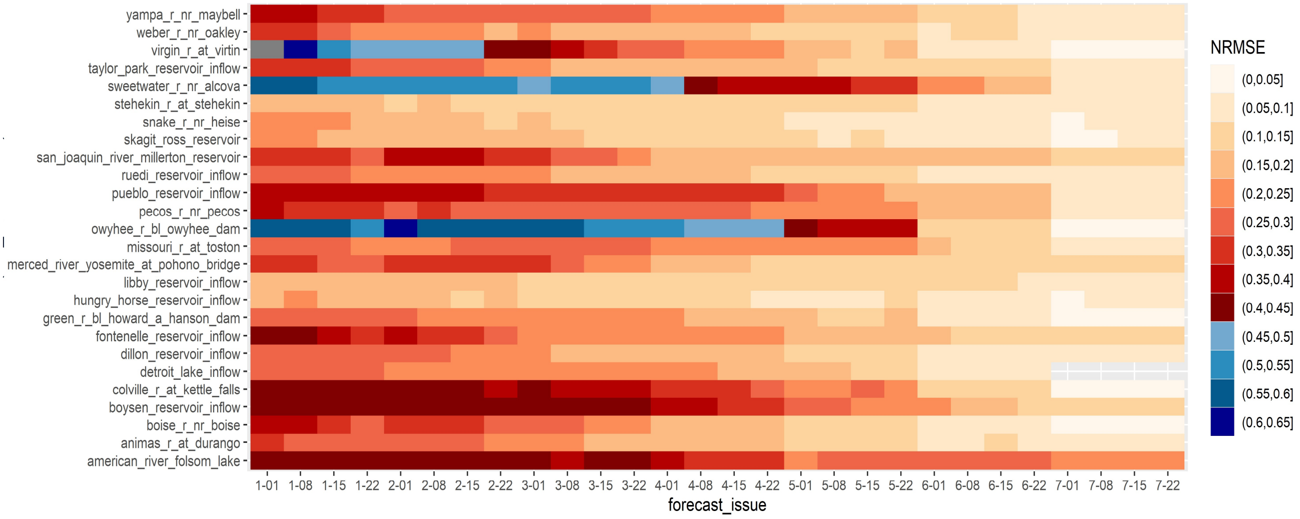


Figure 5. Normalized RMSE of the forecast models per forecast issue date and river basin. RMSE is normalized with respect to mean seasonal volume.

Across the issue dates, predictions for some basins showed relatively higher errors compared to the average errors across other river basins. A more detailed analysis revealed that errors were higher in rivers that exhibit higher inter-annual variability of the seasonal streamflow. To some extent, the higher variability of streamflow is associated with the drainage area of the studied basins. We haven't found any visible association of prediction performance with the elevation or latitudinal location of the basin. Forecasts show relatively lower errors for the basins located in the Cascades region.

Changes Between Stages

Table 2 below outlines the primary modifications made to solutions submitted for the Forecast and Final stages compared to those used in the Hindcast stage. In overall, the model architecture maintains uniformity across the Hindcast, Forecast, and Final stages, while incorporating minor adaptations in predictor selection and training methodologies specific to each stage. Preliminary analysis of forecast accuracies indicated higher deviations for years characterized by abnormal discharge, defined as falling within the 0.33 and 0.67 percentiles of historical observations. As a result, we opted to include only years from 1985 to 2003 falling into these two categories (above and below normal).

Table 2. Changes made during Forecast and Final stages, with respect to the approach submitted during Hindcast Evaluation stage

	Forecast stage	Final stage
changes in set of predictors	+ monthly averaged temperature from nearest ACIS station	+ basin averaged SWE estimates from the the SWANN dataset;
change in selection of training years		For the training period spanning 1984 to 2023, only years with discharge values above or below historical norms were considered

Machine Specifications

Our solution was implemented in the R programming language (version 4.2.1). Data preprocessing was performed using R packages, including '*dplyr*' and '*lubridate*,' while training and evaluation were implemented with the assistance of '*caret*' and '*hydroGOF*' packages. The solution was parallelized via help of the '*foreach*' package. The solution was compiled and run locally using laptop having the following configuration: Intel i7 CPU (8 cores), 16GB RAM.

References

Abatzoglou, J. T. (2011). Influence of the PNA on declining mountain snowpack in the Western United States. *International Journal of Climatology*, *31*(8), 1135–1142. https://doi.org/https://doi.org/10.1002/joc.2137

Leathers, D. J., Yarnal, B., & Palecki, M. A. (1991). The Pacific/North American Teleconnection Pattern and United States Climate. Part I: Regional Temperature and Precipitation Associations. *Journal of Climate*, *4*(5), 517–528. https://doi.org/10.1175/1520-0442(1991)004<0517:TPATPA>2.0.CO;2

Lehner, F., Wood, A. W., Llewellyn, D., Blatchford, D. B., Goodbody, A. G., & Pappenberger, F. (2017). Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability in the U.S. Southwest. *Geophysical Research Letters*, *44*(24), 12,208-212,217. https://doi.org/https://doi.org/10.1002/2017GL076043

Patricola, C. M., O'Brien, J. P., Risser, M. D., Rhoades, A. M., O'Brien, T. A., Ullrich, P. A., Stone, D. A., & Collins, W. D. (2020). Maximizing ENSO as a source of western US hydroclimate predictability. *Climate Dynamics*, *54*(1), 351–372. https://doi.org/10.1007/s00382-019-05004-8

Rice, J. S., & Emanuel, R. E. (2017). How are streamflow responses to the El Nino Southern Oscillation affected by watershed characteristics? *Water Resources Research*, *53*(5), 4393–4406. https://doi.org/https://doi.org/10.1002/2016WR020097

Tootle, G. A., Piechota, T. C., & Singh, A. (2005). Coupled oceanic-atmospheric variability and U.S. streamflow. *Water Resources Research*, *41*(12). https://doi.org/https://doi.org/10.1029/2005WR004381

WMO. (2021). *Guidelines on Seasonal Hydrological Prediction* (2021st ed., Issue 1274). World Meteorological Organization. https://library.wmo.int/idurl/4/57915