JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

AMERICAN WATER RESOURCES ASSOCIATION

Simplified Cross-Validation in Principal Component Regression (PCR) and PCR-Like Machine Learning for Water Supply Forecasting

Sean W. Fleming, and David C. Garen

Research Impact Statement: Seasonal river forecast models are used to manage water in the western US. Our study found alternative ways to measure their accuracy, facilitating design of next-generation prediction systems.

ABSTRACT: Cross-validated principal component regression (PCR) is widely used in day-to-day operational forecasting systems for seasonal river runoff volume in western North America. Complexities are increasing in both predictor datasets (including climate-science products) and in predictive models employed instead of linear regression within the PCR framework (including artificial intelligence), potentially complicating cross-validation for model evaluation. We explored these issues with 300 modeling experiments on two high-impact and hydroclimatically diverse basins in the western United States, the Truckee River (Sierra Nevada) and Rio Grande headwaters (southern Rockies), using five different PCR and PCR-like machine learning models. The results suggest out-of-sample error is satisfactorily estimated by applying cross-validation to only the final, supervised learning, step of PCR/PCR-like procedures. The outcome facilitates streamlined algorithms and potentially reduced computational times for more complex emerging model architectures and datasets; provides reassurance around a possible inability to perform genuinely complete cross-validation when predictors include certain complex and externally sourced data sources; and may reflect mitigation of overtraining by geophysical process-informed model development protocols normally used during feature selection in operational water supply forecast (WSF). The results provide practical guidance helping support the design of next-generation WSF models.

(KEYWORDS: water supply forecasting; water management; statistical modeling; machine learning.)

INTRODUCTION

Water supply forecasts (WSFs) in the western United States (U.S.) are predictions, issued beginning in winter, of upcoming spring-summer river runoff volume. Operational WSFs are crucial here for informing water management, including agriculture, hydropower, flood planning, ecosystem management, and municipal water management. Some of these activities are governed by legal decisions and international treaties, attracting close scrutiny of WSFs, and even modest WSF accuracy improvements can yield millions of dollars of benefit per year for a single basin (e.g., Hamlet et al. 2002). Furthermore, population growth is increasing water demand, and climate change may reduce manageable water supply, primarily through warmer winters giving lower mountain snowpack (see Bureau of Reclamation 2016). Considerations like these have motivated research to improve WSF skill, and gauging the effectiveness of these modeling developments requires reliable measurements of WSF accuracy.

Paper No. JAWR-21-0015-N of the *Journal of the American Water Resources Association* (JAWR). Received February 9, 2021; accepted April 30, 2022. **Discussions are open until six months from issue publication**.

National Water and Climate Center, Natural Resources Conservation Service, US Department of Agriculture, Portland, Oregon, USA (Correspondence to Fleming: sean.fleming@usda.gov).

Citation: Fleming, S.W., and D.C. Garen. 2022. "Simplified Cross-Validation in Principal Component Regression (PCR) and PCR-Like Machine Learning for Water Supply Forecasting." *Journal of the American Water Resources Association* 1–8. https://doi.org/10.1111/1752-1688.13007.

Leave-one-out cross-validated (LOOCV) principal component regression (PCR) is one of the most prevalent techniques in production systems used by government agencies and other organizations tasked with producing WSFs operationally (e.g., Fleming and Gupta 2020). It was adapted to WSF by the U.S. Department of Agriculture Natural Resources Conservation Service (NRCS) to facilitate linear regression modeling under predictor multicollinearity (Garen 1992; James et al. 2013). It has since been widely adopted for operational WSF in the U.S. and Canada, and as a WSF modeling tool in hydrology, snow, and climate research (e.g., Eldaw et al. 2003; Hsieh et al. 2003; Risley et al. 2005; Regonda, Rajagopalan, and Clark 2006; Regonda, Rajagopalan, Clark, and Zagon 2006; Kennedy et al. 2009; Perkins et al. 2009; Moradkhani and Meier 2010; Oubeidillah et al. 2011; Rosenberg et al. 2011; Gobena et al. 2013; Najafi and Moradkhani 2016; Harpold et al. 2017; Lehner et al. 2017; Fleming and Goodbody 2019; Glabau et al. 2020). PCR contains two separate procedures: principal component analysis (PCA) to project (typically, multicollinear) input data into a new coordinate system where the variables are mutually uncorrelated, that is, unsupervised learning for feature extraction; followed by classical stepwise linear regression modeling using these time series as potential predictors, that is, supervised learning for feature selection and predictive modeling. For several reasons, out-of-sample skill is estimated by LOOCV, which experience has shown to be reliable in WSF applications (for details see, e.g., Garen 1992; Pagano et al. 2004; Rosenberg et al. 2011; Lehner et al. 2017; Fleming and Goodbody 2019).

However, data-driven WSF methods are evolving with the appearance of more complex and diverse suites of predictors and more sophisticated predictive models. Some of this growth complicates the foregoing picture of what PCR is, how to deploy it for WSF, and how to perform cross-validation in a meaningful and efficient manner.

Consider five examples, which we return to below. Hsieh et al. (2003) used PCA to extract hemispheric-scale climate information from datasets consisting of time series of tropical Pacific sea surface temperature anomaly data at each location in a large-scale grid; a second, separate PCA, of gridded watershed-scale precipitation data, to obtain an index of initial soil moisture conditions; and several existing indices of large-scale climate variability from publicly accessible databases. These diverse analytical products were gathered to form a predictor pool for an artificial neural network (ANN)-based predictive WSF model. Carrier et al. (2013) developed a support vector machine-based WSF model, using as predictors a combination of instrumental climate indices and reconstructed (tree ring-derived) paleoclimate metrics, both extracted from existing climate-science community databases. Fleming and Goodbody (2019) introduced a WSF metasystem in which six statistical and machine learning prediction models each received an individually optimal feature set, derived from raw input precipitation and mountain snowpack data by PCA and a stochastic global search algorithm, along with algorithms to enforce a priori physicality constraints. Rosenberg et al. (2011) employed standard linear PCR for WSF, but used as predictors the high-dimensional output of a physics-based spatially distributed snow model. Our fifth example is Gobena and Gan (2010), who used long-range forecasts from a seasonal numerical climate model as predictive input to a robust M-regression-based WSF model.

A question that arises with these more involved emerging WSF frameworks is whether cross-validation is required across the entire PCR process (i.e., beginning with sample removal in the raw input dataset), as is common practice for classical linear PCR in conventional WSF applications, or if it is sufficient to perform cross-validation solely on the supervised learning (predictive modeling) portion of the process. The former is straightforward when a conventional PCR-in-WSF workflow is used, but it grows cumbersome, nuanced, inefficient, or even infeasible for more complex and sophisticated predictor sets developed through elaborate features engineering or derived in turn from other models.

The above examples illustrate some of these potential complications. The intricate, domain expertise-driven manual features engineering employed by Hsieh et al. (2003) could be clumsy to automate in a cross-validation process. Though Carrier et al. (2013) did not explicitly use PCR or PCA in their WSF modeling procedure, publicly available paleoclimate reconstructions and instrumental climate indices they used as inputs are developed in turn by climate scientists who often use PCA or PCR behind the scenes in their own data processing (e.g., Mantua et al. 1997; Cook et al. 1999). Another issue is the potential naivete of the cross-validation procedure. Consider the heavily processed gridded climate data used by Hsieh et al. (2003), paleoclimate reconstructions used by Carrier et al. (2013), or seasonal numerical climate model forecasts used by Gobena and Gan (2010). These are end products of extensive studies performed by third-party subject-matter experts and, being climate models and data, naturally contain complex internal dynamical structure. Leaving out one sample at a time of the resulting time series when using it as a potential input among others in WSF PCR model development and testing may only superficially be LOOCV. To truly exclude information from that sample time when building an out-of-sample model during WSF cross-

validation, one might instead have to recreate the input climate data products as if information at that sample time had never existed at any point in the analytical processes used to create them, for example, as if the tree-ring width corresponding to a particular year had never been measured, and then repeating for every sample time. Doing so is generally infeasible. Still another question is whether PCA on much higher-dimensional input datasets, such as the snow model in Rosenberg et al. (2011), may grow computationally intensive enough to appreciably slow the iterative cross-validation procedure. Similarly, the multiple semi-independent modeling system optimizations in Fleming and Goodbody (2019) are sufficiently time-consuming to require parallel computing, and reduction in CPU time could be advantageous. Still another potential consideration is the role of machine learning in place of conventional linear regression in PCR, as in Hsieh et al. (2003), Carrier et al. (2013), and Fleming and Goodbody (2019). This changes the fundamental goal of PCR-like modeling: multicollinearity is not usually considered problematic per se for supervised machine learning, and PCA is used instead primarily to reduce input data dimensionality, giving a more parsimonious, regularized, interpretable, and efficiently trained artificial intelligence. Additionally, given the greater potential vulnerability of machine learning to overtraining in the absence of adequate regularization procedures, one might hypothesize that in PCR-like machine learning applications, discrepancies between in-sample and cross-validated performance might be mainly attributable to overtraining in the artificial intelligence, not the PCA.

All things considered, there are several reasons to consider estimating out-of-sample WSF skill by only subjecting the regression or regression-like portion of the PCR or PCR-like modeling process to cross-validation. But how much validity may be lost from estimates of forecast accuracy? Addressing this question is useful for informing WSF best practices going forward.

DATA AND METHOD

A data matrix, $\mathbf{X} = x_{m,n}$ contains standardized observations of M predictive variables at N sample times (in operational WSF systems, these are typically annual time series of observational precipitation and snow data at various locations, see below). As per standard applications of PCR to WSF, unrotated PCA is performed on $\mathbf{X}$ to ultimately obtain a scores matrix, $\mathbf{A} = a_{m,n}$, containing the PC time series for each of M PCA modes at N times; these scores time series are by construction mutually uncorrelated and are used as candidate features in regression or machine-learning predictions of the vector of streamflow volumes at the corresponding N sample times, $\boldsymbol{q} = q_n$.

Out-of-sample estimates of this best-fit model's prediction error were then estimated by cross-validation. Two scenarios were considered. In (1) full LOOCV, a sample for a given time is dropped from $\mathbf{X}$ and $\boldsymbol{q}$, PCA is re-performed and the supervised learning model is re-fit, a prediction is made with that new model using input data from the sample time omitted during its construction, and the process is repeated for all other sample times to create a length-N LOOCV $\boldsymbol{q}$ estimate used as the basis for calculating fit metrics like root mean square error (RMSE) and coefficient of determination (R^2). In (2) supervised learning-only LOOCV, the process is identical except $\mathbf{A}$ is calculated only once, during initial model development using all the data; during cross-validation, one sample at a time is omitted from this set of PCA-derived features when forming the re-fitted suite of N cross-validation sub-models and length-N LOOCV $\boldsymbol{q}$ time series. These scenarios are summarized, for a given set of retained inputs and PCA modes, in Algorithms 1 and 2 below.

We applied both scenarios to two forecast points in the NRCS operational WSF system, (1) the Truckee River at Farad (Sierra Nevada snowmelt-fed outlet of Lake Tahoe) and (2) the Rio Grande near Del Norte (headwater location in the southern Rocky Mountains fed by snowmelt and minor spring-summer rainfall). Predictive inputs were SNOTEL observations of wintertime-to-date accumulated precipitation and forecast-date SWE; antecedent streamflow is additionally used for the Truckee. We considered early-season (January 1, for Rio Grande) and late-season (April 1, for Truckee) forecast issue dates. The target was U.S. Geological Survey (USGS) streamgage measurements of flow volume accumulated over the established primary management periods of April–September for Rio Grande and April–July for Truckee, with adjustments by NRCS to approximately naturalize flows, that is, correct for withdrawals and other local-scale water management processes. Data over a standard $N = 30$ hydroclimatic normal period (1986–2015) were employed.

This overall setup corresponds to existing operational PCR models at NRCS and elsewhere, helping ensure relevance to practical WSF applications. Similarly, our predictor variate choices and dataset sizes ($M = 25$ for Truckee, 10 for Rio Grande) closely reflect those in the current NRCS PCR models for these locations (e.g., Garen 1992; Perkins et al. 2009; Gobena et al. 2013; Fleming and Goodbody 2019). Tables S1 and S2 provide further details of these datasets, which are publicly available from NRCS (2021).

Algorithm 1: Scenario 1 (full) LOOCV procedure

Step 1 Assemble available $n = 1,N$ samples of target q_n and of $m = 1,M$ input variables at the same N times, $x_{m,n}$

Step 2 Perform PCA on input variables

- **Step 2a** Standardize each length (N) input time series to zero mean and unit variance on the basis of its sample mean and variance and combine into data matrix, $\mathbf{X}$
- **Step 2b** Calculate correlation matrix, $\mathbf{C} = N^{-1}\mathbf{X}\,\mathbf{X}^{T}$
- **Step 2c** Find eigenvectors $\mathbf{E}$ of $\mathbf{C}$, and scores matrix $\mathbf{A} = \mathbf{E}^{T}\mathbf{X}$ for N samples and M modes

Step 3 Train predictive model on $\mathbf{A}$ and $\boldsymbol{q}$

- **Step 3a** Select subset of PCA modes in $\mathbf{A}$ to use as predictive features, $\mathbf{A'} = a_{m,n}$, $m \in (1,2,\ldots,M)$, $n = 1,N$
- **Step 3b** Train a supervised learning algorithm, f, to predict expectation values of target, $<\boldsymbol{q}>$, from $\mathbf{A'}$

Step 4 Obtain out-of-sample predictive error estimate by leave-one-out cross-validation

- **Step 4a** For each $t=1,N$ calculate expectation value of target, $<q_t>$, using a supervised model retrained on all data except those from t^{th} sample
  - **Step 4a(i)** Create data subset leaving out one sample, ${}^{LOO}q_n = q_n \;\forall\; n \neq t$ and ${}^{LOO}x_{m,n} = x_{m,n} \;\forall\; n \neq t$
  - **Step 4a(ii)** Standardize each length (N-1) input variable time series to zero mean and unit variance on the basis of its sample mean and variance, and combine into ${}^{LOO}\mathbf{X}$
  - **Step 4a(iii)** Calculate correlation matrix, ${}^{LOO}\mathbf{C} = (N\text{-}1)^{-1}\,{}^{LOO}\mathbf{X}\,{}^{LOO}\mathbf{X}^{T}$
  - **Step 4a(iv)** Find eigenvectors ${}^{LOO}\mathbf{E}$ of ${}^{LOO}\mathbf{C}$, and scores matrix ${}^{LOO}\mathbf{A} = {}^{LOO}\mathbf{E}^{T}\,{}^{LOO}\mathbf{X}$
  - **Step 4a(v)** Create features matrix by selecting the same PCA modes from ${}^{LOO}\mathbf{A}$ as retained in Step 3a, ${}^{LOO}\mathbf{A'} = {}^{LOO}a_{m,n}$, $m \in (1,2,\ldots,M)$, $n = 1,N\text{-}1$
  - **Step 4a(vi)** Train predictive model, ${}^{LOO}f$, on ${}^{LOO}\mathbf{A'}$ and ${}^{LOO}\boldsymbol{q}$
  - **Step 4a(vii)** Retrieve left-out observational data, ${}^{CV}q_t = q_{n=t}$ and ${}^{CV}x_{m,t} = x_{m,n=t}$
  - **Step 4a(viii)** Standardize the left-out sample for each of the M input variables, ${}^{CV}x_{m,t}$, using sample means and variances of ${}^{LOO}x_{m,n}$ found in Step 4a(ii), and combine into ${}^{CV}\mathbf{X}$
  - **Step 4a(ix)** ${}^{CV}\mathbf{A'} = {}^{LOO}\mathbf{E}^{T}\;{}^{CV}\mathbf{X}$ using the same modes $m \in (1,2,\ldots,M)$ retained in Steps 3a and 4a(v) and the same ${}^{LOO}\mathbf{E}$ calculated in Step 4a(iv)
  - **Step 4a(x)** Estimate $<{}^{CV}q_{n=t}>$ using model, ${}^{LOO}f$, forced by the features in ${}^{CV}\mathbf{A'}$
- **Step 4b** Obtain performance measures (RMSE, R^2) by comparing $<{}^{CV}\boldsymbol{q}>$ and $\boldsymbol{q}$ time series

Algorithm 2: Scenario 2 (partial) LOOCV procedure

Step 1 Assemble available $n = 1,N$ samples of target q_n and of $m = 1,M$ input variables at the same N times, $x_{m,n}$

Step 2 Perform PCA on input variables

- **Step 2a** Standardize each length (N) predictor time series to zero mean and unit variance on the basis of its sample mean and variance and combine into data matrix, $\mathbf{X}$
- **Step 2b** Calculate correlation matrix, $\mathbf{C} = N^{-1}\mathbf{X}\,\mathbf{X}^{T}$
- **Step 2c** Find eigenvectors $\mathbf{E}$ of $\mathbf{C}$, and scores matrix $\mathbf{A} = \mathbf{E}^{T}\mathbf{X}$ for N samples and M modes

Step 3 Train predictive model on $\mathbf{A}$ and $\boldsymbol{q}$

- **Step 3a** Select subset of PCA modes in $\mathbf{A}$ to use as predictive features, $\mathbf{A'} = a_{m,n}$, $m \in (1,2,\ldots,M)$, $n=1,N$
- **Step 3b** Train a supervised learning algorithm, f, to predict expectation values of target, $<\boldsymbol{q}>$, from $\mathbf{A'}$

Step 4 Obtain out-of-sample predictive error estimate by leave-one-out cross-validation

- **Step 4a** For each $t=1,N$ calculate expectation value of target, $<q_t>$, using a supervised model retrained on all data except those from the t^{th} sample
  - **Step 4a(i)** Create subset of target and selected features leaving out one sample, ${}^{LOO}q_n = q_n \;\forall\; n \neq t$ and ${}^{LOO}\mathbf{A'} = \mathbf{A'} \;\forall\; n \neq t$, where $\mathbf{A'}$ is the original scores matrix found in Step 3a
  - **Step 4a(ii)** Train predictive model, ${}^{LOO}f$, on ${}^{LOO}\mathbf{A'}$ and ${}^{LOO}\boldsymbol{q}$
  - **Step 4a(iii)** Retrieve observational data for target and selected features at the left-out sample time, ${}^{CV}q_{n=t}$ and ${}^{CV}\mathbf{A'} = \mathbf{A'}$ for $n = t$ only
  - **Step 4a(iv)** Estimate $<{}^{CV}q_{n=t}>$ using model, ${}^{LOO}f$, forced by features, ${}^{CV}\mathbf{A'}$
- **Step 4b** Obtain performance measures (RMSE, R^2) by comparing $<{}^{CV}\boldsymbol{q}>$ and $\boldsymbol{q}$ time series

For both scenarios at both forecast points, we performed the following model-fitting exercises: (1) using all input variables and the resulting leading PCA mode as a predictor; (2) using all input variables and corresponding PCA modes 1 through 4 as predictors; and (3) using a genetic algorithm to optimize feature selection, spanning both input variable and corresponding PCA mode choices, and retaining up to two modes, a practical choice given that in operational WSF practice most PCR models retain only one or two PCA modes (see discussion below). Where a genetic algorithm is used, min(LOOCV RMSE) was the cost function, though other choices are of course possible. For instance, NRCS WSF modeling procedures have

historically used in-sample standard error during model development, including PCA mode selection, reserving LOOCV procedures for robustly reporting expected out-of-sample prediction performance for the finalized model and generating corresponding prediction intervals (e.g., Garen 1992). However, using cross-validated prediction error for PCA mode selection is more consistent with general statistics-community usage of PCR (e.g., James et al. 2013). This approach tends to yield a smaller number of retained modes and a more parsimonious model that is less likely to be overtrained; such considerations grow still more important when nonlinear machine learning prediction algorithms are substituted for linear regression, as is increasingly common practice.

Further, the entire foregoing process was repeated using each of five different prediction models within the overall PCR structure: (1) linear regression (i.e., classical PCR), (2) quantile regression, (3) random forests, (4) a support vector machine, and (5) a feed-forward error-backpropagation ANN. Algorithms, regularization steps, and other hyperparameters and procedures were generally as described in Fleming and Goodbody (2019); these steps are complex, and in the interest of conciseness, readers are referred there for further details. Fleming et al. (2021) provide additional details around the properties and performance of these five specific WSF models. Ten reruns were performed in each instance and the results were pooled, to account for stochasticity in several of the supervised learning procedures (e.g., random initial weights with nonlinear optimization in neural network training) and the evolutionary algorithm-based feature selection (e.g., random gene mutations).

RESULTS AND DISCUSSION

Outcomes from the 300 resulting models were broadly similar. Figure 1 gives a representative example; the remainder is provided in Figures S1–S12. As expected for any model of any type that is fit or calibrated to observational data, prediction error typically increases from in-sample to out-of-sample estimates. The size of that gap varies across forecast locations and dates and the predictive modeling method. In any given case, however, out-of-sample error estimates are virtually indistinguishable between the two cross-validation scenarios.

This conclusion may initially be surprising. Features available to the supervised learning step are calculated in the unsupervised learning step. In principle, using a slightly different dataset in each PCA during cross-validation should lead to fluctuations in PCA outcomes, which ought in turn propagate to the best-fit predictive model structure and parameters and thus, presumably, net predictive error.

That in practice this does not appear to be a significant effect is, however, intuitively consistent with other considerations. Cross-validated RMSE and R^2 correspond by definition to predictive models, like linear regression, feed-forward error-backpropagation ANNs, and the like. In contrast, PCA is not a predictive model having predictive errors per se. Rather, it is a decomposition of data into orthogonal basis functions, loosely akin to a Fourier transform, for example; and similarly, the original data can be fully reconstructed from these basis functions with no error. Notwithstanding variants developed for different tasks like missing data imputation, PCA is fundamentally a means for finding structure in data, not predicting data. While in PCR and PCR-like models, only a subset of PCA modes is retained as regression predictors and therefore some of the information in the original data is lost, this decision to keep or discard specific PCA modes reflects a standard question of feature selection in the subsequent predictive modeling step (e.g., classical forward stepwise linear regression modeling, or evolutionary algorithm-guided feature selection in a support vector machine). Consequently, it seems intuitively reasonable to treat PCA as an offline data pre-processing step that does not benefit from cross-validation the way the subsequent regression or regression-like step does. As noted above, this distinction seems still more relevant when the regression-like step uses highly flexible nonlinear machine learning techniques, which can be more vulnerable to overtraining. The wider statistical literature tends to confirm that cross-validation is most clearly meaningful for supervised learning tasks; it can be performed for unsupervised learning methods like PCA, but in that case, its interpretation is more subtle, theoretical and practical complications can be significant, and the best approach may be unclear (e.g., Bro et al. 2008).

Our result may also reflect established best practices for WSF applications of PCR and their ramifications for regularization. User protocols are typically in place at operational institutions for WSF model development. This injection of WSF-specific subject-matter expertise represents an often-underdiscussed, but typically valuable, human component of all real-world operational WSF systems (Weber et al. 2012; Wood et al. 2020). These protocols emphasize judicious PCA mode selection with an eye to geophysical interpretability, effectively corresponding to a form of theory-guided data science (Karpatne et al. 2017). Specifically, the final models usually retain only the leading PCA mode, which is (given typical WSF predictors like those used here) a convenient

FIGURE 1. Illustrative example of results: root mean square error (kaf) for 50 models developed using five supervised learning methods with principal component analysis predictor data pre-processing for Rio Grande January 1 water supply forecast using genetic algorithm-based feature selection. SL leave-one-out cross-validated (LOOCV) refers to cross-validation on the supervised learning (SL) portion of the principal component regression (PCR) or PCR-like modeling process; full LOOCV is cross-validation across both unsupervised and supervised steps. Outcomes are similar for R^2. Details vary substantially between supervised modeling techniques, modeling runs, and forecast locations and dates. However, general relationships between SL and full LOOCV are generally consistent across all 300 models.

watershed-scale index of observed winter climate conditions, or occasionally the first two modes, where the second often captures aquifer-stream interactions; only very rarely is a third retained (e.g., Fleming et al. 2021). These restrictive a priori decisions on number of PCA modes retained provides, in effect, a geophysically informed limit to overtraining potentially occurring in the unsupervised learning portion of the PCR procedure. Such basic geophysical signals can in general be reliably extracted from available SNOTEL and naturalized USGS data using PCA of environmental records sampled over a standard hydroclimatic normal period or something similar. That is, typical WSF practice of using about three decades of data yields stable PCA-derived signal estimates. This was confirmed in practice: for the first one or two PCA modes, for instance, eigenvectors were typically very similar, and often almost identical, between the full dataset and the 30 length $N - 1$ datasets constructed during cross-validation in the scenario (1) models.

As a corollary, we observed that differences between in-sample and cross-validated (either full or partial) WSF errors are larger as more PCA modes are retained. That is, relative to using only the leading mode as a predictor, using all of the first four modes improved in-sample but worsened out-of-sample performance. This finding has practical implications for whether to use in-sample or cross-validated errors for PCA mode selection in WSF, as it suggests the former will lead to retention of more modes (due to lower in-sample error) and ultimately an overtrained solution (captured by higher out-of-sample error). As noted above, using cross-validated

regression error as the basis for PCA mode selection is also consistent with general statistics community practice. However, this does not imply that historical NRCS and other PCR implementations using in-sample regression error to select PCA modes gave overtrained models, because accompanying model development protocols force geophysically based and conservative PCA mode selections (see above). It is well-recognized that truncating PCA modes itself regularizes the subsequent regression compared to using all input data, and the more severe the truncation, the stronger the overfitting mitigation; and more broadly, that using physical process knowledge to constrain relationships captured by machine learning contributes additional regularization (e.g., Zhang and Zhang 1999; Karpatne et al. 2017). Nonetheless, it could be prudent to use cross-validated error for feature optimization, especially if more automated ("over-the-loop"; e.g., Wood et al. 2020; Fleming et al. 2021) WSF frameworks are adopted. This recommendation seems consistent with preliminary experiments (not shown) which appear to confirm, again irrespective of whether full or partial cross-validation is used, that setting the genetic algorithm objective function to min(in-sample RMSE) tends to slightly increase overtraining relative to min(LOOCV RMSE), particularly for machine learning-based supervised models.

CONCLUSIONS

Numerical experiments suggest WSF skill estimates are largely indistinguishable between cross-validation across the full PCR or PCR-like machine learning process vs. cross-validation performed only on the final (predictive modeling) step of that procedure. This is particularly apparent when other sources of model performance variability are taken into account, like slightly different outcomes when model development processes have a stochastic component, such as some algorithms for machine learning or feature optimization, and in particular when various different methods are adopted for the predictive modeling step, for example, linear regression vs. random forests (Figure 1). PCA might therefore best be viewed as an offline data-compression and signal-boosting technique in PCR/PCR-like WSF.

The result may help inform future WSF model development in three ways. First, it allows for more streamlined workflows and more computationally efficient algorithms. Second, it gives some reassurance that cross-validation may provide reliable WSF accuracy estimates when it is cumbersome to truly leave out all input information for a given timestep during the LOOCV process, as might be the case for some climatological products for instance. Third, because the result partly reflects regularization provided by geophysically guided modeling protocols leading to conservative PCA mode selections, it emphasizes the continued usefulness of manual hydrologic expertise during model development, even (or especially) in emerging machine-learning based approaches.

DATA AVAILABILITY STATEMENT

Data used in this study are available at NRCS (2021); see above text and Tables S1 and S2 for further information.

AUTHOR CONTRIBUTIONS

Sean W. Fleming carried out conceptualization, data curation, formal analysis, investigation, methodology, project administration, resources, software, validation, visualization, writing—original draft, writing—reviewing and editing of the manuscript. David C. Garen carried out investigation, methodology, writing—original draft, writing—reviewing and editing of the manuscript.

SUPPORTING INFORMATION

Additional supporting information may be found online under the Supporting Information tab for this article: This information includes Tables S1 and S2 describing the model input data, and Figures S1 through S12 summarizing outcomes from the modeling experiments.

LITERATURE CITED

Bro, R., K. Kjeldahl, A.K. Smilde, and H.A.L. Kiers. 2008. "Cross-Validation of Component Models: A Critical Look at Current Methods." *Analytical and Bioanalytical Chemistry* 390: 1241–51.

Bureau of Reclamation. 2016. *SECURE Water Act Section 9503(c) — Reclamation Climate Change and Water 2016*. Denver CO: Bureau of Reclamation, Policy and Administration.

Carrier, C., A. Kalra, and S. Ahmad. 2013. "Using Paleo Reconstructions to Improve Streamflow Forecast Lead Time in the Western United States." *Journal of the American Water Resources Association* 49: 1351–66.

Cook, E.R., D.M. Meko, D.W. Stahle, and M.K. Cleaveland. 1999. "Drought Reconstructions for the Continental United States." *Journal of Climate* 12: 1145–62.

Eldaw, A.K., J.D. Salas, and L.A. Garcia. 2003. "Long-Range Forecasting of the Nile River Flows Using Climatic Forcing." *Journal of Applied Meteorology* 42: 890–904.

Fleming, S.W., D.C. Garen, A.G. Goodbody, C.S. McCarthy, and L.C. Landers. 2021. "Assessing the New Natural Resources Conservation Service Water Supply Forecast Model for the American West: A Challenging Test of Explainable, Automated, Ensemble Artificial Intelligence." *Journal of Hydrology* 602: 126782.

Fleming, S.W., and A.G. Goodbody. 2019. "A Machine Learning Metasystem for Robust Probabilistic Nonlinear Regression-Based Forecasting of Seasonal Water Availability in the US West." *IEEE Access* 7: 119943–64.

Fleming, S.W., and H.V. Gupta. 2020. "The Physics of River Prediction." *Physics Today* 73: 46–52.

Garen, D.C. 1992. "Improved Techniques in Regression-Based Streamflow Volume Forecasting." *Journal of Water Resources Planning and Management* 118: 654–69.

Glabau, B., E. Nielsen, A. Mylvahanan, N. Stephan, C. Frans, K. Duffy, J. Giovando, and J. Johnson. 2020. "Climate and Hydrology Datasets for RMJOC Long-Term Planning Studies, Second Edition, Part II: Columbia River Reservoir Regulation and Operations – Modeling and Analyses." River Management Joint Operating Committee. www.bpa.gov/p/Generation/Hydro/Documents/RMJOC-II_Part_II.PDF.

Gobena, A.K., and T.Y. Gan. 2010. "Incorporation of Seasonal Climate Forecasts in the Ensemble Streamflow Prediction System." *Journal of Hydrology* 385: 336–52.

Gobena, A.K., F.A. Weber, and S.W. Fleming. 2013. "The Role of Large-Scale Climate Modes in Regional Streamflow Variability and Implications for Water Supply Forecasting: A Case Study of the Canadian Columbia Basin." *Atmosphere-Ocean* 51: 380–91.

Hamlet, A.F., D. Huppert, and D.P. Lettenmaier. 2002. "Economic Value of Long-Lead Streamflow Forecasts for Columbia River Hydropower." *ASCE Journal of Water Resources Planning and Management* 128: 91–101.

Harpold, A.A., K. Sutcliffe, J. Clayton, A. Goodbody, and S. Vazquez. 2017. "Does Including Soil Moisture Observations Improve Operational Streamflow Forecasts in Snow-Dominated Watersheds?" *Journal of the American Water Resources Association* 53: 179–96.

Hsieh, W.W., L.J. Yuval, A. Shabbar, and S. Smith. 2003. "Seasonal Prediction with Error Estimation of Columbia River Streamflow in British Columbia." *Journal of Water Resource Planning and Management* 129: 146–49.

James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning with Applications in R*. New York: Springer.

Karpatne, A., G. Atluri, J.H. Faghmous, M. Steinback, A. Banerjee, A. Ganguly, S. Shekhar, N. Samatova, and V. Kumar. 2017. "Theory-Guided Data Science: A New Paradigm for Scientific Discovery from Data." *IEEE Transactions on Knowledge and Data Engineering* 29: 2318–31.

Kennedy, A.M., D.C. Garen, and R.W. Koch. 2009. "The Association between Climate Teleconnection Indices and Upper Klamath Seasonal Streamflow: Trans-Niño Index." *Hydrological Processes* 23: 973–84.

Lehner, F., A.W. Wood, D. Llewellyn, D.B. Blatchford, A.G. Goodbody, and F. Pappenberger. 2017. "Mitigating the Impacts of Climate Nonstationarity on Seasonal Streamflow Predictability in the US Southwest." *Geophysical Research Letters* 44: 12208–17.

Mantua, N.J., S.R. Hare, Y. Zhang, J.M. Wallace, and R.C. Francis. 1997. "A Pacific Interdecadal Climate Oscillation with Impacts on Salmon Production." *Bulletin of the American Meteorological Society* 78: 1069–79.

Moradkhani, H., and M. Meier. 2010. "Long-Lead Water Supply Forecast Using Large-Scale Climate Predictors and Independent Component Analysis." *Journal of Hydrologic Engineering* 15: 744–62.

Najafi, M.R., and H. Moradkhani. 2016. "Ensemble Combination of Seasonal Streamflow Forecasts." *Journal of Hydrologic Engineering* 21: https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250.

NRCS. 2021. *Report Generator 2.0*. Portland, OR: National Water and Climate Center, Natural Resources Conservation Service, US Department of Agriculture. https://wcc.sc.egov.usda.gov/reportGenerator.

Oubeidillah, AA, Tootle, GA, Moser, C, Piechota, T, and Lamb, K. 2011. "Upper Colorado River and Great Basin Streamflow and Snowpack Forecasting using Pacific Oceanic–Atmospheric Variability." *Journal of Hydrology* 410: 169–177.

Pagano, T.C., D.C. Garen, and S. Sorooshian. 2004. "Evaluation of Official Western US Seasonal Water Supply Outlooks, 1922–2002." *Journal of Hydrometeorology* 5: 896–909.

Perkins, T.R., T.C. Pagano, and D.C. Garen. 2009. "Innovative Operational Seasonal Water Supply Forecasting Technologies." *Journal of Soil and Water Conservation* 64: 15–17.

Regonda, S.K., B. Rajagopalan, and M. Clark. 2006. "A New Method to Produce Categorical Streamflow Forecasts." *Water Resources Research* 42. https://doi.org/10.1029/2006WR004984.

Regonda, S.K., B. Rajagopalan, M. Clark, and E. Zagon. 2006. "A Multimodel Ensemble Forecast Framework: Application to Spring Seasonal Flows in the Gunnison River Basin." *Water Resources Research* 42. https://doi.org/10.1029/2005WR004653.

Risley, J.C., M.W. Gannett, J.K. Lea, and E.A. Roehl Jr. 2005. "An Analysis of Statistical Methods for Seasonal Flow Forecasting in the Upper Klamath River Basin of Oregon and California." Scientific Investigations Report 2005-5177, US Geological Survey, Reston, VA.

Rosenberg, E.A., A.W. Wood, and A.C. Steinemann. 2011. "Statistical Applications of Physically Based Hydrologic Models to Seasonal Streamflow Forecasts." *Water Resources Research* 47. https://doi.org/10.1029/2010WR010101.

Weber, F.A., D.C. Garen, and A.K. Gobena. 2012. "Invited Commentary: Themes and Issues from the Workshop, 'Operational River Flow and Water Supply Forecasting'." *Canadian Water Resources Journal/Revue Canadienne Des Ressources Hydriques* 37: 151–61.

Wood, A., L. Woelders, and J. Lukas. 2020. "Streamflow Forecasting." In *Chap. 8 in Colorado River Basin Climate and Hydrology: State of the Science*, edited by J. Lukas and E. Payton, 287–333. Boulder, CO: Western Water Assessment, University of Colorado Boulder.

Zhang, H., and Z. Zhang. 1999. "Feedforward Networks with Monotone Constraints." Proceedings of the IEEE International Joint Conference On Neural Networks, Washington DC, July 10–16 1999, Volume 3, 1820–23.

Supplementary Information

Fleming and Garen, JAWRA, 2022

Station name	Variable type	Measurement date/date range
Big Meadow	Instantaneous SWE	April 1
Big Meadow	Accumulated precipitation	October 1-March 31
CSS Lab	Instantaneous SWE	April 1
CSS Lab	Accumulated precipitation	October 1-March 31
Donner Summit	Instantaneous SWE	April 1
Independence Camp	Instantaneous SWE	April 1
Independence Camp	Accumulated precipitation	October 1-March 31
Independence Creek	Instantaneous SWE	April 1
Independence Creek	Accumulated precipitation	October 1-March 31
Independence Lake	Instantaneous SWE	April 1
Independence Lake	Accumulated precipitation	October 1-March 31
Mt Rose Ski Area	Instantaneous SWE	April 1
Mt Rose Ski Area	Accumulated precipitation	October 1-March 31
Squaw Valley GC	Instantaneous SWE	April 1
Squaw Valley GC	Accumulated precipitation	October 1-March 31
Tahoe City Cross	Instantaneous SWE	April 1
Tahoe City Cross	Accumulated precipitation	October 1-March 31
Truckee Num 2	Instantaneous SWE	April 1
Truckee Num 2	Accumulated precipitation	October 1-March 31
Ward Creek Num 2	Instantaneous SWE	April 1
Ward Creek Num 3	Instantaneous SWE	April 1
Ward Creek Num 3	Accumulated precipitation	October 1-March 31
Webber Lake	Instantaneous SWE	April 1
Webber Peak	Instantaneous SWE	April 1
Truckee River at Farad	Accumulated antecedent flow volume	October 1-March 31

Table S1 Pool of $M = 25$ predictors used for April 1 forecast of yearly April-July flow volume for the Truckee River at Farad. For WSF modeling runs where genetic algorithm is used for feature optimization, these predictors form a candidate pool; for runs where automated feature selection is not used, all these predictors are employed. Station name refers to NRCS SNOTEL or CCSS automated snow and climate monitoring station or snow course location for accumulated precipitation and snow water equivalent (SWE) data, or to USGS gage location for streamflow volume data. $N = 30$ annual values over 1986-2015 were used for each predictor. This predictor list is strongly guided by long-term NRCS experiential knowledge with WSF at this location as captured in its current operational forecast models, and the overall predictor selections and WSF problem setup are typical of data-driven operational WSF models in western North America generally. See article main text and Fleming et al. (2021) for further details. All data are freely available at NRCS (2021).

Station name	Variable type	Measurement date/date range
Beartown	Instantaneous SWE	January 1
Beartown	Accumulated precipitation	October 1-December 31
Lily Pond	Instantaneous SWE	January 1
Lily Pond	Accumulated precipitation	October 1-December 31
Middle Creek	Instantaneous SWE	January 1
Middle Creek	Accumulated precipitation	October 1-December 31
Slumgullion	Instantaneous SWE	January 1
Slumgullion	Accumulated precipitation	October 1-December 31
Upper San Juan	Instantaneous SWE	January 1
Upper San Juan	Accumulated precipitation	October 1-December 31

Table S2 As in Table S1, but for $M = 10$ predictors used in January 1 forecast model for yearly April-September flow volume at the Rio Grande near Del Norte. See Table S1 caption, and main article text, for additional details.

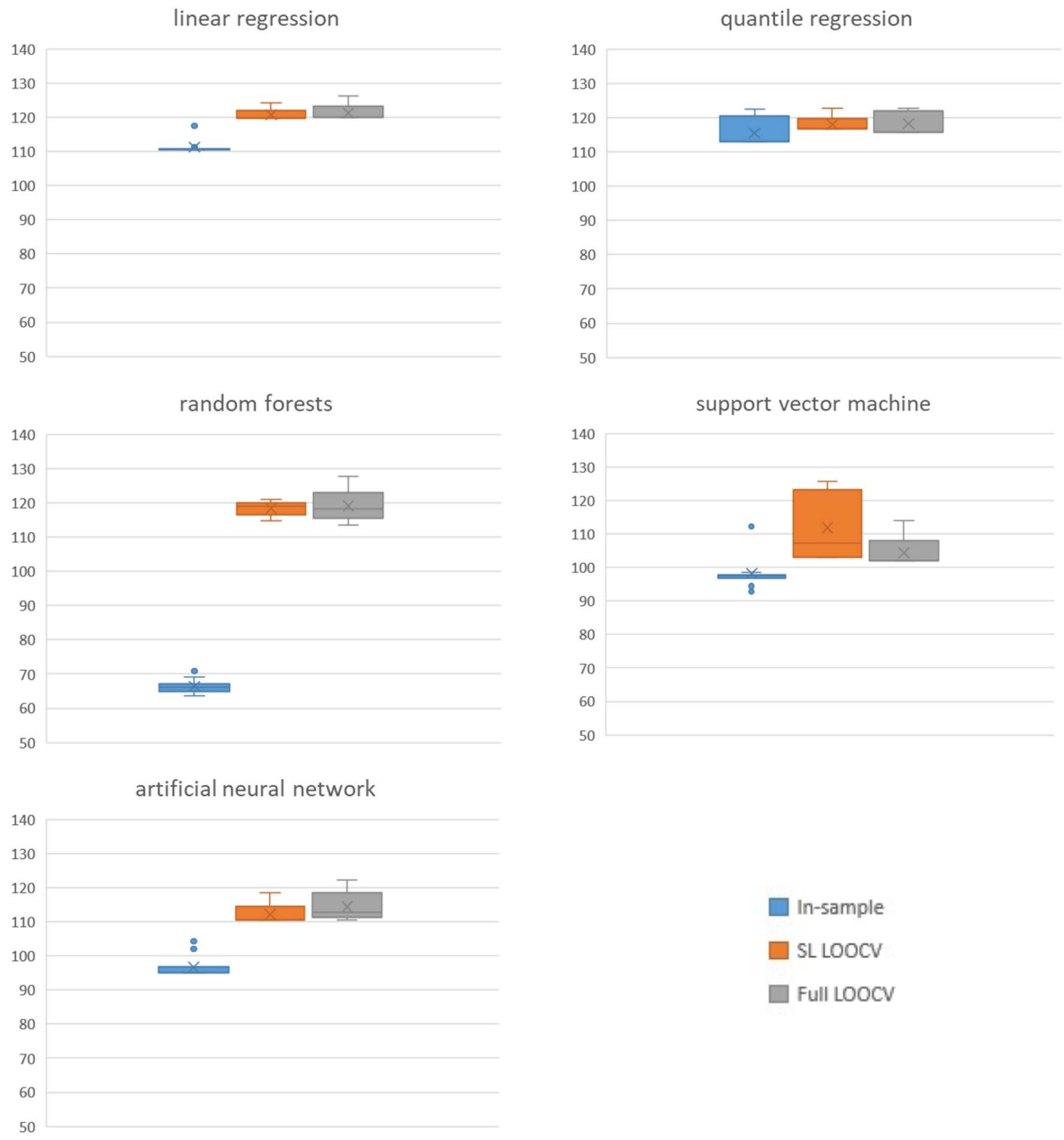

Figure S1. RMSE (kaf) for 50 models developed using five supervised learning methods with PCA predictor data pre-processing for Rio Grande January 1 WSF using genetic algorithm-based feature selection retaining up to two PCA modes. SL LOOCV refers to cross-validation on the supervised learning portion of the PCR or PCR-like modeling process; full LOOCV is cross-validation across both unsupervised and supervised steps. This is the most realistic (see text of main article and Fleming et al., 2021) set of scenarios for PCR/PCR-like WSF model development and implementation at NRCS. (As in Figure 1 of main article)

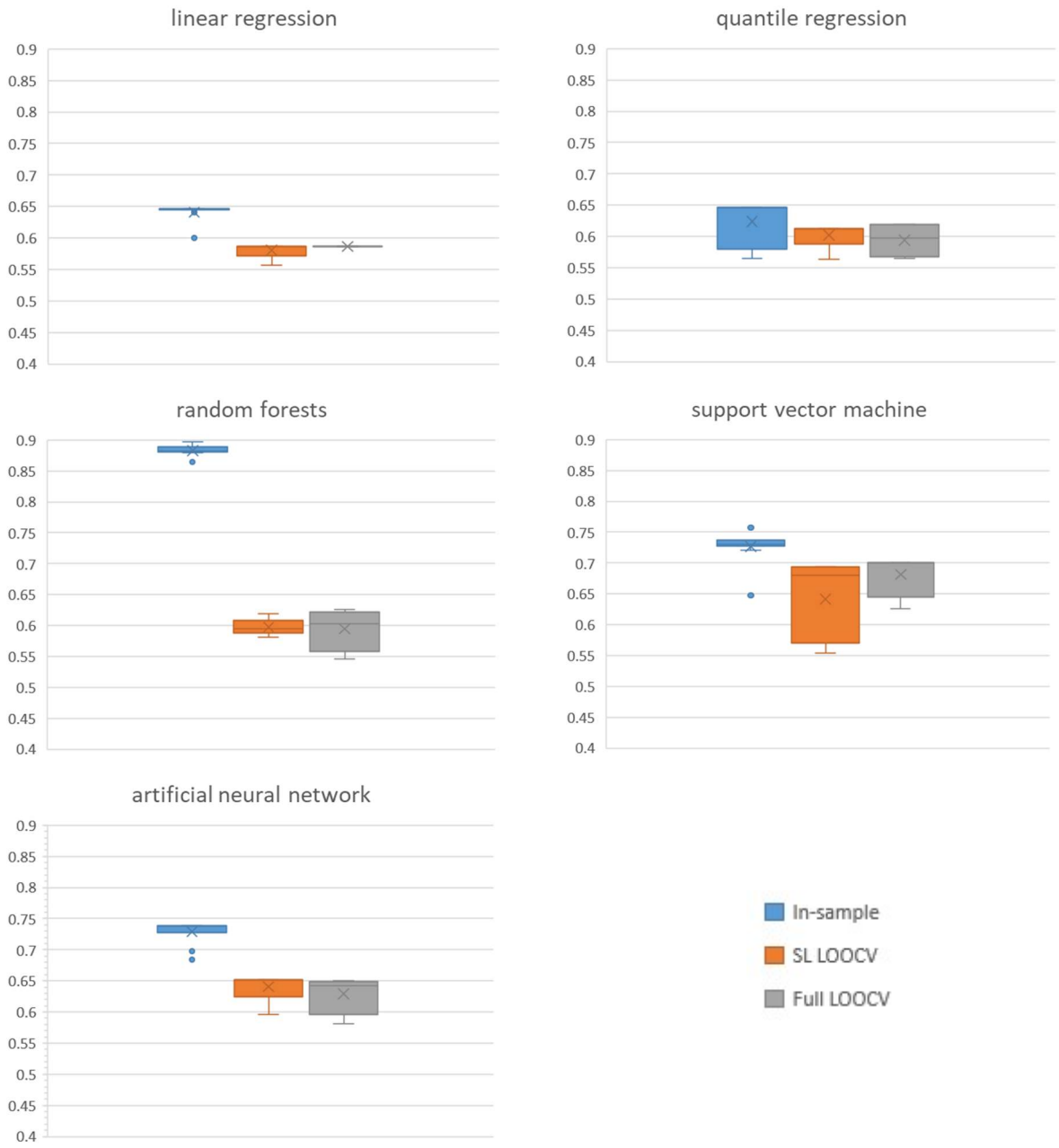

Figure S2. As in Figure S1 but for R^2

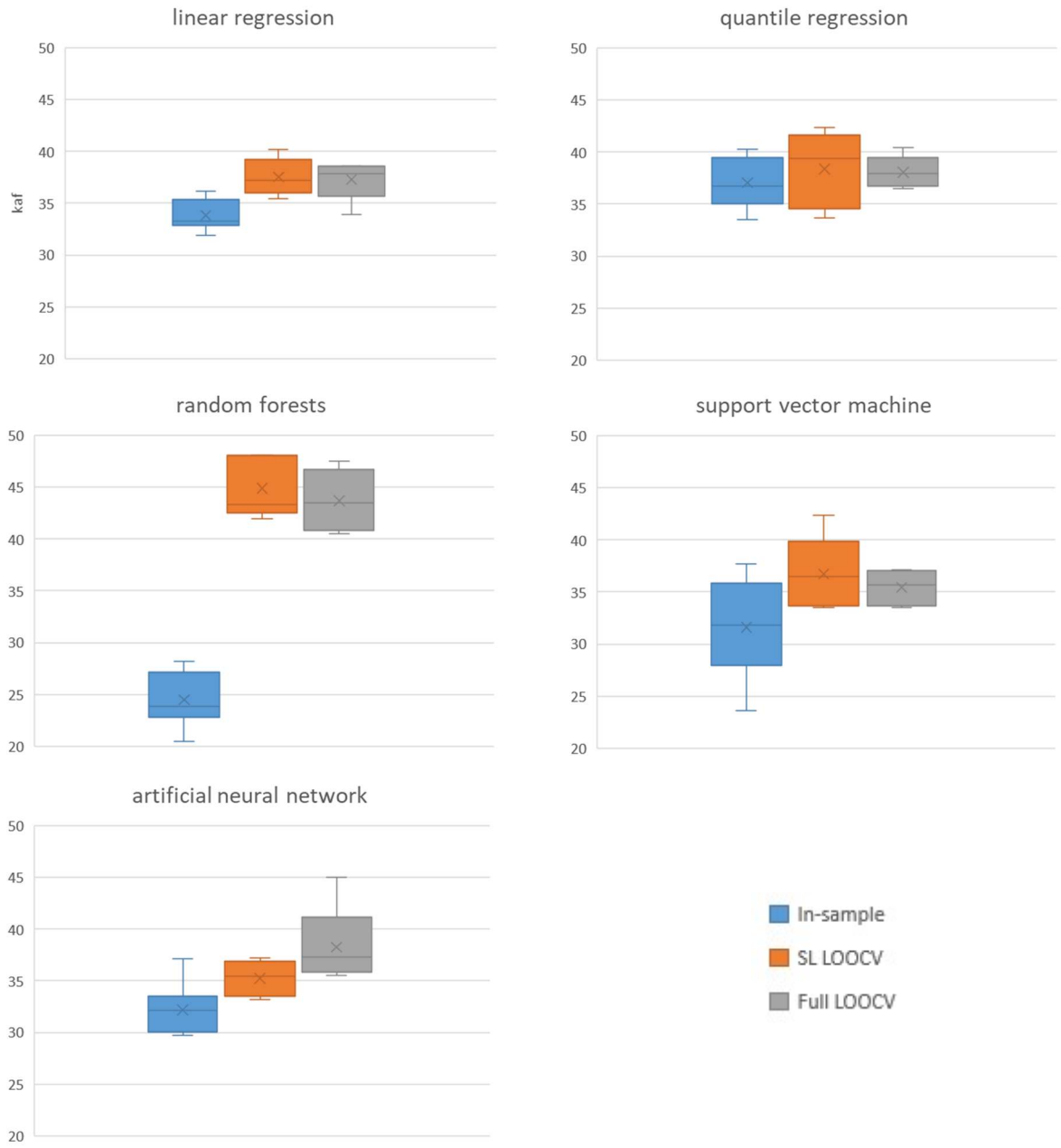

Figure S3. As in Figure S1 but for Truckee River April 1 WSF

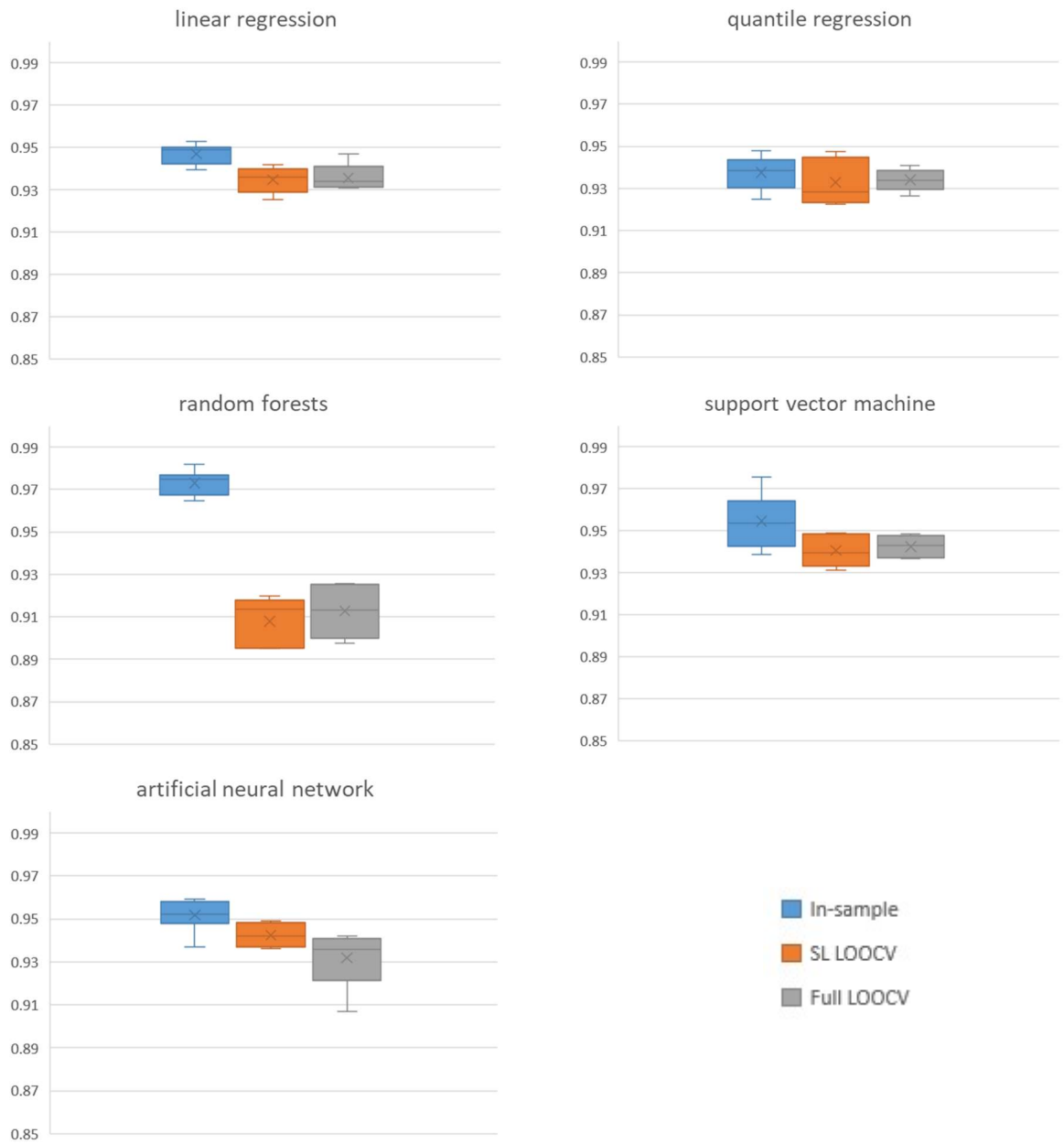

Figure S4. As in Figure S3 but for R^2

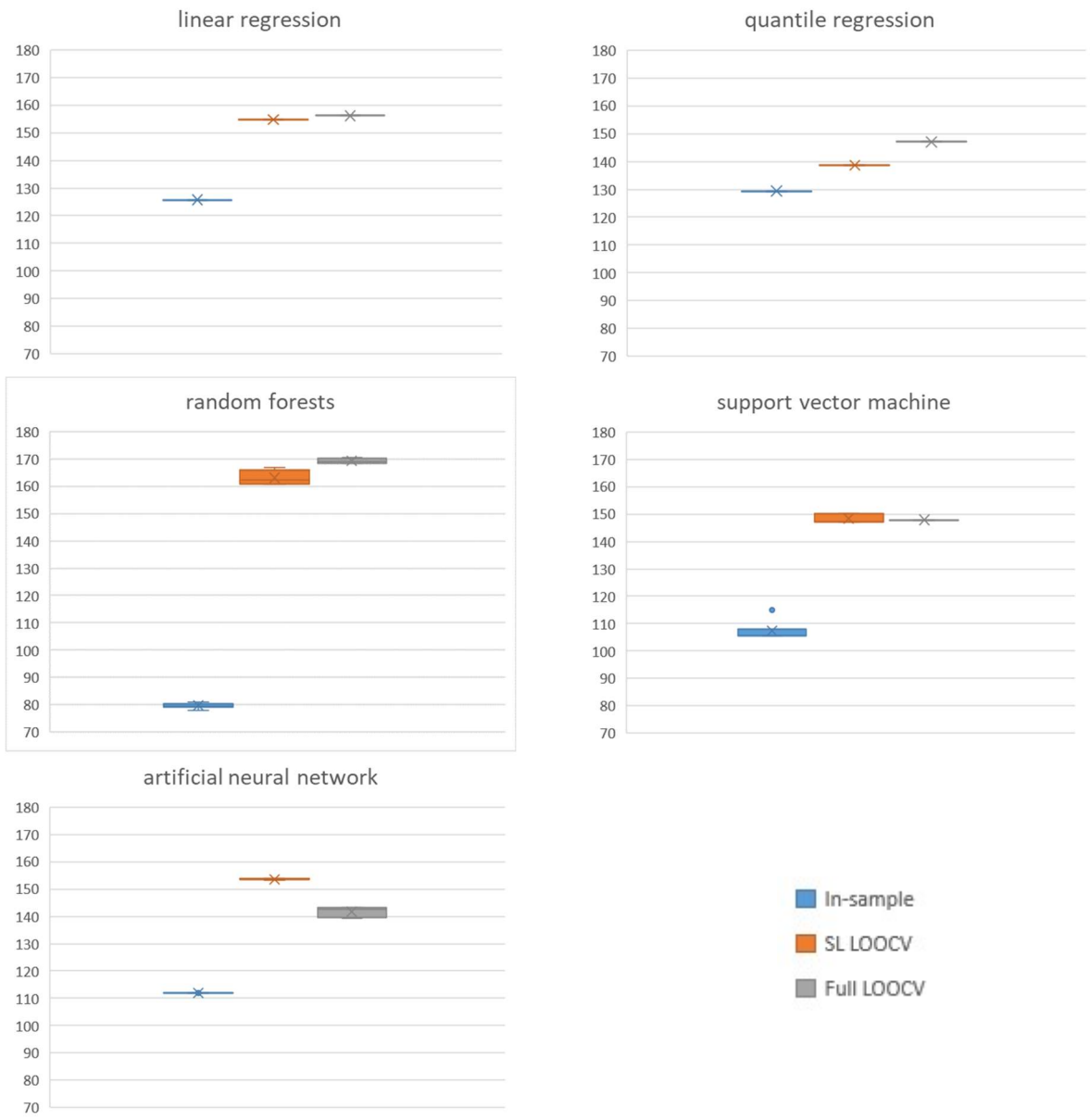

Figure S5. As in Figure S1 but using PCA modes 1 through 4 inclusive as the features delivered to the supervised learner, with no automated feature selection. As such, stochasticity inherent to genetic algorithm-based feature optimization is by construction absent in these modeling scenarios, and outcomes for linear and quantile regression are therefore deterministic, unlike Figures S1-S4. The three machine learning-based (random forests, support vector machine, artificial neural network) methods all retain stochasticity in the initialization and training process, giving a range of outcomes somewhat akin to Figures S1-S4, but with much less variability due to the aforementioned absence of stochastic feature optimization. In practical WSF applications only the leading one or, occasionally, two PCA modes are retained (see main article text), with higher modes generally corresponding to noise, so imposing use of the top four modes as features in PCR/PCR-like models is non-parsimonious, forces a fit to noise, and may represent a worst-case scenario, among those considered here, around overtraining and in-sample vs. out-of-sample performance characteristics.

Figure S6. As in Figure S5 but for R^2.

Figure S7. As in Figure S5 but for Truckee River April 1 WSF.

Figure S8. As in Figure S7 but for R^2.

Figure S9. As in Figure S5 but using the leading PCA mode as the sole feature delivered to the supervised learner. As in Figures S5-S8 (and unlike Figures S1-S4) stochastic genetic algorithm-based feature optimization is not used. Degree of stochasticity in outcomes from the three machine learning algorithms (random forests, support vector machine, and artificial neural network) has some loose tendency to be slightly lower than in Figure S5, because the Figure S9 models are more parsimonious due to the three fewer input features used relative to Figure S5, giving a lesser number of machine learning parameters to be estimated stochastically during the training process for each of those models. Note that retention of a single mode is the most common, but not sole, outcome encountered in mainstream operational applications of PCR/PCR-like models to WSF (see main article text).

Figure S10. As in Figure S9 but for R^2.

Figure S11. As in Figure S9 but for Truckee River April 1 WSF.

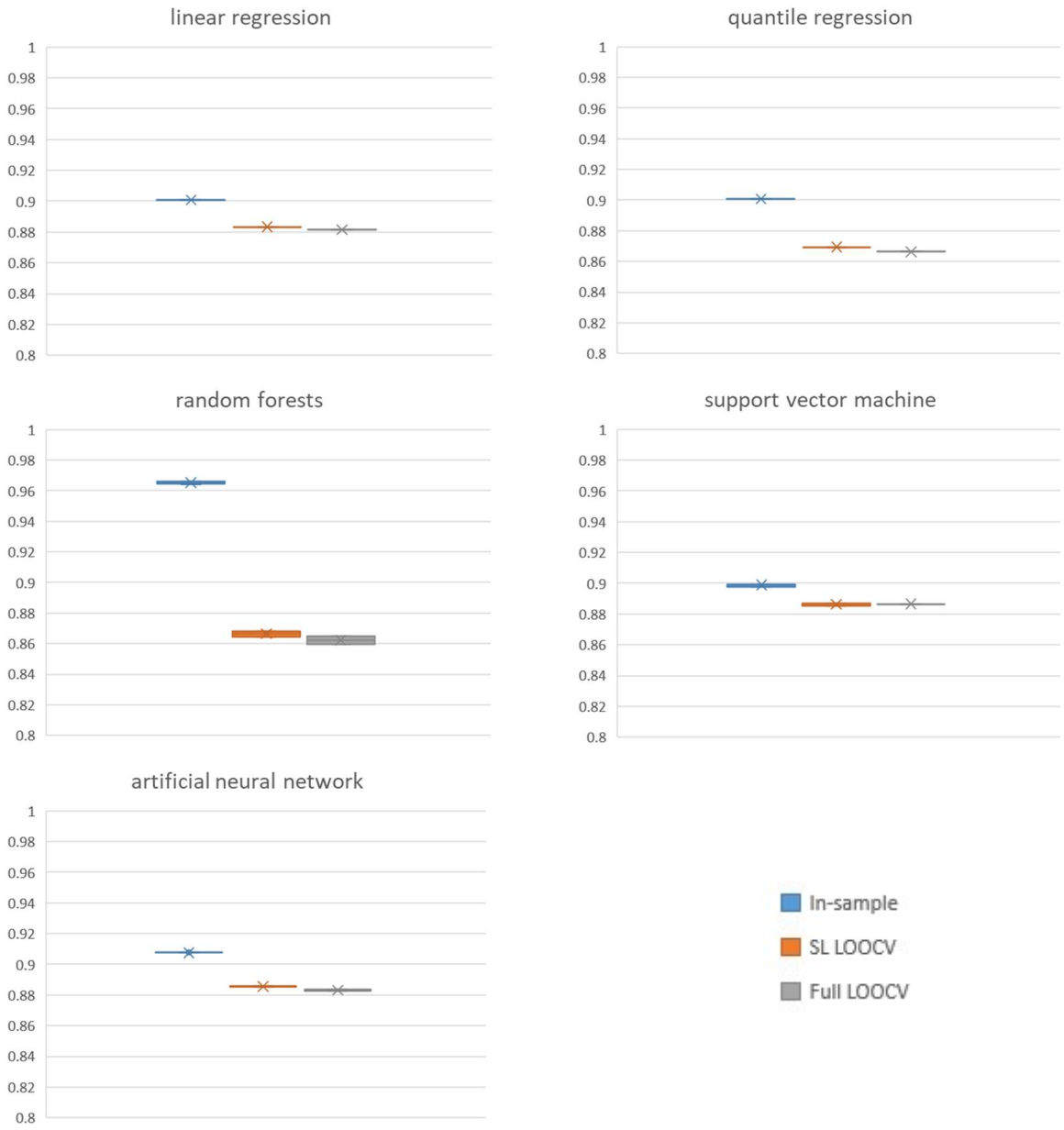

Figure S12. As in Figure S11 but for R^2.