IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received July 10, 2019, accepted August 19, 2019, date of publication August 22, 2019, date of current version September 9, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2936989

A Machine Learning Metasystem for Robust Probabilistic Nonlinear Regression-Based Forecasting of Seasonal Water Availability in the US West

SEAN W. FLEMING[1] AND ANGUS G. GOODBODY[2]
[1]White Rabbit R&D LLC, Corvallis, OR 97333, USA
[2]Natural Resources Conservation Service, U.S. Department of Agriculture, Portland, OR 97232, USA

Corresponding author: Sean W. Fleming (whiterabbit.research.llc@gmail.com)

This work was supported in part by funding to White Rabbit R&D LLC from the US Department of Agriculture through Elyon International Inc.

ABSTRACT Hydroelectric power generation, water supplies for municipal, agricultural, manufacturing, and service industry uses including technology-sector requirements, dam safety, flood control, recreational uses, and ecological and legal constraints, all place simultaneous, competing demands on the heavily stressed water management infrastructure of the mostly arid American West. Optimally managing these resources depends on predicting water availability. We built a probabilistic nonlinear regression water supply forecast (WSF) technique for the US Department of Agriculture, which runs the largest stand-alone WSF system in the US West. Design criteria included improved accuracy over the existing system; uncertainty estimates that seamlessly handle complex (heteroscedastic, non-Gaussian) prediction errors; integration of physical hydrometeorological process knowledge and domain-specific expert experience; ability to accommodate nonlinearity, model selection uncertainty and equifinality, and predictor multicollinearity and high dimensionality; and relatively easy, low-cost implementation. Some methods satisfied some of these requirements but none met all, leading us to develop a novel, interdisciplinary, and pragmatic prediction metasystem through a carefully considered synthesis of well-established, off-the-shelf components and approaches, spanning supervised and unsupervised machine learning, nonparametric statistical modeling, ensemble learning, and evolutionary optimization, focusing on maintaining but radically updating the principal components regression framework widely used for WSF. Testing this integrated multi-method prediction engine demonstrated its value for river forecasting; USDA adoption is a landmark for transitioning machine learning from research into practice in this field. Its ability to handle all the foregoing design criteria and requirements, which are not unique to WSF, suggests potential for extension to complex probabilistic prediction problems in other fields.

INDEX TERMS Machine learning, regression analysis, forecast uncertainty, hydroelectric power generation, water resources, environmental management, industry applications.

I. INTRODUCTION

President Teddy Roosevelt's 1901 description of the American West, ''Whoever controls the stream practically controls the land,'' remains true today. The combination of generally dry but highly variable climate, high water demand, immense economic scale and its sensitivity to water and energy availability, and strong technical capacity and resourcing, has made the western US a proving ground for new water management technologies.

Forecasts of spring-summer river runoff volumes, issued starting the previous winter and based largely but not exclusively on mountain snowpack measurements, are a cornerstone of the vast organizational and engineering infrastructure around water in this region. The implications of this predictive environmental information span agricultural, industrial, and municipal water supplies, hydroelectric power generation, and environmental and legal constraints,

The associate editor coordinating the review of this article and approving it for publication was Bora Onat.

such as international treaty requirements and legal decisions around required ecological flows. Even modest incremental improvements in water supply forecast (WSF) skill can yield over $(US)100 million per year in benefit for a single river basin [1].

Implications of WSF accuracy to hydropower generation and electricity pricing in the Pacific Northwest are an interesting example of the overall worth and some of the socioeconomic ripple effects of WSFs. Hydroelectric power producers must operate reservoirs to meet a dynamic portfolio of social, economic, and environmental objectives, including power generation – which grows increasingly challenging under, for example, a shift from winter heating demand peaks in the past to record-high summer cooling demands due to climate change, and the need to use hydropower to even out the high-frequency variability of growing generation from green energy sources like wind power [2], [3] – as well as dam safety, flood control, and water licensing. Legal obligations also play an important day-to-day role in operating choices around hydroelectric reservoirs; examples from just the Columbia River Basin alone include court decisions on required ecological minimum flows such as the Biological Opinions (BiOps) for the Federal Columbia River Power System, and coordinated Canada-US reservoir operation and management under the Columbia River Treaty. Foreknowledge of reservoir inflows is crucial to reliably attaining these goals, so hydropower producers in the region rely on WSFs to safely run and in some cases optimize their systems, and certain large power producers, such as Bonneville Power Authority and BC Hydro, additionally run their own operational river forecasting systems and teams to complement information available from government agencies [4]–[6]. Moreover, in a hydropower context, WSFs amount to predictions of the supply of de facto ''fuel'' available to generate energy [7]. Some hydroelectric utilities and power brokers (e.g., [8]) in the Pacific Northwest have used WSFs to help set the pricing of futures contracts on electricity in the western interconnection, and disparities in WSF information provided by different agencies using different input data and modeling approaches and used by different utilities have been assessed by power traders for potential competitive advantage.

The largest stand-alone WSF system in the American West is that of the US Department of Agriculture's Natural Resources Conservation Service (NRCS), spanning over 600 forecast locations [9], [10]. The probabilistic WSF system currently used by NRCS was revolutionary in the industry when introduced in the early 1990s due to its adoption of both principal components regression (PCR) to address input multicollinearity issues typical of WSF problems, and a probabilistic forecasting philosophy such that best-estimate predictions were accompanied by statistically rigorous prediction intervals. However, specific performance limitations, including difficulties reproducing nonlinear functional relationships or heteroscedastic and non-Gaussian prediction bounds without extensive and subjective manual interventions in the modeling process, as well as logistical considerations such as budget and staffing restrictions, argue for a fresh approach based on modern automated data science concepts.

We modernized the NRCS WSF system using machine learning. In practice, this had to be accomplished in such a way as to (a) solve known technical limitations with the existing system; (b) accommodate disciplinary subject matter expertise and experience around river and reservoir inflow prediction; and (c) integrate the specific operational requirements of a WSF-issuing federal government agency in general and the NRCS specifically. These three overall design principles in turn led to multiple design criteria, briefly discussed below in Section I.B. Some advanced statistical and machine learning techniques satisfied some of these criteria but none met all, leading us to develop a new but pragmatic framework, integrating multiple solution pathways drawn from a diverse range of existing, off-the-shelf methods in several disciplines including machine learning, advanced statistical modeling, and process-based water resource, climate, and ecological modeling.

When considering prediction algorithms to include in our technique or to use as a performance benchmark against which to meaningfully compare it, it is necessary to recognize that we do not have a blank canvas to work on, and that the successful intersection of machine learning with operational water supply forecasting requires some methodological and study design choices to be made that may not be entirely obvious. Though the technique and the lessons learned designing it are likely to be more broadly relevant (see below), it is nonetheless built for a specific purpose: operational forecasting of seasonal water supplies in the US West by a federal government agency that has been performing this task since the 1920s. To be accepted by the water resource science and engineering community and its forecast product users, any predictive method must align with the established body of knowledge and practice in that specific field. The consequences for forecast model design are twofold: our new system cannot be developed completely from scratch; and the vast majority of available data-driven prediction algorithms are not on the table for potential adoption. Rather, three general considerations – institutional requirements around what is and what is not a logistically feasible applied science and engineering solution to this specific prediction task, the largely successful decades-long track record of the existing PCR-based forecast system, and that system's consistency with the large body of environmental and geophysical science knowledge around water resources and their prediction – point in combination to a solution that involves building upon the existing PCR framework. By this we mean multicollinearity mitigation and dimensionality reduction through independent signal extraction from specific known classes of geophysical predictor datasets, followed by a phenomenological modeling process relating selected signals to the predictand through some form of regression-like input-output mapping, along with an automated process for optimally choosing which candidate input

variables, and signals derived from them in the initial data pre-processing step, to retain in the final model. Even within the constraints imposed by this powerful guidance around the range of viable solutions, however, there is abundant opportunity to radically update and upgrade the existing framework with a diverse selection of machine learning, ensemble modeling, and evolutionary computing approaches.

That said, the design criteria in Section I.B – which are largely centered around a combination of prediction accuracy improvements, increased flexibility and robustness, combining established practice domain-specific principles and practices with artificial intelligence and evolutionary computing methods, and low cost and risk around institutional adoption – is pertinent to other applied prediction problems. As such, while the resulting technique is most immediately relevant to environmental management and optimization of natural resources, such as hydroelectric power generation, it may also suggest viable practical approaches for certain problems commonly encountered by applied scientists and engineers in applications of machine learning to prediction of complex open systems in other fields as well.

A. PRIOR WORK

WSF originated in the 1920s, based on manual snow measurements by mountaineering-savvy scientists and engineers and simple back-of-the-envelope water volume calculations based on those data. Modern WSF prediction systems build on those long-established fundamentals using either process-based or data-driven approaches. Process simulation approaches are mathematical models that deterministically represent the large number of geophysical and biophysical processes (forest and crop evapotranspiration, snowpack accumulation and melt, rainfall and snowmelt infiltration, groundwater interactions, etc.) and corresponding environmental parameters that control river runoff production for a given watershed. In contrast, data-driven models do not explicitly represent underlying physical processes, instead using empirical fits between inputs like snowpack, precipitation, and climate data, and outputs like seasonal river runoff volume. In operational practice, data-driven approaches consist of statistical models, typically multiple linear regression or principal components regression using heuristic prediction bounds based on out-of-sample (cross-validated) standard error as a measure of predictive error variance. At NRCS, the regression predictand is most commonly April-July aggregated river flow or reservoir inflow volume, forecasts begin to be issued in January or February and may continue into May or June, and predictor variates typically consist of mountain snowpack measurements taken just before the forecast is made along with a few other environmental variables; details vary between rivers and between forecast-issuing agencies, but the generalities are the same across the North American west. Alternative linear statistical regression techniques, such as M-regression and partial least squares regression, have been explored in the research literature but are close variations on the same theme. Memory-based (Box-Jenkins, ARIMAX, etc.) time series models are used across the physical, life, and social sciences for data-driven prediction and also have an important place in hydrologic and climate science, but in general they have not been found to be a good match to the specific problem of WSF in the American West; the yearly sampling interval of spring-summer flow volume time series normally exceeds the decorrelation timescale of streamflow data, and as previously noted (and discussed in further detail below), it is well-established that predictive skill here is mainly derived from regression upon springtime mountain snowpack data along with a few other, in most cases contemporaneously measured, environmental variables.

Overall operational WSF community experience has been that process simulation models provide valuable physical insights and diagnostics, and temporally high-frequency model products that can be useful for some applications, but that data-driven models typically are much cheaper in terms of both setup and operating costs, can be computationally more reliable and much faster, are more amenable to incorporating new predictor data types such as newly discovered climate indices, match or exceed the prediction performance of process simulation models, and provide more reasonable prediction uncertainty information (e.g., [11], [12]). For these reasons, they tend to be the most widespread type of operational WSF model; those organizations that run process simulation models usually also run data-driven models either officially or for supplemental information (e.g., California Department of Water Resources, National Weather Service Colorado River Basin River Forecast Center, BC Hydro) and several organizations run only data-driven WSFs operationally, even if process simulation models are available to them (e.g., US Bureau of Reclamation and US Army Corps of Engineers forecasts in the Columbia River Basin, NRCS, Alberta Environment).

Under this WSF categorization scheme, machine learning, that aspect of artificial intelligence concerned with detecting patterns in data and using these to make predictions, is also considered a data-driven approach. Machine learning was first assessed for hydrological applications, primarily the related but distinct, and much shorter-timescale, problem of flood forecasting in the 1990s [13], and research on the topic has abounded since then. Literature documenting machine learning for seasonal water supply forecasting is comparatively recent and sparse. Examples include neural networks and support vector machines [14], [15]. Broadly speaking, research community experience has been that machine learning solutions provide deterministic prediction quality as good as or better than both linear statistical and process simulation models.

Nevertheless, there has been a "glaring lack of" [16] migration of these seemingly promising research outcomes into mainstream operational hydrology, that is, into government agencies or companies having a responsibility, with some degree of associated accountability, to produce river forecasts on a routine basis for internal or external clients who

rely upon them to make real-world decisions having significant consequences. A few instances of successful adoption in operational roles for flood forecasting have been recently documented [17], but overall, water resource scientists and engineers are far more likely to use AI to navigate the most traffic-free route to work in the morning using a smartphone application than to apply it to the water resource prediction problems they face when they get there.

This general failure of machine learning to transition widely into practical hydrologic applications stems from several issues [16], [17]. Its black-box nature suggests a lack of interpretability or ability to ingest or respect existing knowledge of the underlying physics of the system being modeled. A lack of emphasis on addressing uncertainty has also been identified as a key limitation in environmental applications. Quantitative prediction uncertainty estimates are a core product of all modern water, weather, and climate prediction systems, for example, yet are not an integral part of many machine learning methods, which often tend to focus exclusively on obtaining the best possible deterministic prediction. In fact, some major current trends in machine learning, such as mining big data for predictive patterns using deep learning-based massive neural networks, for example, seem to be moving further away from the awareness of specific physical problem knowledge and statistical predictive error estimation required by applied scientists and engineers in problems like river forecasting. Further, government agencies have strong professional and organizational accountabilities around reliable generation of readily explainable river forecasts. This can lead to risk aversion around using unfamiliar approaches like machine learning to replace proven technologies that have, over years or decades, achieved buy-in with large and diverse public stakeholder communities. Only those machine learning-based solutions that work closely with experienced operational WSF professionals to address their central concerns and requirements are likely to be operationally adopted. Given the general lack of transitioning of machine learning into mainstream, non-academic, operational hydrometeorological prediction, a new approach seems to be required.

B. DESIGN CRITERIA

The new system was required to simultaneously meet several criteria:

(1) Improvements in forecast accuracy were expected.

(2) Improved potential for automation was desired.

Building and running models in the existing system is labor-intensive, can involve many subjective choices, and is not conducive to certain long-term NRCS institutional goals, such as more frequent issuing of forecasts. Though domain experts feel that no WSF system should be entirely hands-off, steps were desired to minimize unnecessary time expenditures and streamline model-building and operations.

(3) It was similarly necessary to retain the relatively low cost of a traditional statistical WSF solution, without resorting to extensive and expensive computational or staffing resources for modeling system development, implementation, and operation.

(4) The system had to address three known technical limitations with the existing PCR-based method: (a) nonlinear functional forms, and (b) heteroscedastic and (c) non-Gaussian residuals. Practical WSF experience shows it can involve nonlinear relationships between regression predictors (i.e., features) and the predictand (i.e., target), and linear approximations sacrifice predictive capacity and can occasionally even contribute to non-physical outcomes, such as negative-valued flow volume predictions. Additionally, prediction uncertainty is typically greatest when flow volumes are high, and are also often characterized by model errors that are not normally distributed around the best estimate. These attributes are not accommodated by most traditional statistical methods and instead require the capability to generate, when needed, prediction bounds that are asymmetric about the estimate and having a width that varies from year to year. Failure to do so can lead to misleading information about the confidence of the prediction and, in some cases, prediction intervals that contain non-physical negative-valued flows in dry years. Using predictand transforms prior to modeling is a standard statistical trick for applying linear stationary Gaussian techniques to nonlinear, nonstationary, non-Gaussian datasets and is used in the current NRCS system, but in practice it requires slow and subjective manual intervention and can lead to model relationships that are also non-physical.

(5) The resulting prediction system was expected to achieve an overall balance between visibly demonstrating innovation and performance improvements, while also being constructed from established building blocks using proven tools. For instance, on one hand the operational hydrology community and forecast product users would typically view the use of a completely new, cutting-edge artificial intelligence method in an operational WSF system as irresponsible and unprofessional, and it was also felt important to retain some broad design elements of the existing, successful, and broadly accepted NRCS linear statistical WSF system. On the other hand, innovation and progress were required; for example, the new system assimilates some recently introduced machine learning variants having specific properties needed for achieving some of the required design criteria, uses a novel framework for integrating off-the-shelf components, and where necessary employs certain statistical and algorithmic methods or developments that do not appear to have been widely reported in the hydrology or machine learning literature.

(6) A multi-method ensemble approach was required. Equifinality is a central issue in virtually all forms of environmental prediction, including water supply forecasting. Equifinality is a non-unique dependence of various model fit measures on model philosophy, structure, or parameter values, such that many (either subtly or significantly different) models perform similarly overall, leading to model selection uncertainty and undermining model interpretation and credibility. That is, while for a given river and prediction

performance measure, one model among several developed will always be ''best,'' the margins between models are often small, and which model is best typically varies between fit measures and specific examples. More generally, every modeling technique has advantages and limitations, complicating identification of any one approach as being the sole correct model for a given application. Multi-model ensembles are used in many fields to address this issue; such ensembles typically provide more robust and consistent prediction performance relative to the constituent models within the ensemble by blending the capabilities and damping the limitations of each, and they can be capable of outperforming the individual ensemble members through mutual error cancelation. A modular and expandable architecture was also desired, such that individual methods could be switched out, or a forecast generated externally, possibly even by another agency using completely different methods, could be ingested into the multi-method ensemble.

(7) Ability to accommodate both high-dimensional multicollinear predictor data and potential for multiple independent input signals was required. Predictive information for a WSF model at a given time step includes present and sometimes previous values of a range of geophysical variables. These datastreams have much redundant information (such as whether it is a high-snow or low-snow year) that can be compressed into a compact feature space, improving the reliability of linear predictive models and reducing the required complexity of a machine learning-based nonlinear regression model. However, the capacity to retain higher-order non-redundant information potentially contained in the input fields (like year-to-year carryover of water in large groundwater aquifers within a watershed) can also be vital in certain river basins for enhancing WSF accuracy.

(8) Integrating a measure of physics-awareness into the machine learning solution was important to establishing buy-in for the new system from the NRCS and its clients. Though machine learning is to some degree fundamentally black-box, steps can be taken to help ensure that outcomes are interpretable in terms of, or at least honor, certain key aspects of the known process physics. For example, there is a track record of using machine learning to discover underlying process physics in complex open environmental systems, such as identifying controls on the strength of the Indian Monsoon, understanding the origin of nonlinear memory processes in watershed dynamics, and discovering key controls on the formation and erosion of beaches in coastal engineering [18]–[20]. Of key concern here was allowing certain known properties of the physical problem to be enforced. In particular, the selection of particular ML methods having specific computational characteristics, like non-negativity or monotonicity, can be used in combination with an understanding of the physical background to the specific forecasting application at hand, like knowledge of the general relationships between snowpack and runoff and what functional forms are and are not reasonable for those relationships, to create machine learning solutions that are physically defensible.

C. PAPER ORGANIZATION

Section II summarizes the system developed in accordance with the multiple, interdisciplinary goals outlined in I.B above. Section III discusses some practical details of WSF and provides an example application of the new system. Section IV provides a summary, identifies directions for additional future work, and discusses some of the broader potential implications of the prediction system.

These implications extend both to WSF generally – an increasingly challenging and high-stakes problem given projections of a 55% increase in global water demand by mid-century, while nearly two billion are without adequate water even today – as well as to other, non-WSF, prediction tasks having similar or analogous technical and logistical requirements. Most of the design requirements laid out in I.B tend to be typical of geophysical and environmental prediction problems, and we expect that probabilistic nonlinear regression prediction tasks in other fields might also benefit from some of the developments presented here.

II. INTEGRATED MULTI-METHOD PREDICTION ENGINE

A. OVERALL FRAMEWORK

A probabilistic forecast model estimates the probability distribution of the future predictand, y, conditional upon the current values of the predictor vector, $P[y(n + \Delta t)]|X = X(n)$, where $X = \{x_1(n), x_2(n), \ldots, x_M(n)\}$ are M predictor time series and Δt is the forecast lead time. In practice, and particularly for continuous-valued predictands, the desired product often consists of a best estimate of y, taken to be some measure of central tendency (most often but not always the mean; see below) of P, and an associated estimate of the uncertainty in that best available prediction. We frame much of the following prediction system description in these terms for convenience of presentation and consistency with NRCS needs, but note that these methods can generate additional probabilistic forecasting products, like the probability that the predictand will exceed a threshold value or occupy a certain category.

The overall concept we use, illustrated in Fig. 1, contains several main elements: unsupervised statistical learning for extracting dominant features from high-dimensional input data, a multi-method core drawing on statistical and machine learning techniques for relating the extracted features to the predictand, and evolutionary methods for automated generation of optimal model suites, that is, input data and feature selections on a per-model basis. This overall system design directly reflects the way that the water resource science and engineering community frames and structures statistical WSF, which can be summarized as follows.

The job of a probabilistic WSF system is to provide a best-estimate prediction, with quantitative estimates of prediction

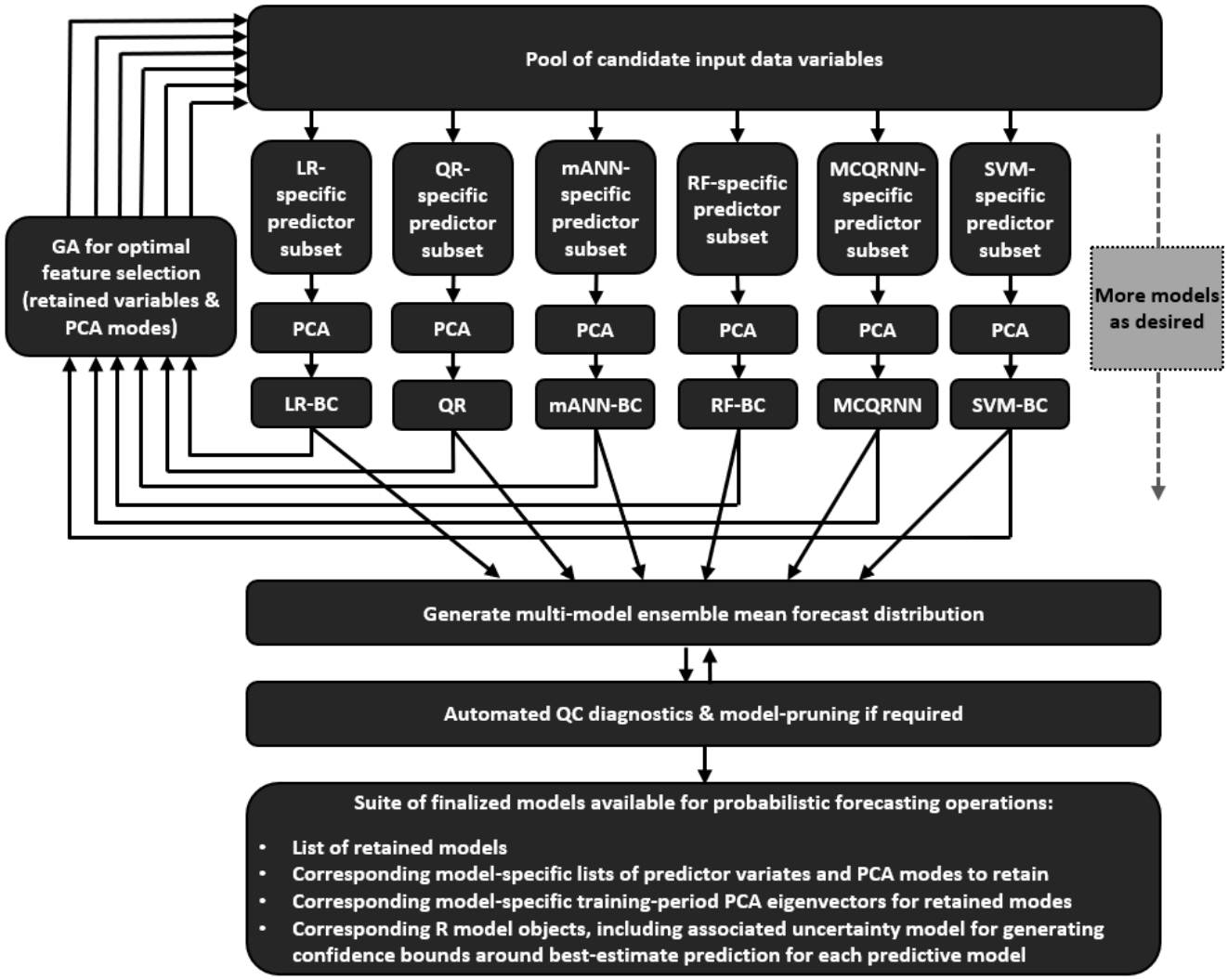


FIGURE 1. Approximate representation of some major components and workflows in the model-building step. Items listed in the bottom box are the main products of the model-building process and are subsequently used in forecast operations. Not all system components and workflows are illustrated here. Acronyms are as defined in the text. In summary, (i) the prediction metasystem starts with the unsupervised learning method of principal components analysis for extracting independent signals from the predictor dataset; (ii) six probabilistic regression and regression-like methods, carefully selected for specific capabilities from the large available number of advanced statistical modeling and supervised machine learning techniques, are each fitted separately to relate those extracted signals to the predictand, including algorithms for hyperparameter tuning as needed and partial parallelization across processor cores to improve computational efficiency; (iii) a genetic algorithm is wrapped around (i) and (ii) to optimize feature extraction and selection for each of the six regression and regression-like methods separately; (iv) when the six individual regressions are finalized as per (iii), a simple averaging process then combines the probabilistic predictions from each of them into a multi-method mean forecast distribution; and (v) a final quality control step is undertaken, mainly to help ensure the predictions obey certain physical constraints.

error, of spring-summer total runoff volume for a given location on a river of interest. Official forecasts are issued once per month, starting in February or earlier, and continuing through April or later, though more frequent updating is routinely undertaken. Typical WSF predictor variates can include snowpack, soil moisture, and precipitation data from each of multiple measurement locations within or near the watershed area. For the NRCS, these points are mostly SNOTEL sites, a network of remote high-elevation environmental measurement stations with automated data collection and telemetry (for additional detail, see Section III). Additional WSF predictor data can include soil moisture measurements, antecedent streamflow observations, indices of interannual climate variation such as El Niño-Southern Oscillation (ENSO), output from process-based snow models, and snow estimates from airborne or satellite remote sensing. Selection of specific candidate input variables for a WSF model for a given location and forecast date is necessarily based on hydrologist expertise, and includes factors like spatial proximity, incorporating redundancy through multiple partially correlated SNOTEL sites in the event of sensor or telemetry failures during forecast operations, capturing local-scale environmental heterogeneity across a watershed using a variety of SNOTEL sites, and many other considerations. Details vary considerably from watershed to watershed depending on local climatic and geologic characteristics.

For a given predictand, each forecast issue date has its own regression model and corresponding set of predictor variates, which evolve over the forecast season. For instance, the forecast made on 1 February of April-July flow volume for a certain river might include the 1 November to 31 January average ENSO index as one of its predictor variates, because for complex and incompletely understood reasons of global climate dynamics it serves as a proxy and therefore partial predictor of total seasonal snow accumulation through the end of winter, and therefore of spring-summer river flow volume. By April, however, the winter snow accumulation has typically ceased and direct measurements of it have obviously become available, so the forecast made on 1 April of April-July aggregated flow volume tends to drop large-scale climate indices as predictors, favoring 1 April snowpack observations instead. A given river therefore has a suite of separate models, one for each forecast issue date. Further, for a given forecast date and river, multiple target periods (April-July, May-June, etc.) may be considered as predictands, leading to additional models in some cases. That said, a combination of geophysical predictability and water management considerations are such that the 1 April forecast of April-July volume is, for most rivers, the cornerstone of the model suite.

To help illustrate, a realistic and common example of how a WSF regression (or regression-like) model is structured and used operationally in the US West would be a forecast, issued on 1 March 2019, of 1 April through 31 July 2019 total runoff volume at a certain location on a certain river of interest, using eight predictor variables: 1 March 2019 snow water equivalent measurements at eight SNOTEL sites to capture the most recent available mountain snowpack information across the upstream watershed area draining to that location on the river. Typically, the training dataset for a regression or regression-like model of this type would be roughly 30 samples, one for each of the same number of years. For a 1986-2015 training set, for instance, the first sample would consist of the observed 1 April 1986 to 31 July 1986 total flow volume (the single predictand); and 1 March 1986 measurements of mountain snowpack for each of the eight SNOTEL stations (the eight predictors). The second sample would be 1987 measurements of the same variables, and so on for the 28 subsequent samples. In linear regression modeling, these samples are used to estimate regression coefficients and, typically, assess the statistical significance of individual predictors; the linear PCR method currently used at NRCS and elsewhere additionally uses PCA pre-processing of the eight inputs (which would usually be highly correlated) to extract mutually uncorrelated signals used as the predictors in the linear regression. Further, a semi-quantitative combination of the statistical significance of each predictor in the regression under standard assumptions, a tree-based algorithm for selecting which of the input variables to retain, and qualitative hydrologist judgement and intervention are currently used at NRCS to find a quasi-optimal model (see Section III for more detail).

These basic facts of how the water resource science and engineering community performs WSF in the US West motivate the overall framework depicted in Fig. 1, which can be viewed as a modernized and upgraded version of the existing NRCS system (see Section III). From experience, a data-driven WSF system requires methods for addressing predictor multicollinearity, identifying multiple input signals with potential WSF predictive value, an objective means for identifying the most promising predictor variables from a pool of broadly reasonable candidates, and relating these to forthcoming water supply availability using a regression-like model. These tasks are performed here using a combination of an unsupervised learning algorithm for feature extraction, an evolutionary algorithm for feature selection, and a suite of regression models embedded within that semi-automated feature generation and selection framework that were chosen for specific characteristics known to be important from WSF experience, such as ability to handle nonlinearity and heteroscedastic or non-normal error distributions, as well as other logistical considerations, such as a proven track record, as described above in the system design criteria (Section I.B).

Modeling is split into model-building and model-operation phases. This is typical of traditional statistical prediction and many machine learning applications, and applied science and engineering models in general, but departs from some contemporary directions in machine learning for big-data applications. These phases are described here for clarity.

Model-building is a de facto inverse modeling problem: for a given WSF forecast task (that is, a certain combination of river and forecast date) the modeler selects, on the basis of hydrologic expertise, an input variable candidate pool consisting of a few decades of annually sampled data for certain geophysical variables at certain locations, and makes a few basic modeling choices; the prediction engine then forms an optimal (in some practical sense; see below) suite of regression models for the dataset and saves them with some associated modeling choices and performance metrics. The model-operation phase is a de facto forward model run using the optimized prediction engine: the saved modeling suite for a given problem is retrieved and then run using a new input data sample, corresponding to the observed values of the predictors that year. The small sample sizes (roughly 20-40 samples; see example application below), slow trickle of new data (one sample per year for each predictor and the predictand), and established WSF protocols around periodic recalibration once every few years including re-evaluation of candidate predictor choices based on practical considerations like measurement site suitability after wildfires, land use change, and so forth, are such that online sequential machine learning approaches, though useful for certain big-data settings including, potentially, some environmental applications [21], do not appear to offer significant value in a seasonal WSF context.

The system was implemented in the R scientific computing environment, chosen for its combination of diverse packages, ease of use, widespread and long-standing

adoption and therefore (it is hoped) robustness to obsolescence, and free, open-source status. The various components of the integrated prediction system (each machine learning method, for instance) consisted of existing and well-documented R packages, directly available for easy download and installation from a CRAN mirror site; these were tied together in custom R scripts. The specific R packages used are identified along with citations as they arise in the following discussion. Construction emphasized a modular and flexible framework into which new methods, or probabilistic prediction products from completely different external sources (such as physical process simulation models), can be integrated in the future if desired, leaving as many development and refinement options open as pragmatically possible.

B. FEATURE CREATION BY UNSUPERVISED LEARNING

The dimensionality and multicollinearity problem is addressed using principal component analysis (PCA) data pre-processing. PCA is a pattern recognition technique that compresses the information content of a large dataset into a series of mutually uncorrelated modes that efficiently concentrate the total dataset variance. A classical eigenanalysis method is employed here. Other matrix factorization approaches, such as singular value decomposition or non-negative matrix factorization, might also be used and could be explored for adoption in a WSF system in future work. However, PCA is by far the most widely known and proven of these techniques, and its track record in WSF applications leads to its selection here.

The length-N time series (corresponding to the training set; or combined training and testing set in a cross-validation framework) of each of M predictor variables is standardized to zero mean and unit variance and arranged into an M by N data matrix:

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1N} \\ \vdots & \ddots & \vdots \\ x_{M1} & \cdots & x_{MN} \end{bmatrix} \tag{1}$$

The M by M correlation matrix is then:

$$C = \frac{1}{N} X X^T \tag{2}$$

where X^T denotes the transpose of X. Eigenanalysis is performed on C, giving eigenvectors arranged in a M by M matrix, E, and corresponding eigenvalues arranged in a vector, λ:

$$E = \begin{bmatrix} e_{11} & \cdots & e_{1M} \\ \vdots & \ddots & \vdots \\ e_{M1} & \cdots & e_{MM} \end{bmatrix} \tag{3}$$

$$\lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_M \end{bmatrix} \tag{4}$$

The eigenvectors provide orthogonal basis functions, and the eigenvalues define the proportion of variance explained by each mode. The PCA scores are:

$$A = E^T X \tag{5}$$

where T again denotes the transpose such that each row of E^T contains an eigenvector and each column of E^T is indexed to one of the original variables in X, and A is a M by N matrix consisting of the projections of the original time series into the new coordinate system defined by the unit vectors in E:

$$A = \begin{bmatrix} a_{11} & \cdots & a_{1N} \\ \vdots & \ddots & \vdots \\ a_{M1} & \cdots & a_{MN} \end{bmatrix} \tag{6}$$

The principal component time series corresponding to each of the PCA modes are mutually uncorrelated and are used as candidate features in multi-method nonlinear regressions (see below), with an emphasis on the few leading modes which by construction explain the most variance within the predictor matrix.

C. MULTI-METHOD ENSEMBLE: CONCEPT

We address the model selection problem using multi-method ensembles. Model averaging can produce more accurate predictions than the individual ensemble members through mutual error cancellation and, in particular, leads to more consistently reliable predictions.

It has a strong pedigree across several distinct fields. For instance, machine learning examples include bagging and boosting, which guard against overtraining or combine multiple weak learners into a strong learner (e.g., [22]); the no-free-lunch theorem is also an expression of underlying model selection ambiguities [23]. Ensemble learning continues to be an active field of AI research (e.g., [24]–[28]). In statistics, multi-model inference involves addressing model selection problems with linear combinations of different but similarly-performing linear models, sometimes weighting constituent models using information theoretic or Bayesian criteria [29], [30]. In risk analysis, multiple probability density functions generated by different quantitative models or human expert opinions are routinely combined into a consensus distribution using linear opinion pooling [31]. In some areas of science and engineering, the underlying physics or its optimal explicit representation for a given scale and purpose of application can be ambiguous, leading to multiple plausible process simulation models, and the most accurate and consistent outcomes are generated by an ensemble mean across these [32].

A common thread across these fields is that the more diverse the models in the ensemble, the better. In machine learning, for instance, both random forests and bagged CART models are ensemble regression trees, but the decorrelated trees used in random forests are preferred. Similarly, in weather, water, and climate prediction, for example, if strongly methodologically distinct models are used (capturing different physics or combining physics-based and data-driven models) then only 3-6 models are needed to

realize the benefits of ensemble modeling [32], [33]. Though perhaps counter-intuitive, poorly performing ensemble members should be retained (up to a point) because even a poor performer can contribute valuable prediction information at certain timesteps corresponding to certain conditions that it may capture reasonably well, but that are not captured by other (and under most other circumstances, better-performing) models. In practice, equally weighted linear combinations often perform as well as or better than more complex approaches based on weighting by individual model performances, particularly when samples sizes are small [31]–[33].

Here, for each of R probabilistic supervised learning methods, the expectation value at time n, and prediction bounds around it expressed as desired quantiles of the forecast distribution, are averaged with corresponding outcomes from the other R-1 methods to collectively form the ensemble probabilistic prediction:

$$\begin{aligned} \langle E[y(n)]\rangle &= \frac{1}{R}\sum_{r=1}^{R} E[y(n)]_r, \\ \langle Q_{0.10}[y(n)]\rangle &= \frac{1}{R}\sum_{r=1}^{R} Q_{0.10}[y(n)]_r, \\ &\cdots \end{aligned} \tag{7}$$

where $Q_{0.10}[y(n)]$, for example, is the 10^{th} percentile of a probability density function having mode $E[y(n)]$. That is, the final probabilistic forecast is the average of the forecast probability density functions from the R individual probabilistic methods.

If one or more of the R individual regression or regression-like modeling methods is in turn the outcome of an ensemble learning process, such as a random forest or bagged neural network, then (7) produces an ensemble of ensembles. This is known in some fields (such as hurricane forecasting or climate change projections) as a super-ensemble. That is, the method entails, in part, a multi-level hierarchy of multi-model ensembles, with a low-level ensemble learner (e.g., a random forest or bagged neural network) being one of several predictors integrated into a higher-level multi-method ensemble.

By the same token, our multi-method approach is distinguished from some common ensemble techniques in machine learning, such as bagging and boosting using multiple slightly different versions of the same model (e.g., a bootstrapped regression tree), by the use of R strongly distinct classes of probabilistic regression and regression-like methods, as described below. This approach of using significantly different constituent regression methods to form a multi-method ensemble is again similar to, and inspired by, that used for both statistical and process simulation-based models in fields like weather and climate prediction, as described above.

To help ensure clarity in the following discussion, we generally use the term ''method'' to refer to one of the six regression and regression-like modeling techniques integrated into the multi-method ensemble, to help distinguish these from the isolated models (e.g., an isolated regression tree) that occur within those particular methods which are in turn ensemble learners (e.g., a random forest).

D. CONSTITUENT PROBABILISTIC REGRESSION/ REGRESSION-LIKE METHODS

For reasons described above, selection of specific methods for our current ($R = 6$) suite is guided by capacities to capture nonlinear relationships, to quantitatively represent forecast uncertainty through generation of prediction bounds accommodating heteroscedastic and non-Gaussian residuals, and/or to permit some level of physics-awareness by respecting basic problem constraints like monotonicity and non-negativity when desired. A track record of successful prior application in several fields, and preferably some prior experimental application to geophysical prediction problems, is also strongly valued.

A three-layer, feed-forward, error-backpropagation artificial neural network (ANN), that is, a multi-layer perceptron (MLP), was chosen for its track record as the most widely used supervised machine learning algorithm:

$$E[y(n)]_{ANN} = b^{(2)} + \sum_{j=1}^{J} w_j^{(2)} \left\{ h\left[\sum_{m=1}^{M'} w_{m,j}^{(1)} a_m(n) + b_j^{(1)} \right] \right\} \tag{8}$$

where $a_m(n)$ is the value of the m^{th} mode PCA time series at time n; $M' \leq M$ is the number of retained PCA modes; w and b are weights and biases, respectively; the superscripts (1) and (2) denote consecutive network layers; J is the number of hidden-layer neurons; and h is a nonlinear activation function, commonly tanh. Such an MLP is, in principle, a universal approximator.

Two MLP variants were selected. The monotone artificial neural network (mANN) offers a user-selectable option of enforcing a monotonicity constraint for specified predictors ($r = 1$). This contributes to regularization and enables use of, and ensures compliance with, the underlying physics of certain problems where the relationship between the m^{th} candidate predictor and the predictand is known to be (potentially nonlinear but) monotonically non-increasing or non-decreasing:

$$\frac{\partial E[y(n)]}{\partial a_m(n)} \geq 0 \,\forall\, n \quad or \quad \frac{\partial E[y(n)]}{\partial a_m(n)} \leq 0 \,\forall\, n \tag{9}$$

This is accomplished by exponentiating the neural network weights [34]. mANN was implemented using `monmlp` [35], which employs a quasi-Newton (Broyden-Fletcher-Goldfarb-Shanno) algorithm for supervised training; additional regularization options included are bagging and stopped training. We employ a heuristic post-processed approach in Box-Cox transform space [36] to generate associated prediction bounds

accommodating non-Gaussian and heteroscedastic residuals:

$$y(n)^{(\psi)} = \begin{cases} \dfrac{y(n)^{\psi} - 1}{\psi}, & \psi \neq 0 \\ \ln[y(n)], & \psi = 0 \end{cases} \tag{10}$$

where ψ is a parameter and $y(n)^{(\psi)}$ denotes the Box-Cox transform of y at time, n. Both the observed and mANN-predicted time series of y are transformed; the predictions used in this process are obtained using k-fold cross-validation to better capture out-of-sample forecast accuracy. These two transformed datasets are then differenced to obtain the residual time series, which is normally distributed in Box-Cox transform space. The root mean square error (RMSE) can therefore be used as a convenient approximate metric of the standard deviation of the transform-space residuals. The transform-space α^{th} quantile forecast is estimated as:

$$Q_{\alpha}[y(n)]^{(\psi)} = E[y(n)]^{(\psi)} + z\left(RMSE_{CV}^{(\psi)}\right) \tag{11}$$

where z are corresponding z-scores under the normal distribution. An inverse Box-Cox transform is applied to the result to obtain final estimates of the α^{th} quantile prediction bounds. Forward and inverse Box-Cox transforms were performed using the `forecast` package [37], which also finds optimal ψ. Note that the one-parameter Box-Cox transform uses, and returns, only positive values; thus, if the best-estimate prediction is strictly positive-valued, the prediction bounds around it determined using (10) and (11) will also be positive-valued, a desired characteristic of a WSF system.

The second MLP ($r = 2$) is a monotone composite quantile regression neural network (MCQRNN) [38]. This method incorporates nonlinear quantile regression, such that both $E[y(n)]$ and $Q_{\alpha}[y(n)]$ for user-specified quantiles of the forecast distribution are directly generated. We take $E[y(n)]$ to be $Q_{0.50}[y(n)]$ so that MCQRNN is a form of rank-based (median) regression, giving a best-estimate prediction that is comparatively robust to outliers. The technique ensures non-crossing quantiles, an occasional issue for small sample sizes of noisy data. In addition to allowing enforcement of monotonicity constraints as in (9), the forecast distribution can also be forced to be non-negative, further contributing to guaranteed physical plausibility of outcomes in certain common applications including WSF:

$$E[y(n)] \geq 0 \,\forall\, n \quad \textit{and}\ Q_{\alpha}[y(n)] \geq 0 \,\forall\, \alpha,\, n \tag{12}$$

The method is implemented using `qrnn` [38], which employs a Newton-type training algorithm and weight penalty regularization. While both MLPs fit a model of the general form of (8), in practice the final models and their predictions are strongly distinct.

Random forests (RF) was selected ($r = 3$) due to its widespread acceptance as a contemporary nonlinear machine learning algorithm; its fundamental difference from MLPs, contributing to a more methodologically diverse ensemble of models; its relative ease and robustness of application to a wide range of problems; and reduced (relative to some other methods) model development and implementation complications, such as overtraining or extensive manual hyperparameter tuning:

$$E[y(n)]_{RF} = \frac{1}{L}\sum_{l=1}^{L} < y_l(n) >,$$
$$< y_l(n) > = d_l\left[a_{m=1,M'}(n)\right] \tag{13}$$

where $< y_l >$ here denotes the prediction from one of L regression trees and the best estimate is an ensemble of these. A tree is formed by recursive binary partitioning without pruning, leading to a number of terminal leaves, each corresponding to the mean value of the response variable over some disjoint subset of the predictors. Each subset is defined by a predictor function, d, so as to minimize residual sum of squares at each decision point. An ensemble of mutually decorrelated trees is generated through random selection of both the retained samples (bagging) and the retained explanatory variates. RF was implemented using `randomForest` [39], and prediction bounds were obtained using the post-processed Box-Cox transform-space heuristic described above for mANN.

Whereas ANN and RF can be viewed as soft-computing methods that emulate the information processing capabilities of biological or social processes (ANN: the brain's network of neurons and synapses; RF: the intuitive decision-tree model of human choice), the support vector machine (SVM; $r = 4$), also a widely successful machine learning technique, is very different and therefore adds further to the desired methodological diversity of the model ensemble. It is based on abstract geometric constructs, in particular an ε-insensitive loss function and a kernel function taken here to be a radial basis function, that turn relatively low-dimensional nonlinear regression problems into high-dimensional linearly separable classification problems (e.g., [40]) solved by fitting the hyperplane:

$$\mathbf{u} \cdot \mathbf{v} - \mathrm{b} = 0 \tag{14}$$

that maximizes the margin between itself and the nearest observations, which constitute its namesake support vectors, where $\mathbf{v}$ is a set of predictor vectors built from the original features through the kernel function and $\mathbf{u}$ is a set of weights that defines the normal vector to the hyperplane. We used `e1071` [41]. Cross-validated prediction bounds were estimated using a Box-Cox transform-space heuristic.

Nonlinear relationships are one of our central concerns here, yet linear methods have persisted in data-driven modeling because typically they are tractable, easily interpreted, and often provide serviceable approximations. We therefore additionally include two linear techniques, while acknowledging that in some applications they can be omitted on the basis of a priori problem-specific physical information (known strongly nonlinear relationships) or a posteriori

performance assessment (see pruning step below):

$$E[y(n)] = \beta_0 + \sum_{m=1}^{M'} \beta_m a_m(n) \tag{15}$$

where β are model coefficients. Linear quantile regression ($r = 5$) is a nonparametric (distribution-free) technique that accommodates heteroscedastic and non-Gaussian errors, and is robust to outliers due to its use of the median as the best estimate, $E[y(n)]_{QR}$ [42]. Quantile regression was implemented using `quantregGrowth` [43], which ensures non-crossing quantiles. It is somewhat akin to MCQRNN (see above), but of course the structural model form is fundamentally different, coefficients are fit for each quantile by linear programing rather than by nonlinear optimization, and the quantile lines are determined sequentially rather than simultaneously [38]. Finally, conventional linear regression (which in combination with PCA pre-processing amounts to principal components regression, PCR) was included ($r = 6$) due to its central role in the established theory of linear statistical modeling and more than a century of widespread application across all fields of science and engineering, including extensive use in WSF. $E[y(n)]_{LR-BC}$ is taken to be the mean value of the predictand conditional on the current values of the predictor variates, and corresponding regression parameters are estimated by least squares. A common heuristic estimate of linear regression prediction bounds is provided by (11) but in non-transform space, consistent with the standard regression assumption of Gaussian homoscedastic residuals. As a serviceable back-of-the-envelope approach to modifying standard linear regression to generate non-Gaussian heteroscedastic prediction bounds, we instead apply the post-processed Box-Cox transform space-based approach we use here for mANN, RF, and SVM. Experimentation suggested this approach is generally effective at producing time-variant and asymmetric prediction bounds when needed, but also automatically reduces to near-Gaussian homoscedastic bounds when appropriate, rendering an otherwise conventional LR (or in effect, PCR) more flexible. Similarly, experimentation demonstrated that all the nonlinear machine learning methods captured linear relationships when appropriate.

E. OPTIMAL FEATURE SELECTION USING EVOLUTIONARY COMPUTING

Following the lead established by computational statistics, biology, and economics, a genetic algorithm (GA) is used here to solve the NP-hard [44] combinatorial optimization problem of optimal predictor selection [45]–[47], which in our application requires simultaneously selecting input datasets from a pool of candidates and, for a given trial input dataset, the corresponding PCA modes to retain. This approach avoids restrictive and often unrealistic statistical (e.g., distributional) assumptions for assessing the significance of individual candidate predictors in linear statistical models, it is suitable for application to machine learning methods that do not have clear parametric tests for judging which inputs are significant, and it combines the selection of candidate input variables and PCA modes into a single unified step. This GA-based feature selection is done separately for each model, as experimentation showed each technique could prefer its own optimal combination of both input variables and retained PCA modes. Evolutionary fitness of a trial solution is judged by its (arithmetic-space) cross-validated RMSE. For computational efficiency, we restricted candidate PCA modes considered by the GA to a user-selectable, not necessarily consecutive [48], subset, starting with the mode explaining the most variance. The `genalg` package was invoked for GA implementation [49], [50] and uses the basic genetic operators of elitism, single-point crossover with roulette-wheel mating pair selection, and mutation; the `rbga.bin` functionality was employed, corresponding to binary discrete optimization (switching genes corresponding to individual candidate input variables and PCA modes on or off). The final gene sequence for a given model (e.g., mANN) encodes the optimal feature extraction and selection.

F. MODEL OUTPUT AGGREGATION AND PRACTICAL QUALITY CONTROL

After probabilistic predictions from the $R = 6$ optimized individual regression methods are aggregated using (7), the final step in our framework tests the solution for key criteria and adjusts the ensemble composition if needed. Model selection uncertainty is such that while some regression techniques, out of all those conceivably available, might be ruled out a priori (strangers), it is difficult to uniquely determine which subset potentially appropriate to a problem (the family) is the best (family we like). Thus, we introduce a quality control (QC) process in which we invite the family (the short-listed ensemble of regression and regression-like methods described in the foregoing section) to Thanksgiving dinner, and only if absolutely necessary, kick out relatives who misbehave.

For a given application, we might define some inadmissible behaviors on a per-method basis, such as sub-par performance for method r', judged by one or more metrics, relative to the other methods individually and/or collectively, e.g.:

$$\left.\begin{aligned} RMSE_{r=r'} &> \frac{1}{R-1}\sum_{r\neq r'} RMSE_r + \delta, \\ AIC_{r=r'} &> \max\left(AIC_r\right) + \epsilon \;\forall\; r \neq r' \end{aligned}\right\} : remove\ r'$$

$$\left.\begin{aligned} RMSE_{r=r'} &\leq \frac{1}{R-1}\sum_{r\neq r'} RMSE_r + \delta, \\ AIC_{r=r'} &\leq \max\left(AIC_r\right) + \epsilon \;\forall\; r \neq r' \end{aligned}\right\} : keep\ r' \tag{16}$$

where $\delta, \in$ are tolerances, and AIC is the Akaike information criterion. (Opinions vary around AIC-type measures for nonlinear modeling where degrees of freedom do not exactly correspond to the number of model parameters [51]–[53]; obviously, one may substitute other metrics if preferred). However, all models are flawed, particularly in real-world applications to complex systems; and even models having

a poor value for some summary skill metric can contribute useful predictive information in certain cases (we sometimes see a model that is generally mediocre but outperforms other ensemble members for timesteps when the predictand takes on, for example, an extreme value).

So, individually problematic behavior from a given method is not necessarily cause for removal, and another approach is to assess the ensemble mean forecast distribution and determine whether this end product meets criteria of interest. If it does not, we iteratively step through the ensemble members, pruning the gravest contributor to the problem one at a time until the corresponding multi-method ensemble is satisfactory. Any test thought important to the specific application can be used, such as consistently plausible behavior, e.g., a strictly non-negative forecast distribution for an application where a negative-valued predictand is physically impossible:

$$\begin{aligned} &min\left[\langle Q_\alpha[y(n)]\rangle\ \forall n, \alpha\right] < 0 : \textit{prune ensemble} \\ &min\left[\langle Q_\alpha[y(n)]\rangle\ \forall n, \alpha\right] \geq 0 : \textit{accept ensemble} \end{aligned} \quad (17)$$

where α again denotes the set of specified quantiles, Q, of the forecast distribution, P, for which results are desired and $< >$ is the ensemble mean across regression methods. Overall, this method-trimming philosophy is more realistic than expecting a fixed subset of methods to perform well for all forecast problems, and it is easily automated.

Note that each regression method (monotone composite quantile regression neural network, monotone artificial neural network, random forests, support vector machine, quantile regression with a non-crossing constraint, and linear regression), together with its method-specific feature extraction (principal components analysis) and feature selection (genetic algorithm) steps and predictive bound estimation, constitutes a forecasting system. In combination, and integrated with an ensemble generation and correction process, they form a modular prediction metasystem.

G. HYPERPARAMETER TUNING

The performance of machine learning methods can be sensitive to hyperparameter values. Preliminary system testing suggested that for our WSF application, the most crucial choices are around network topology (mANN and MCQRNN); neural network bootstrapping (mANN); ε, γ, and C in SVM; and the number of generations and the population size in the genetic algorithm.

Experimentation demonstrated that a single hidden layer with one neuron and no bootstrapping, or two neurons (without bootstrapping, MCQRNN; with bootstrapping, mANN) were sufficient, depending on the particular river. Note that after PCA pre-processing, our WSF problem becomes very low-dimensional (one predictand and therefore one MLP output-layer neuron; one to four predictors and a corresponding number of MLP input-layer neurons). More complex topologies did not provide consistently better out-of-sample performance, unnecessarily complicated the training process, and seemed slightly more susceptible to overtraining. A pragmatic algorithm for automated MLP configuration selection could therefore be implemented. The default for given set of predictors is a parsimonious and computationally fast single-neuron, no-bootstrapping configuration. This is tested by a criterion similar to (16). That is, if RMSE or R^2 for this MLP performs significantly worse than the mean RMSE or R^2 across all the non-neural network methods, an alternative configuration with two hidden-layer neurons is fitted; for mANN, this also includes bootstrapping (10 bootstraps were found to be sufficient; we wish to keep this number as low as possible to mitigate run times). The maximum allowable percentage performance deficit relative to the remainder of the models is a user-selectable run control parameter; 25% was found to work sufficiently well in this application. The alternative configuration is kept if either of two conditions are met: (a) its performance deficit is within this specified tolerance; or (b) it provides a lower AIC than the default configuration. Otherwise, the algorithm reverts to the default MLP configuration. If the default configuration meets the performance deficit criterion, no alternative configuration is considered. The user is also free to manually select all MLP hyperparameters, but these two basic configurations and the automated procedure for choosing between them proved satisfactory for our application.

SVM performance was found to be sensitive to some key hyperparameters, so we used the `tune` functionality in the `e1071` package to perform a simple grid search to find best values for a given predictor set. Initial experimentation showed that allowing γ to be determined in this way seemed to lead to overtraining. Manual experimentation suggested a value of about 0.2 provided a good balance, in our application, between allowing sufficient nonlinearity to capture the underlying nonlinear functional forms commonly present in the relationships between predictors and predictands in WSF (see Section III) while minimizing any tendency to memorize the data. The grid search was used to optimize ε and C. The search is inefficient, so removing γ from consideration also improves run times. Users may also deploy automated tuning of all three of these major SVM hyperparameters, or set all SVM hyperparameters manually, if preferred.

The computational scale of the GA problem is largely determined by the population size and the number of generations. Systematic testing was undertaken to track various model performance metrics as a function of GA problem scale. Results differed slightly between test cases, with the larger search spaces associated with larger input variable candidate pools unsurprisingly benefitting from the additional refinements potentially derived from more exhaustive searches using larger population sizes and longer runs, but broadly speaking the results were fairly uniform. Specifically, even the most rudimentary GA optimization (population size of 10 with only 5 generations) provided a large gain in prediction quality relative to no optimization of feature creation and selection; and a population size of 15 to 25, with 7 to 10 generations, provided significantly better results but also marked a point of diminishing returns, with prediction quality

plateauing at a population size of about 25 to 50 with 10 to 15 generations. There was no consistent benefit to using larger GA problem scales (testing considered population sizes as large as 400 and up to 25 generations). Note that the run time cost of larger GA problem scales is large. We therefore selected a population size of 15 with 7 generations as the default for our application, though the user can make alternative choices as desired.

Hyperparameters other than those discussed above were, in general, left at the default values in the corresponding R packages as they either seemed to perform satisfactorily, or had little effect, in our application. For example, the number of trees in a regression forest strongly impacts accuracy in some applications, but testing revealed that departures from the default yielded little consistent effect here.

H. PARALLELIZATION

The nonlinear optimization problem of ANN training, in combination with other iterative model-building procedures (bagging, cross-validation, predictor optimization), initially gave somewhat long run times in some applications. The embarrassingly parallel task of mANN and MCQRNN cross-validation was therefore distributed across multiple processor cores using `foreach` and `doParallel` constructs [54], [55]. The significant efficiency gains obtained were adequate to our present purposes, but several additional parts of the model-building cycle described above are clearly amenable to parallelization and could capitalize on, for instance, distributed cloud computing resources. Conversely, operational forecasting – that is, running a previously completed and archived set of model objects (lowermost box in Fig. 1) using a new (current) sample of the input data vector – is extremely quick under any reasonable computing environment.

III. APPLICATION

The prediction engine has been successfully applied to several test cases drawn from the NRCS WSF domain. Ultimately, hundreds of such applications will be made. We provide three illustrative examples below that demonstrate some of the issues faced in WSF and how the prediction engine addresses them. Detailed results differ from river to river, but the overall conclusions are similar.

A. TEST CASES

The Gila River is a tributary to the Lower Colorado River. The continental divide forms its eastern watershed boundary and separates the Gila and Rio Grande basins. At the upstream location considered here, it is an arid mountain river draining the Mogollon, Pinos Altos, and Black Ranges of southwestern New Mexico. The Upper Gila is relatively pristine, but downstream its waters are heavily diverted for agricultural and municipal water supplies and its flows are supplemented using Colorado River water through the Central Arizona Project. The Deschutes River is a tributary to the Columbia River. It flows from the moist crest of the Cascade Range in central Oregon. Dams and diversions on the Deschutes provide agricultural and municipal water supplies and hydropower generation, and the river has significant recreational and tourism values. It is a geophysically unusual basin insofar as the extensive volcanic aquifers of the Cascades result in close coupling of groundwater and surface water, yielding seasonally stable flows. The Owyhee River is a remote semi-arid watershed with headwaters in northern Nevada that also flows through southern Idaho and southeast Oregon and empties into the Snake River. The US Bureau of Reclamation operates a dam on the Owyhee to provide irrigation water for regional agriculture. Specific existing NRCS forecast points on these rivers, which correspond to US Geological Survey streamgage locations (see below), considered here are the Gila River near Gila, Owyhee River near Rome, and Deschutes River below Snow Creek. In these examples we consider the yearly 1 April forecast of 1 April-31 May (Gila) or 1 April-31 July (Owyhee, Deschutes) flow volume.

The existing US Department of Agriculture operational WSF modeling approach, which has also subsequently been adopted by a variety of other organizations in the US and Canada, uses PCR as noted previously. The NRCS PCR model-building procedure involves a tree-based search approach to prioritizing input variables for inclusion in the model, beginning with a one-variable model and progressively adding more variables, in various combinations, in new models until the standard error no longer improves. Choices around the number of PCA modes to retain for a given set of input variables are guided by starting with a linear regression model using only the leading PCA time series as a predictor, and sequentially including higher PCA modes until the additional predictor is no longer statistically significant using a t-test under the standard assumptions at, usually, $p < 0.10$. Prediction uncertainty quantification assumes a stationary normal distribution with a standard deviation equal to the leave-one-out cross-validated standard error of prediction, centered at the regression prediction.

Issues with this NRCS PCR approach for the Gila River include heteroscedastic and non-Gaussian regression residuals and a nonlinear functional form; that is, several of the central assumptions made in a linear regression are not met. The approach used to address these issues in the official NRCS PCR model for this location is to apply a cube-root predictand transform prior to modeling (other options available in the existing system, which is termed VIPER, include log and square-root transforms). There is no physical basis in hydrology for a cube-root transform, however, and its selection over other commonly used transform types is subjective and somewhat arbitrary. Improvements in prediction accuracy were also desired. The Owyhee River is similarly known from NRCS experience to be problematic as a result of nonlinearity and heteroscedasticity. The Deschutes River, in contrast, obeys the assumptions of linear Gaussian statistical modeling and the existing NRCS PCR model performs well. The NRCS Owyhee and Deschutes models do not employ transforms.

Outcomes from these NRCS PCR (VIPER) models form the baseline against which the prediction engine is compared. While this choice reflects the needs and priorities of the NRCS insofar as the new prediction system must meet or beat the predictive accuracy of the existing system, more generally the PCR model additionally provides a classical linear statistical modeling benchmark for evaluating the performance of the integrated, machine leaning-based, multi-method nonlinear regression metasystem. The NRCS model was refit to the same 1986-2015 historical period (see below) as the prediction engine to help ensure a reasonably apples-to-apples comparison.

Note that the broad framework of this existing NRCS approach, with PCA pre-processing followed by linear regression modeling, input variable selection via a search algorithm, and testing for which PCA modes to include as predictors for a given trial set of input variables, bears similarities to the new prediction engine; this is intentional and reflects some of the fundamental design criteria and overall requirements listed in sections I.B and II.A. That is, the machine learning-based prediction engine can be viewed as a modernized and upgraded version of the existing, proven, and widely accepted operational forecasting model, in which newer predictive analytics methods are judiciously applied to address some limitations of the current NRCS system.

B. PROBLEM SETUP

A 30-year model development period was employed, corresponding to the standard climatic ''normal period'' typically used by the weather and climate communities to calculate mean climate conditions. About 20-40 years of data is a common choice for hydrologic model-fitting, as it provides a reasonable balance among competing considerations. Using longer (50-100 year) records would of course provide larger sample sizes and capture a wider variety of hydroclimatic extremes. On the other hand, longer records would be more likely to capture gradual land use and climate changes that create nonstationarities which could undermine the reliability of model fits intended for the shorter-timescale problem of forecasting one or two seasons ahead, and it could dramatically reduce the number of available input data locations, given that many environmental (e.g., weather, snow, and streamflow) monitoring sites were established relatively recently. The dataset for prediction engine development therefore consists of 30 samples, one per year (see also Section II.A).

The experimental data considered here were obtained from large accumulated centralized databases of routine, long-term, ongoing environmental and natural resource monitoring programs conducted by the NRCS and US Geological Survey, and which currently serve as the basis for NRCS operational WSF. NRCS SNOTEL sites measure SWE using snow pillows, fluid-filled bladders with pressure transducers that weigh the overlying snowpack. SNOTEL stations also have weather stations monitoring precipitation and temperature; some stations have enhanced configurations with sensors for additional environmental variables, such as soil moisture. Data are telemetered to central offices by meteor-burst radio transmission, cellular modems, or satellite, and are then quality-controlled and archived. SNOTEL sites are often very remote and difficult to access, and power is provided by solar cells. NRCS snow surveys, in contrast, involve periodic field visits by ski, snowmobile, or helicopter to a fixed monitoring location, called a snow course, by NRCS staff to manually measure snow depth and density. Streamflow data are collected by the US Geological Survey at streamgages, which in most cases measure water depth using a pressure transducer and combine that information with bathymetry and periodic manual water velocity measurements to find volumetric flow rate; in some cases, they are instead local reservoir inflow volumes that are back-calculated from dam operation information, reservoir surface elevation recordings, and other information. Data were obtained from the freely available online NRCS database, wcc.sc.egov.usda.gov/reportGenerator.

The following input variable candidate pools were determined on the basis of NRCS SNOTEL operations and operational WSF experience, and include concerns like monitoring station proximity, record length and continuity, and reliability, as well as a variety of geophysical considerations as discussed earlier in this paper. For the Gila River, 1 April snow water equivalent (SWE; the amount of water that would be released from the measured snowpack given its observed depth and density) and 1 November through 31 March accumulated precipitation at three SNOTEL sites (SNOTEL station names: Lookout Mountain, Signal Peak, Silver Creek Divide) were selected as potential predictors, forming a candidate pool of six input variables. Similarly, the candidate pool for the Owyhee River was 1 April SWE and 1 November-31 March accumulated precipitation at each of eight SNOTEL stations (Big Bend, Buckskin Lower, Fawn Creek, Granite Peak, Jack Creek Upper, Laurel Draw, Mud Flat, South Mountain), as well as SNOTEL April 1 SWE at Taylor Canyon and SNOTEL 1 November-31 March accumulated precipitation at Jacks Peak, giving a total of 18 available input variables. Potential predictors for the Deschutes River were 1 April SWE and 1 November through 31 March accumulated precipitation at each of two SNOTEL sites (Irish Taylor and Three Creeks Meadow), SWE from a manual snow survey site (Tangent), and its own 1-31 March average flow rate at a specific location known to serve as an indirect but useful index of the aforementioned aquifer storage effects (Deschutes River at Benham Falls), giving a total of six candidate input variables.

The GA selected the optimal combination of input variables from the candidate pool to retain; the leading PCA mode was used without exception, and the GA was given the option of selecting whether to retain higher PCA modes up to the second mode. In WSF applications of this type, the leading PCA mode is known to capture variability in basin-wide overall wintertime precipitation and snowpack levels and is therefore the primary predictor. Higher modes capture either more

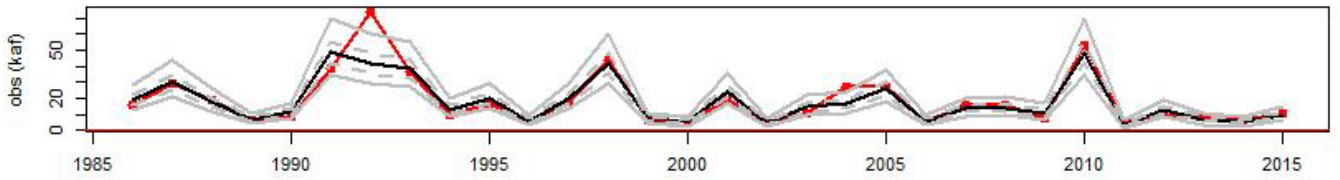


FIGURE 2. Observations (red dashed line with dots), best-estimate predictions (black line), and associated 0.10 and 0.90 quantile (solid gray lines) and 0.30 and 0.70 quantile (dashed gray lines) prediction intervals for spring-summer flow volume in kaf of the Gila River. Horizontal red line denotes zero flow.

subtle patterns of spatiotemporal variability in snowpack and precipitation, or in the case of the Deschutes, aquifer-stream interactions, and may or may not be retained for predictive purposes depending on basin-specific circumstances.

Though fine details of the underlying processes are complex, and accurate water supply forecasts can be difficult to produce for some basins, particularly those in the US desert southwest like the Gila River, general a priori knowledge of the underlying system physics is nevertheless significant. This process understanding leads to some key points for machine learning model setup. As one example, we know what are and are not potential controls on spring-summer runoff volume, and this information is reflected in input variable selection; this is not a data-mining exercise. Another example is that the relationship between the dominant PCA mode and runoff is known to often be nonlinear; this is an expression of hydrologic processes reflecting varying contributions of seasonal snowpack or precipitation vs. internal watershed storage (lakes, soil moisture, aquifers, wetlands, etc) during wet vs. dry years. This relationship is also known to be monotonic: for instance, low snowpack never gives high water supply, all else being equal. Additionally, river flow volume cannot be negative. Available monotonicity (MCQRNN and mANN) and non-negativity (MCQRNN; model pruning procedure of (17)) constraints were therefore invoked. Consequently, the resulting solution respects, and is in turn explainable in terms consistent with, certain key aspects of physical process knowledge. This property helps ensure physical defensibility of the results – found from experience to be essential for credibility with operational hydrologists and forecast product consumers. These constraints on the available solution space were also found to have a noticeable regularization effect on the machine learning solutions, reducing overfitting when invoked, consistent with expectations [34].

Some additional technical notes are as follows. Cross-validation was performed using $k = 1$, i.e., leave-one-out cross-validation (LOOCV), as partial autocorrelation functions of residuals were not significantly different from zero. NRCS service delivery obligations require forecasts framed as a best estimate with associated 0.10, 0.30, 0.70, and 0.90 quantiles. Results discussed below are for the default GA settings we determined for WSF problems in testing (see Section II.G), i.e., a population size of 15 with 7 generations, which required only 15-20 minutes to run on a typical current general consumer-grade four-core PC using partial parallelization (Section II.H). Forward simulation for a new sample in forecasting operations using the saved modeling suite is essentially instantaneous.

C. RESULTS

Table I provides five key performance metrics for the integrated multi-method ensemble mean prediction for this test case. LOOCV RMSE (see Sections II.D and III.B) is an approximate but instructive measure of typical prediction error. Coefficient of determination, R^2, is the square of the Pearson product-moment linear correlation coefficient between the observations and predictions, and it provides the proportion of predictand variance explained by the model. RMSE and R^2 are common measures of deterministic prediction accuracy, i.e., verifying the expectation value of the forecast distribution. Assessing the accuracy of probabilistic prediction information involves, directly or indirectly, verifying higher-order statistics, which can be challenging for modest sample sizes compared to verifying the mean, but we found two approaches in tandem provided a useful view on the quality of the prediction bounds and underlying probability models. One is a quantitative metric, the ranked probability skill score (RPSS). RPSS originated in the weather forecasting community and has spread to other fields including WSF. It rewards both the ability to forecast which category the flow will fall into (correct expectation value of the forecast distribution) and to do so with confidence (narrow prediction bounds about that best estimate); incorrect best-estimate predictions, and forecast intervals that are so wide that the correct value is almost inevitably included no matter what best-estimate prediction is made, are penalized. Following typical geophysical practice, three categories are defined for RPSS evaluation, and the terciles of the observed predictand are used as the category cutoffs. Another approach to assessing the reasonableness of the prediction intervals as well as the overall fit is a qualitative check, i.e., visual comparison of the observations, best-estimate predictions, and prediction intervals; an example is provided for the Gila River in Fig. 2. Another crucial requirement of the system is that it generates physically plausible estimates without user intervention, in particular, that it does not produce negative-valued flow predictions within the relevant state space. Table 1 therefore also indicates whether the

best estimate (BE) or the lowest (0.10 quantile) prediction bound (PB) < 0 for any of the samples.

Table 1 additionally provides this set of performance metrics for the individual regression methods within the multi-method ensemble, and for the existing official NRCS PCR WSF models. As noted previously, outcomes from the NRCS PCR models (the VIPER results in Table 1) form the baseline against which the prediction engine is compared. There are two reasons for this choice: NRCS requires the new prediction system to meet or beat the predictive accuracy of the existing system – and more generally, the principal components regression model, which is merely linear regression preceded by principal components analysis, provides a meaningful classical linear statistical modeling benchmark, particularly given its use for WSF at a variety of institutions across western North America as discussed above, against which to evaluate the performance of our forecasting metasystem.

In the interest of conciseness, we focus our performance assessment and interpretation on seven central outcomes.

1) LR COMPONENT PERFORMANCE RELATIVE TO LINEAR STATISTICAL MODELING BENCHMARK OF EXISTING NRCS WSF SYSTEM

An interesting starting point for comparing the prediction engine against the PCR benchmark is the linear regression that forms one of the six regression and regression-like methods within the multi-method metasystem. As the prediction engine includes PCA pre-processing, the LR component of it amounts to PCR, much like the VIPER modeling environment used classically by NRCS and other water supply forecast agencies.

Importantly, however, the LR component within our multi-method ensemble contains two major updates relative to VIPER: the use of a post-processed Box-Cox transform space-based heuristic for generating prediction intervals that can, when needed, seamlessly accommodate heteroscedastic and non-Gaussian residuals, i.e., produce asymmetric and time-varying prediction bounds; and the use of evolutionary optimization for feature selection. These two changes seem to provide a clear performance advantage of LR over classical PCR (Table 1), providing similar or slightly better deterministic (R^2, RMSE) and probabilistic (RPSS) prediction quality across all three rivers. As expected, however, it does not help with the inability of PCR to accommodate significantly nonlinear relationships, as seen for Owyhee.

2) OTHER CONSTITUENT MODEL PERFORMANCES RELATIVE TO LINEAR STATISTICAL MODELING BENCHMARK OF EXISTING NRCS WSF SYSTEM

The integrated multi-method prediction engine contains a total of six individual probabilistic regression and regression-like models, each independently combined with PCA-based feature creation and genetic algorithm-based feature selection. As noted above in point (1), certain attributes of this metasystem allow even an otherwise conventional linear regression to turn in better performances than a standard PCR approach. How effective is this framework when used with the five other methods, spanning nonparametric statistical modeling and several machine learning methods?

Table 1 demonstrates that, for most cases, all the individual constituent methods within the new prediction metasystem show superior performances to the existing NRCS PCR system. The benefits are by far the most noticeable for the Gila and Owyhee Rivers, which as discussed above are known to be marked by both nonlinear dependence of spring-summer flow volume on wintertime precipitation and snow accumulation, as well as heteroscedastic and non-Gaussian residuals. The prediction engine was specifically built to easily handle these complications (see Sections I and II). It produces superior performance statistics to VIPER, and of particular note, its nonlinear constituent methods – that is, the four machine learning techniques (mANN, RF, SVM, MCQRNN) – produce predictions, and associated prediction intervals, that are always positive-valued, a key physicality requirement for acceptance of machine learning solutions in a water resource prediction application. (We note that it is perhaps ironic that the machine learning methods, which are often viewed as being black-box and without physical interpretation, deliver geophysical predictions that better match known physical processes and constraints.) While the linearity of QR can be a limiting factor in its deterministic prediction accuracy and physicality for some rivers, most notably Owyhee, it also usually turns in the best RPSS values among all the individual methods in the ensemble and relative to the conventional linear PCR modeling benchmark, indicating an ability to contribute the most accurate quantitative estimates of prediction uncertainty.

For the Deschutes, which as noted above is known to be a linear, Gaussian, homoscedastic regression problem for which a traditional statistical model should suffice, the prediction system performs about on par with the existing VIPER approach – so there is no disadvantage to using it even when its full capabilities are not required. In this case, the various machine learning methods, though capable of accommodating nonlinearities, faithfully capture the essentially linear relationships, at this river, between spring-summer flow volume and its predictors. Similarly, the more sophisticated and flexible prediction interval generation techniques used in all the constituent methods within the ensemble, including but not limited to both of the linear statistical methods (QR, LR), automatically reduce to the homoscedastic and Gaussian uncertainty estimates required for this river.

3) MULTI-METHOD ENSEMBLE MEAN PERFORMANCE RELATIVE TO LINEAR STATISTICAL MODELING BENCHMARK OF EXISTING NRCS WSF SYSTEM

The most immediately important litmus test for the prediction engine is that its final product, the multi-method ensemble mean forecast distribution, matches or exceeds the

TABLE 1. Performance of existing linear statistical modeling-based NRCS WSF VIPER system vs. integrated multi-method ensemble.

Metric [a]	VIPER	LR	QR	mANN	RF	MCQRNN	SVM	Ensemble [b]
Gila:								
R^2	0.62	0.69	0.71	0.80	0.73	0.71	0.74	**0.76**
RMSE	9.9	9.0	9.2	7.3	8.5	8.8	8.4	**8.0**
RPSS	0.53	0.59	0.64	0.62	0.64	0.63	0.64	**0.66**
BE<0?	N	N	N	N	N	N	N	**N**
PB<0?	N	N	Y	N	N	N	N	**N**
Owyhee:								
R^2	0.67	0.67	0.65	0.70	0.85	0.64	0.81	**0.83**
RMSE	138	137	146	130	101	147	108	**105**
RPSS	0.49	0.53	0.54	0.45	0.43	0.54	0.59	**0.56**
BE<0?	Y	Y	Y	N	N	N	N	**N**
PB<0?	Y	Y	Y	N	N	N	N	**N**
Deschutes:								
R^2	0.78	0.81	0.82	0.75	0.74	0.81	0.81	**0.83**
RMSE	5.4	5.1	4.8	5.7	6.0	5.0	5.1	**4.8**
RPSS	0.61	0.59	0.66	0.56	0.56	0.60	0.52	**0.61**
BE<0?	N	N	N	N	N	N	N	**N**
PB<0?	N	N	N	N	N	N	N	**N**

[a] R^2: coefficient of determination; RMSE: root mean square error; RPSS: ranked probability skill score, measuring probabilistic forecast performance (accuracy of forecast distribution: 1=perfect, 0=no better than assuming equal likelihood at any n of below-normal, normal, or above-normal conditions as defined by terciles of the empirical cumulative distribution function of the observations); BE/PB<0? indicate whether a negative-valued best estimate or prediction bound occurs at any n.

[b] The ensemble mean forecast distribution consists of the average of the predictand probability distributions generated by all of the constituent models as per (7) and is the final product of the integrated multi-method prediction engine. The initial ensemble was built using the default models, LR, QR, mANN, RF, MCQRNN, and SVM; the automated QC step of (17) removed LR and QR from the final ensemble for Owyhee but no model pruning was required for Gila or Deschutes. VIPER refers to the NRCS official PCR water supply forecast model, and its performance metrics are provided here as a benchmark. See text for detailed explanation.

performance of the existing official NRCS PCR (VIPER) model forecast, a key design criterion for the new forecast technique (Section I.B).

Table 1 shows that it does so, across all rivers, for every performance metric. Significant gains are seen in RMSE, R^2, and RPSS, and strictly non-negative predictions and prediction intervals are generated, without manual user requirements around evaluating the need for, selecting, and implementing predictand transforms. The sole partial exception is RPSS for Deschutes, for which it matches the performance of the linear statistical WSF benchmark model.

The degree to which the mean forecast distribution from the multi-method framework outperforms the conventional linear PCR modeling reference forecast varies between basins in much the same way as discussed for its individual constituent methods in points (1) and (2) above. That is, on the one hand, the advantages of the integrated prediction engine are most pronounced for rivers that have the sorts of probabilistic regression challenges it was intended to tackle, i.e., nonlinear relationships and a need for time-varying and asymmetric prediction bounds at Gila and Owyhee. On the other hand, for Deschutes, where those capabilities are not required, the multi-method ensemble mean essentially reproduces the performance of a conventional linear Gaussian regression model, as desired; although the performance still appears somewhat better than that of the NRCS PCR model, presumably at least in part due to the capabilities discussed in point (1), in particular the use of a genetic algorithm for feature selection.

4) MODEL SELECTION UNCERTAINTY WITHIN THE MULTI-METHOD FRAMEWORK

Model selection uncertainty and equifinality amongst the individual methods within the multi-method framework are strongly evident. No single method – LR, SVM, RF, mANN, QR, or MCQRNN – can be said to be uniquely best across all rivers and performance measures.

In the interest of conciseness, we will not run through the results of every method, river, and fit metric, but a few examples illustrate the point. For instance, RF is the clear winner for deterministic measures (R^2 and RMSE) for Owyhee yet its probabilistic accuracy measure (RPSS) for this watershed is the worst of the six methods; its performance for Deschutes is mediocre compared to the other methods; and its R^2 and RMSE for Gila are middling yet its RPSS for that watershed is tied for top spot among the individual methods. Additionally, even where models have clear flaws in some respects, they also offer distinct advantages in other respects: while MCQRNN turns in the poorest accuracy metrics for Owyhee and may therefore initially seem like a low-performing method, it reliably ensures non-negative predictions and associated prediction intervals, i.e., physically plausible water supply forecasts, such that it is in this crucial respect superior to several of the other constituent methods for this watershed. Similarly, as a linear method, QR can have difficulty dealing with the more nonlinear basins, particularly Owyhee where the method-pruning algorithm of (17) removed it from the ensemble (see details in footnotes to Table 1), but as noted in point (2) above, it consistently turns

in some of the best RPSS performances of all the individual probabilistic regression methods within the multi-method framework and therefore contributes substantial value around prediction uncertainty. Each of the methods offers capabilities and limitations.

5) MUTUAL ERROR CANCELLATION IN THE MULTI-METHOD ENSEMBLE MEAN

The multi-method mean prediction, which meets or beats the linear performance benchmark of the NRCS PCR statistical model (point (3)), also meets – or beats – its constituent methods in many respects.

Of particular note is that the water supply forecast obtained by averaging those made by the six constituent methods delivers a RPSS value of 0.66 for Gila, whereas the RPSS values for the predictions of each of those constituent methods ranged from 0.59 (LR) to 0.64 (QR, RF, and SVM). Similarly, the multi-method ensemble mean prediction R^2 for Deschutes is 0.83, better than that of any of the methods that went into the ensemble. These results are consistent with the well-known mutual error cancellation property of multi-model averaging (see Section II.C).

Even where the ensemble mean prediction does not outperform all of its constituent methods, it typically delivers performance metrics that are among the best for a given river and metric, that is, it has reliably consistent prediction quality, as discussed in point (6) below.

6) GREATER PERFORMANCE CONSISTENCY OF THE MULTI-METHOD ENSEMBLE MEAN RELATIVE TO ITS CONSTITUENT METHODS

A particularly notable asset of the integrated multi-method prediction engine is that, in addition to outperforming the VIPER linear benchmark (see point (3) above), it also provides greater overall consistency and reliability relative to any of the individual methods within it.

The overall implications are dramatic. For each river, irrespective of its geophysical and statistical characteristics, the multi-method ensemble mean is always either the best or second-best performer for all five prediction quality measures, that is, without exception it provides the best or second-best quantitative metrics of both deterministic and probabilistic forecast performance (R^2, RMSE, RPSS) and for binary metrics (BE, PB$<$0?) it was always correct. In contrast, the worst performance for each of the individual constituent methods was always worse than second for each of the rivers, and indeed, some methods capable of turning in a very good performance for one river or metric performed, comparatively, quite poorly on others.

The multi-method ensemble mean addresses model selection uncertainty by capitalizing on the strengths of individual methods and damping their weaknesses, providing a more stable performance than any individual ensemble member. For example, although PB $<$ 0 for QR at Gila, QR was still retained by the QC algorithm of (17) as this individually non-physical result did not ultimately lead to a non-physical ensemble mean prediction, and in fact, combining QR with the outcomes of the other constituent models allowed the ensemble to capitalize on the good aspects of QR performance, in particular high RPSS, while overcoming its poor aspect, i.e., the generation of negative-valued prediction intervals. As another clear example, MCQRNN at Owyhee provided relatively poor quantitative performance measures but its guarantee of BE, PB $\geq$ 0 helped ensure that the multi-method ensemble mean satisfied key physicality constraints, which was a significant challenge for Owyhee.

This consistency is a crucial practical advantage for an operational forecast system intended for ultimate application to over 600 forecast points across the western US. Given the model selection uncertainty and equifinality apparent for the six individual methods in Table 1 and described above, choosing a single best (across all performance measures, and hundreds of rivers) regression/regression-like modeling technique would seem impossible. But by integrating these six, very different, regression methods into a multi-method averaging framework, far greater consistency and reliability in prediction quality (R^2, RMSE, RPSS) and physical plausibility (BE, PB$<$0?) are achieved and the prediction metasystem can be applied with greater confidence across the modeling domain.

7) COROLLARIES TO THE PERFORMANCE CONSISTENCY AND RELIABILITY OF THE MULTI-METHOD ENSEMBLE MEAN

There are two interesting corollaries to point (6).

First, by exactly the same token, several testing runs (not shown for conciseness) also demonstrate the multi-method ensemble mean tends to damp overall performance fluctuations arising from specific architecture and hyperparameter choices for the individual methods (mANN, RF, etc.) and for the GA, provided of course that these choices are broadly reasonable. Such configuration and hyperparameter selections can lead to slightly different final models for each technique, and to correspondingly different relative rankings among the methods as captured by various performance measures, but the ensemble mean across all the methods tends to remain largely stable. In general, for a given test case, noteworthy changes in the ensemble mean performance only occurred if major changes in procedure were made, such as increasing the scale of the GA problem by two orders of magnitude or omitting the GA optimization altogether (see also hyperparameter tuning discussion in Section II.G).

Second, recall from Section II.C and II.D that some of the individual methods within the multi-method framework are themselves ensemble learners, specifically, RF, and for configurations where bagging is employed mANN. That is, as discussed previously, the multi-method framework is in part an ensemble of ensembles, or super-ensemble to borrow a term from the weather and climate modeling community. The multi-method ensemble framework is obviously not in competition with individual ensemble learning methods like RF to replace them, as it integrates and depends upon

these methods. However, it is interesting to contemplate the implications of the performance of the multi-method prediction engine relative to each of its individual constituent methods (points 4, 5, and 6 above) in light of the fact that some of those methods are in turn ensemble learners. The prediction engine therefore becomes, in part, a multi-level hierarchy of ensemble learners, at least with respect to RF and bagged mANN, and the far greater consistency of the multi-method ensemble mean, and for some metrics and rivers its better performance measures through error cancellation (see points (5) and (6) above), show that (i) while an individual ensemble learner benefits from internal model averaging, such that, typically, RF performs better than a single isolated regression tree or a bagged mANN performs better than a single isolated neural network, there is still significant model selection uncertainty and equifinality when comparing these classes of individual ensemble learner (that is, RF vs. bagged mANN) against each other, (ii) still higher levels of model aggregation beyond that employed within a RF or bagged mANN, in particular that implemented in the multi-method engine across six distinct modeling technologies including but not limited to both RF and bagged mANN, is useful for addressing that model selection uncertainty and equifinality, and (iii) the more diverse the methods within a multi-method average, the better its performance and in particular the reliability of that performance, consistent with prior experience in the machine learning, statistical modeling, and geophysical and environmental modeling communities.

IV. CONCLUSION

We describe a study in which a number of supervised and unsupervised machine learning, nonparametric statistical, ensemble modeling, and evolutionary optimization methods were integrated into a prediction metasystem and used to radically update and improve an existing principal components regression framework for water supply forecasting in the US West. It has direct implications to various aspects of water resource management, including optimal hydroelectric power generation and planning, and potentially, sustainable practices in certain water-intensive tech-sector areas like chip manufacturing and server farms. More broadly, the technique may have applications to probabilistic nonlinear regression problems in other applied science and engineering fields having similar statistical requirements. The study also provides a successful demonstration of the way in which the performance and reliability of established geophysical prediction techniques can be upgraded using artificial intelligence and other modern data analytics methods, through a process of carefully merging the bodies of knowledge and practice of machine learning with those of environmental science; lessons learned from this example may have implications to successful AI implementation in applied science and engineering areas beyond environmental prediction.

Specifically, a new, largely machine learning-based prediction system was developed using a relatively multi-disciplinary philosophy to replace the existing linear statistical regression framework for the largest stand-alone water supply forecast system in the western US. A central aspect of both the existing and new systems are that they are probabilistic, generating predictions consisting of a best estimate with associated prediction intervals. The new approach is an integrative, modular, multi-method framework for probabilistic nonlinear regression modeling that is suitable for a wide class of prediction problems and incorporates – not replaces – a careful selection of well-established concepts in predictive analytics and ensemble modeling drawn from the machine learning, statistics, risk assessment, and hydrologic and climate modeling communities. In particular, it amounts to a prediction metasystem that capitalizes on the successful aspects of the existing, well-proven system by upgrading its principal components regression framework using a range of modern machine learning and evolutionary computing techniques; enhances flexibility and efficiency by automating many steps and incorporating various nonlinear regression methods and error estimation techniques chosen for their ability to handle a wide range of problem types, including nonlinear relationships and heteroscedastic and non-Gaussian prediction errors, without requiring user intervention or excessive tuning; and employs a multi-method ensemble spanning a diverse selection of regression and regression-like modeling techniques to create a relatively robust and stable estimator and to help sidestep model selection uncertainty. Testing of the new system suggests it is more amenable to automation and produces more accurate forecasts than the well-established and finely tuned current operational system.

A wide range of additional work is under consideration, including experimentation with new predictor types such as outputs from snow and climate models and airborne and satellite remote sensing products, upgrading some additional elements of the prediction framework, adding other types of machine learning-based nonlinear regression methods to the multi-method ensemble, and integrating WSFs from physics-based process simulation models, such as those issued by other agencies like the National Weather Service, into the multi-method ensemble. In the latter case, the prediction engine would grow from a modeling framework into a broader platform for integrating a wide variety of prediction products from various sources generated using different methods; this is broadly consistent with some contemporary research directions in hydrologic modeling [56], [57] and in principle could enable a re-introduction of the informal multi-agency forecast coordination process that used to take place in the US West.

Demonstrating that a suitable machine learning approach to WSF, developed jointly using both geophysical and AI knowledge and tools, can successfully transition from research to operational agency applications has implications for improving other WSF systems, in the US West and elsewhere. Globally, over a billion people are currently without adequate access to water, and estimates call for an increase of 55% in water demand by 2050 due to population

and economic growth [58]. Climate change may also be a concern, potentially leading to increasingly unpredictable water supplies [59]. Successful management and planning of water for basic living needs, water-intensive agricultural and industrial production, hydroelectric power generation, and ecological and legal requirements, will demand increasingly powerful geophysical prediction tools as the margins between water supply and demand narrow.

The relatively multifaceted and interdisciplinary approach taken here, in which diverse required design criteria were achieved by integrating multiple existing techniques into a type of regression metasystem that combines the qualities of the constituent methods, may also be useful to prediction of other types of complex open systems. Generation of quantitative prediction intervals that accommodate complex (in particular, heteroscedastic and non-Gaussian) predictive errors, integration of some basic physical process considerations, nonlinearity, high dimensionality, model selection uncertainty and equifinality, reduced need for manual user intervention and increased amenability to automation, and low cost are all requirements for many problems. Examples include other geophysical problems, ecosystem modeling, and economic systems. As noted above, predictive analytics methods that accommodate one or a few of these requirements and complications are common; methods that simultaneously accommodate all of them are not.

Some ability to impose solution constraints suggested by physical knowledge of the process being modeled may be worth emphasizing given current interest in, and criticism of, machine learning solutions. Water supply forecasting is not the only applied science and engineering field in which machine learning has encountered difficulty transitioning from academic research to widespread mainstream use, or finding its place within the standard industrial toolkits of those fields. Another well-documented example is materials science, and questions like small sample sizes, addressing uncertainty, and in particular an expectation for some level of physics-awareness, may set such applied science and engineering applications apart from many other AI uses [60]. Some of these issues tie into still-broader questions: when IBM surveyed 5,000 businesses about using artificial intelligence, 82% expressed interest yet two-thirds of those companies indicated they were reluctant to proceed, with the leading roadblock being that the resulting machine learning solutions suffered from a lack of explainability in terms of underlying (e.g., physical) processes [61]. Our study does not solve these problems, and indeed, solutions may be application area-specific, such as methods for optimal design of chemistry experiments to develop new materials [62]. Nevertheless, some aspects of the approach adopted here for helping ensure physicality, such as guaranteeing non-negative predictions or monotonically nonlinear functional relationships through appropriate selection of specific machine learning methods from the huge range of AI techniques available, and developing post-modeling QC steps to retroactively adjust ensemble composition when needed, are likely to be transportable to other fields and at a minimum provide further demonstration that it is possible to incorporate some physical process knowledge into machine learning solutions in a practical way.

ACKNOWLEDGMENT

We thank V. Muggeo (Dept. Economics, Business, and Statistics, University of Palermo), A. Cannon (Climate Research Division, Environment and Climate Change Canada), D. Goodsman, A.E. Garcia, and M. Vesselinov (Los Alamos National Laboratory), C. Schwarz (Dept. Statistics and Actuarial Science, Simon Fraser University), A. Ellenson (College of Earth, Ocean, and Atmospheric Sciences, Oregon State University), W. Hsieh (Dept. Physics and Astronomy, University of British Columbia), and B. Meredig (Citrine Informatics Inc. and Dept. Materials Science and Engineering, Stanford University) for valuable conversations. White Rabbit R&D LLC contributions were funded by NRCS through Elyon International Inc. The work also benefited tremendously from extensive discussion with, and support from, NRCS personnel: C. McCarthy, D. Garen, R. Tama, J. Lea, C. Brown, and M. Strobel. Four anonymous reviewers and the IEEE associate editor provided helpful comments and suggestions.

REFERENCES

[1] A. F. Hamlet, D. Huppert, and D. P. Lettenmaier, "Economic value of long-lead streamflow forecasts for Columbia River hydropower," *ASCE J. Water Resour. Planning Manage.*, vol. 128, pp. 91–101, Mar. 2002.

[2] S. W. D. Turner, N. Voisin, J. Fazio, D. Hua, and M. Jourabchi, "Compound climate events transform electrical power shortfall risk in the Pacific Northwest," *Nature Commun.*, vol. 10, 2019, Art. no. 8. doi: 10.1038/s41467-018-07894-4.

[3] U.S. Energy Information Administration. *Northwest Heat Wave Leads to Record Levels of Summer Electricity Demand*. Accessed: Apr. 24, 2019. [Online]. Available: https://www.eia.gov/todayinenergy/detail.php?id=32612

[4] F. Weber, D. Garen, and A. Gobena, "Invited commentary: Themes and issues from the workshop 'operational river flow and water supply forecasting,"' *Can. Water Resour. J./Revue Canadienne Ressources Hydriques*, vol. 37, pp. 151–161, Jan. 2012.

[5] D. J. Druce, "Incorporating daily flood control objectives into a monthly stochastic dynamic programing model for a hydroelectric complex," *Water Resour. Res.*, vol. 26, pp. 5–11, Jan. 1990.

[6] A. McManamon, "Inflow forecasting at Bonneville Power Administration," presented at the Centre Energy Advancement Through Tech. Innov. (CEATI) Inflow Forecasting Workshop, Knoxville, TN, USA, Nov. 2007.

[7] D. A. Harpman, "Exploring the economic value of hydropower in the interconnected electricity system," Bureau Reclamation, Denver, CO, USA, Tech. Rep. EC-2006-03, 2006.

[8] Powerex. *Trading With Powerex*. Accessed: Apr. 24, 2019. [Online]. Available: https://www2.powerex.com/TradingWithPowerex.aspx

[9] T. R. Perkins, T. C. Pagano, and G. C. Garen, "Innovative operational seasonal water supply forecasting technologies," *J. Soil Water Conservation*, vol. 64, pp. 15–17, 2009.

[10] G. C. Garen, "Improved techniques in regression-based streamflow volume forecasting," *J. Water Resour. Planning Manage.*, vol. 118, pp. 654–670, Nov. 1992.

[11] D. E. Robertson, P. Pokhrel, and Q. J. Wang, "Improving statistical forecasts of seasonal streamflows using hydrological model output," *Hydrol. Earth Syst. Sci.*, vol. 17, no. 2, pp. 579–593, 2013.

[12] A. K. Gobena and T. Y. Gan, "Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system," *J. Hydrol.*, vol. 385, pp. 336–352, May 2010.

[13] A. W. Minns and M. J. Hall, "Artificial neural networks as rainfall-runoff models," *Hydrol. Sci. J.*, vol. 41, no. 3, pp. 399–417, 1996.

[14] W. W. Hsieh, J. Li, A. Shabbar, and S. Smith, "Seasonal prediction with error estimation of Columbia River streamflow in British Columbia," *J. Water Resource Planning Manage.*, vol. 129, no. 2, pp. 146–149, 2003.

[15] A. Kalra, W. P. Miller, K. W. Lamb, S. Ahmad, and T. Piechota, "Using large-scale climatic patterns for improving long lead time streamflow forecasts for Gunnison and San Juan River Basins," *Hydrol. Processes*, vol. 27, no. 11, pp. 1543–1559, 2013.

[16] R. J. Abrahart, F. Anctil, P. Coulibaly, C. W. Dawson, N. J. Mount, L. M. See, A. Y. Shamseldin, D. P. Solomatine, E. Toth, and R. L. Wilby, "Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting," *Prog. Phys. Geogr., Earth Environ.*, vol. 36, no. 4, pp. 480–513, 2012.

[17] S. W. Fleming, D. R. Bourdin, D. Campbell, R. B. Stull, and T. Gardner, "Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river," *J. Amer. Water Resour. Assoc.*, vol. 51, pp. 502–512, Apr. 2015.

[18] A. J. Cannon and I. G. McKendry, "A graphical sensitivity analysis for statistical climate models: Application to Indian monsoon rainfall prediction by artificial neural networks and multiple linear regression models," *Int. J. Climatol.*, vol. 22, pp. 1687–1708, Nov. 2002.

[19] S. W. Fleming, "Artificial neural network forecasting of nonlinear Markov processes," *Can. J. Phys.*, vol. 85, no. 3, pp. 279–294, 2007.

[20] T. Beuzen, K. D. Splinter, L. A. Marshall, I. L. Turner, M. D. Harley, and M. L. Palmsten, "Bayesian networks in coastal engineering: Distinguishing descriptive and predictive applications," *Coastal Eng.*, vol. 135, pp. 16–20, May 2018.

[21] A. R. Lima, W. W. Hsieh, and A. J. Cannon, "Variable complexity online sequential extreme learning machine, with applications to streamflow prediction," *J. Hydrol.*, vol. 555, pp. 983–994, 2017.

[22] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[23] D. H. Wolpert, "The lack of a priori distinctions between learning algorithms," *Neural Comput.*, vol. 8, no. 7, pp. 1341–1390, 1996.

[24] Z.-H. Zhou, *Ensemble Methods: Foundations and Algorithms*. New York, NY, USA: CRC Press, 2012.

[25] P. I. Rani and K. Muneeswaran, "Facial emotion recognition based on eye and mouth regions," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 30, no. 7, 2016, Art. no. 1655020. doi: 10.1142/S021800141655020X.

[26] A. Verma and S. Mehta, "A comparative study of ensemble learning methods for classification in bioinformatics," in *Proc. IEEE 7th Int. Conf. Cloud Comput., Data Sci. Eng.*, Noida, India, Jan. 2017, pp. 155–158.

[27] Y. Tao, Y. J. Chen, X. Fu, B. Jiang, and Y. Zhang, "Evolutionary ensemble learning algorithm to modeling of warfarin dose prediction for Chinese," *IEEE J. Biomed. Health Inform.*, vol. 23, no. 1, pp. 395–406, Jan. 2019.

[28] F. Lv, M. Han, and T. Qiu, "Remote sensing image classification based on ensemble extreme learning machine with stacked autoencoder," *IEEE Access*, vol. 5, pp. 9021–9031, 2017.

[29] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York, NY, USA: Springer, 2002.

[30] J. A. Hoeting, D. Madigan, A. E. Raftery, and C. T. Volinsky, "Bayesian model averaging: A tutorial," *Stat. Sci.*, vol. 14, pp. 382–417, Nov. 1999.

[31] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis," *Risk Anal.*, vol. 19, pp. 187–203, Apr. 1999.

[32] R. Hagedorn, F. J. Doblas-Reyes, and T. N. Palmer, "The rationale behind the success of multi-model ensembles in seasonal forecasting—I. Basic concept," *Tellus A, Dyn. Meteorol. Oceanogr.* , vol. 57, no. 3, pp. 219–233, 2005.

[33] M. R. Najafi and H. Moradkhani, "Ensemble combination of seasonal streamflow forecasts," *J. Hydrol. Eng.*, vol. 21, no. 1, 2016, Art. no. 04015043. doi: 10.1061/(ASCE)HE.1943-5584.0001250.

[34] H. Zhang and Z. Zhang, "Feedforward networks with monotone constraints," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Washington, DC, USA, vol. 3, Jul. 1999, pp. 1820–1823.

[35] A. J. Cannon. (2017). *monmlp: Monotone Multi-Layer Perceptron Neural Network. R Package Version 1.1.4*. [Online]. Available: https://CRAN.R-project.org/package=monmlp

[36] L. Feyen, M. Kalas, and J. A. Vrugt, "Semi-distributed parameter optimization and uncertainty assessment for large-scale streamflow simulation using global optimization," *Hydrol. Sci. J.*, vol. 53, no. 2, pp. 293–308, 2008.

[37] R. J. Hyndman. (2017). *forecast: Forecasting Functions for Time Series and Linear Models. R Package Version 8.2*. [Online]. Available: http://pkg.robjhyndman.com/forecast

[38] A. J. Cannon, "Non-crossing nonlinear regression quantiles by monotone composite quantile regression neural network, with application to rainfall extremes," *Stochastic Environ. Res. Risk Assessment*, vol. 32, no. 11, pp. 3207–3225, 2018. doi: 10.1007/s00477-018-1573-6.

[39] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[40] Y. Liang, Q. S. Xu, H. D. Li, and D. S. Cao, *Support Vector Machines and Their Application in Chemistry and Biotechnology*. Boca Raton, FL, USA: CRC Press, 2011.

[41] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, and F. Leisch. (2017). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R Package Version 1.6-8*. [Online]. Available: https://CRAN.R-project.org/package=e1071

[42] R. Koenker and K. Hallock, "Quantile regression," *J. Econ. Perspect.*, vol. 15, no. 4, pp. 143–156, 2001.

[43] Muggeo VMR. (2018). *quantregGrowth: Growth Charts via Regression Quantiles. R Package Version 0.4-3*. [Online]. Available: https://cran.r-project.org/web/packages/quantregGrowth/index.html

[44] A. Das and D. Kempe, "Algorithms for subset selection in linear regression," in *Proc. 40th Annu. ACM Symp. Theory Comput. (STOC)*, Victoria, BC, Canada, May 2008, pp. 45–54.

[45] K. G. Balcombe, "Model selection using information criteria and genetic algorithms," *Comput. Econ.*, vol. 25, pp. 207–228, Jun. 2005.

[46] V. Calcagno and C. de Mazancourt, "glmulti: An R package for easy automated model selection with (generalized) linear models," *J. Stat. Softw.*, vol. 34, no. 12, pp. 1–29, 2010.

[47] V. Trevino and F. Falciani, "GALGO: An R package for multivariate variable selection using genetic algorithms," *Bioinformatics*, vol. 22, pp. 1154–1156, May 2006.

[48] I. T. Jolliffe, "A note on the use of principal components in regression," *J. Roy. Stat. Soc. C*, vol. 31, no. 3, pp. 300–303, 1982.

[49] E. Willighagen and M. Ballings. (2015). *genalg: R Based Genetic Algorithm. R Package Version 0.2.0*. [Online]. Available: https://CRAN.R-project.org/package=genalg

[50] P. Cortez, *Modern Optimization With R*. Cham, Switzerland: Springer, 2014.

[51] N. Murata, S. Amari, and S. Yoshizawa, "Network information criterion-determining the number of hidden units for an artificial neural network model," *IEEE Trans. Neural Netw.*, vol. 5, no. 6, pp. 865–872, Nov. 1994.

[52] J. Ye, "On measuring and correcting the effects of data mining and model selection," *J. Amer. Stat. Assoc.*, vol. 93, no. 441, pp. 120–131, 1998.

[53] U. Anders and O. Korn, "Model selection in neural networks," *Neural Netw.*, vol. 12, no. 2, pp. 309–323, 1999.

[54] Microsoft Corporation and S. Weston. (2017). *foreach: Provides Foreach Looping Construct for R., R Package Version 1.4.4*. [Online]. Available: https://CRAN.R-project.org/package=foreach

[55] Microsoft Corporation and S. Weston. (2017). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package, R Package Version 1.0.11*. [Online]. Available: https://CRAN.R-project.org/package=doParallel

[56] J. Quilty, J. Adamowski, and M.-A. Boucher, "A stochastic data-driven ensemble forecasting framework for water resources: A case study using ensemble members derived from a database of deterministic wavelet-based models," *Water Resour. Res.*, vol. 55, pp. 175–202, Jan. 2019.

[57] H. Tyralis, G. Papacharalampous, A. Burnetas, and A. Langousis, "Hydrological post-processing using stacked generalization of quantile regression algorithms: Large-scale application over CONUS," *J. Hydrol.*, vol. 577, Oct. 2019, Art. no. 123957. doi: 10.1016/j.jhydrol.2019.123957.

[58] *The United Nations World Water Development Report 2015: Water for a Sustainable World*, United Nations World Water Assessment Programme, UNESCO, Paris, France, 2015.

[59] L. Brekke, K. Werner, D. Laurine, and D. Garen, "Climate change impacts on water supply predictability," in *American Meteorological Society Short Course, Hydrologic Prediction and Verification Techniques With a Focus on Water Supply*. Seattle, WA, USA, Jan. 2011.

[60] B. Meredig, "Solving industrial materials problems with machine learning," presented at the Amer. Phys. Soc. March Meeting, Los Angeles CA, USA, Mar. 2018.

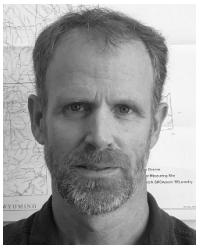
[61] J. Kahn. (Dec. 21, 2018). *Artificial Intelligence has Some Explaining to Do, Bloomberg Businessweek*. [Online]. Available: www.bloomberg.com

[62] J. H. Martin, B. D. Yahata, J. M. Hundley, J. A. Mayer, T. A. Schaedler, and T. M. Pollock, "3D printing of high-strength aluminium alloys," *Nature*, vol. 549, pp. 365–369, Sep. 2017.

SEAN W. FLEMING received the B.Sc. degree in geophysics from the Department of Geophysics and Astronomy, The University of British Columbia, in 1994, the M.S. degree in geophysics from the College of Oceanic and Atmospheric Sciences, Oregon State University, in 1997, the M.S. degree in geology and civil engineering from the Department of Geoscience, Oregon State University, and also from the Department of Civil and Environmental Engineering, Oregon State University, in 1998, and the Ph.D. degree in geophysics from the Department of Earth, Ocean, and Atmospheric Sciences, The University of British Columbia, in 2004.

He operates White Rabbit R&D LLC, an Oregon-based consultancy focusing on practical machine learning applications to multidisciplinary applied science problems (www.facebook.com/westcoastdatascience). He has more than two decades of experience in the public, private, academic, and NGO sectors in the U.S., Canada, U.K., and Mexico, including performing operational forecasting for a large hydroelectric utility and managing a federal government research unit, and he has been working with AI applications to environmental systems for almost 20 years. He is also a Courtesy Professor and Graduate Faculty with Oregon State University, where he contributes practical industry perspectives to collaborative research with full-time faculty and helps supervise graduate students. His work has been extensively published in the physics, geophysics, water resources, climate, biology, environmental management, and engineering research literature. He is also strongly active in science outreach, publishing a general-audience book with Princeton University Press (*Where the River Flows: Scientific Reflections on Earth's Waterways*), engaging the public through events like a Smithsonian lecture, OSU Science Pub talks, and a live NPR interview, and writing for Scientific American and Wired.

He holds professional registrations with the Canadian Association of Physicists, the Canadian Meteorological and Oceanographic Society, and the Association of Professional Engineers and Geoscientists of British Columbia. He is a long-time member of the American Geophysical Union and the American Physical Society.



ANGUS G. GOODBODY was born in Brooklyn, New York, NY, USA, in 1971. He received the B.A. degree in geography and geology from Macalester College, St. Paul, MN, USA, in 1994, and the M.S. degree in watershed science from Colorado State University, Fort Collins, CO, USA, in 2004.

From 2000 to 2004, he was a Research Assistant with the Department of Earth Resources, Colorado State University, supporting field campaigns for NASA's Cold Lands Processes Experiment. From 2004 to 2007, he was a Research Hydrologist with the Rocky Mountain Research Station, supporting a variety of data collection and analysis projects at the Fraser Experimental Forest. From 2007 to 2008, he was an operational Forecast Hydrologist with NOAA's Northwest River Forecast Center, providing daily flood and water supply forecasts for the Columbia Basin. Since 2008, he has been a Forecast Hydrologist with the National Water and Climate Center, part of the U.S. Department of Agriculture (USDA)'s Snow Survey and Water Supply Forecasting Program, providing systems support and operational water supply forecasts for the Colorado, Rio Grande, and Columbia basins. He has coauthored several articles over his professional tenure. His research interests span topics on snowpack processes, distribution and extent, and hydrologic forecasting processes and systems to support and improve operational water supply planning in the western United States.

Mr. Goodbody has been a member of the American Geophysical Union, since 2004.