# APPENDIX A

Appendix A contains the following:

> Fleming SW, Garen DC, Goodbody AG, McCarthy CS, Landers LC. 2021. Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: a challenging test of explainable, automated, ensemble artificial intelligence. *Journal of Hydrology*, 602, 126782.

This article provides a hydrology-focused overview of M$^4$, its intent and design criteria, its performance in retrospective and live operational testing in 20 test cases spanning 11 sites in the western US and Alaska, geophysical explainability of its results, possible implications of climate change, and directions going forward.

Journal of Hydrology 602 (2021) 126782

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

# Assessing the new Natural Resources Conservation Service water supply forecast model for the American West: A challenging test of explainable, automated, ensemble artificial intelligence

Sean W. Fleming [a,b,c,d,*], David C. Garen [a,1], Angus G. Goodbody [a], Cara S. McCarthy [a], Lexi C. Landers [a]

[a] National Water and Climate Center, Natural Resources Conservation Service, US Department of Agriculture, Portland, OR, USA
[b] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA
[c] Water Resource Graduate Program, Oregon State University, Corvallis, OR, USA
[d] Department of Earth, Ocean, and Atmospheric Sciences, University of British Columbia, Vancouver, BC, Canada

ARTICLE INFO

This manuscript was handled by Marco Borga, Editor-in-Chief, with the assistance of Andrea Castelletti, Associate Editor

*Keywords:*
Water supply forecasting
Operational hydrology
Explainable machine learning
Hydrologic modeling
Water management

ABSTRACT

Western US water management is underpinned by spring-summer water supply forecasts (WSFs) from hydrologic models forced primarily by winter mountain snowpack data. The US Department of Agriculture Natural Resources Conservation Service (NRCS) operates the largest such system regionally. NRCS recently developed a next-generation WSF prototype, the multi-model machine-learning metasystem ($M^4$). Here, we test this ensemble artificial intelligence (AI)-based prototype against challenging theoretical and practical criteria for accepting a new operational WSF model. In 20 hindcasting test-cases spanning diverse environments across the western US and Alaska, on average out-of-sample $R^2$ and RPSS improved over 50% and RMSE improved 13% relative to current benchmarks. The $M^4$ ensemble mean forecast also performed more consistently than any of its diverse constituent models and in several cases outperformed all of them. Live operational testing at a subset of sites during the 2020 forecast season additionally demonstrated logistical feasibility of workflows, as well as geophysical explainability of results in terms of known hydrologic processes, belying the black-box reputation of machine learning and enabling relatable forecast storylines for clients. This was accomplished using WSF-focused pragmatic solutions, like "popular votes" for different candidate predictors among the constituent forecast systems, and graphical visualization of reduced-dimension, AI-extracted nonlinear feature-target relationships. We also found that certain $M^4$ technical design elements, including autonomous machine learning (AutoML), hyperparameter pre-calibration, and theory-guided data science, collectively permitted automated ("over-the-loop") training and operation. Overall, the analyses confirmed $M^4$ meets requirements for NRCS operational adoption. This finding signals that, despite negligible operational-community uptake of machine learning so far, suitably purpose-designed novel AI systems have capacity to transition into large-scale practical applications with service-delivery organizations; it appears $M^4$ will be the largest AI migration into operational river forecasting to date. It may ultimately provide a broader integration platform for harnessing multiple data and model types.

## 1. Introduction

### 1.1. River runoff volume forecasting in the western US

Water scarcity defined the history of the western US and remains one of its most complex and pressing public issues: economic, food, environmental, and energy security here all depend critically on river runoff (e.g., Rosenberg et al., 2011; Reisner, 1986). Effective water management in this region relies on operational water supply forecasts (WSFs) (Glantz, 1982; Kalra et al., 2013; Grantz et al., 2005; Hoekema and Ryu,

* Corresponding author at: Applied R&D Technical Lead, National Water and Climate Center, Natural Resources Conservation Service, US Department of Agriculture, 1201 NE Lloyd Boulevard, Suite 802, Portland, OR 97232-1274, USA.
*E-mail address:* sean.fleming@usda.gov (S.W. Fleming).
[1] Emeritus

https://doi.org/10.1016/j.jhydrol.2021.126782
Received 23 March 2021; Received in revised form 15 June 2021; Accepted 30 July 2021
Available online 11 August 2021
0022-1694/Published by Elsevier B.V.

2013). These are predictions of spring-summer runoff volume on a river-by-river basis, typically issued at the start of every month with periodic updates, starting in early winter and continuing through late spring, generated by government agencies and other service-delivery organizations (SDOs; see Serafin et al., in preparation) having strict accountabilities around delivering timely and reliable information. Operational WSFs are required under treaties governing management of international rivers like the Columbia, Colorado, and Rio Grande basins; are stipulated in legal decisions, like Biological Opinions (BiOps) in the Klamath Basin; and are central input to engineering models and decision support systems used in optimal reservoir management for competing needs around flood control, agricultural and urban water supply, water-intensive industrial and technology-sector manufacturing, navigation, hydroelectric generation, and ecological flows. Operational WSFs also influence reservoir facility construction plans and guide choices around annual crop selection and amount of land left fallow, water rights rentals, and negotiation of forward contracts for hydropower, among other economic planning choices.

WSFs can have good skill despite poor weather forecast accuracy over the same seasonal-scale prediction horizons because of large lags between the overall annual cycles of meteorological forcing and watershed response in western North America. For most rivers here, flows peak in spring and summer, coinciding with peak water demand, and are driven mainly by melting of mountain snowpack accumulated the previous winter. In WSF practice, snowpack is measured and provided as a primary input to river hydrology models implemented and operated on a subbasin-by-subbasin basis. These models fall into two categories: process-simulation models that explicitly represent the underlying physics of watershed-scale runoff generation, and data-driven phenomenological models that account for the physics implicitly using empirical input–output mappings of predictors to predictands. A wide variety of specific models fall under these broad umbrellas, each with advantages and disadvantages (Singh and Woolhiser, 2002; Perkins et al., 2009; Gelfan and Motovilov, 2009; Bourdin et al., 2012; Weber et al., 2012; Cunderlik et al., 2013; Hrachowitz and Clark, 2017; Fleming and Gupta, 2020).

Even incremental improvements in WSF skill can provide well over $100 million per year in additional public benefit for a single river in the western US (Yao and Georgakakos, 2001; Hamlet et al., 2002). WSF improvements are also critically needed due to narrowing margins between increasing water demand under growing populations, and decreasing manageable water supply under climate change, which is reducing snowpack through warmer winter temperatures (e.g., Barnett et al., 2005; Clarke et al., 2015; BOR, 2016). This climate change-induced decline in the hydrologic role of snowpack in western US watersheds also reduces the inherent seasonal predictability of river runoff (Harrison and Bales, 2016; Harpold et al., 2020). The implications of these skill losses are reminiscent of biology's Red Queen hypothesis, which posits that evolutionary progress is required of a species to simply maintain its status relative to competitors. In effect, geophysical modelers are competing against climate change which, by decreasing predictability of seasonal runoff, forces continual forecasting innovation to maintain constant skill. It follows that skill improvements require even more aggressive advances. Indeed, increased water management flexibility is a leading goal in the Bureau of Reclamation's US West-wide climate change adaptation strategy, with improved hydrometeorological forecasting as a central element (BOR, 2016).

Collectively, these considerations have led to intense ongoing interest in improving WSF models in western North America. Related research directions are diverse; some examples include Garen (1998); Mahabir et al. (2003); Hsieh et al. (2003); McGuire et al. (2006); Wood and Lettenmaier (2006); Kennedy et al. (2009); Gobena and Gan (2009); Gobena and Gan (2010); Rosenberg et al. (2011); Gobena et al. (2013); Robertson et al. (2013); Fleming and Dahlke (2014); Demargne et al. (2014); Pagano et al. (2009); Pagano et al. (2004); Trubilowicz et al. (2015); Harpold et al. (2016); Najafi and Moradkhani (2016); Beckers et al. (2016); Mendoza et al. (2017); Lehner et al. (2017); Fleming and Goodbody (2019); and Peñuela et al. (2020).

### 1.2. NRCS water supply forecasting, the next-generation model, and machine learning

The US Department of Agriculture Natural Resources Conservation Service (NRCS) has been monitoring snowpack and predicting runoff in the western US since the Dust Bowl of the 1930s (Perkins et al., 2009). It operates the SNOTEL mountain climate and snow monitoring network, with over 850 sites across the region. Additionally, its current operational WSF platform is the largest stand-alone system regionally, and to our knowledge the largest data-driven system globally, with over 600 forecast locations in the Colorado, Missouri, Columbia, Rio Grande, Klamath, and other basins (Fig. 1).

NRCS uses several WSF models. The primary method is a probabilistic form of principal component regression (PCR), implemented in the NRCS VIPER software platform. It was adapted to WSF by NRCS to facilitate linear regression under predictor multicollinearity (Garen, 1992; James et al., 2013). PCR has since been widely adopted for operational WSF, and as a WSF modeling tool in hydrology, snow, and climate research (e.g., Moradkhani and Meier, 2010; Oubeidillah et al., 2011; Hsieh et al., 2003; Najafi and Moradkhani, 2016; Eldaw et al., 2003; Rosenberg et al., 2011; Gobena et al., 2013; Risley et al., 2005; Regonda et al., 2006a; Regonda et al., 2006b; Kennedy et al., 2009; Harpold et al., 2016; Lehner et al., 2017; Perkins et al., 2009; Beckers et al., 2016; Fleming and Goodbody, 2019; Glabau et al., 2020). Though successful, the technique is decades old, has known technical issues, and required revisiting for potential upgrades or replacement. A particular point of interest was potential adoption of machine learning (ML), a branch of artificial intelligence (AI; here we use ML and AI interchangeably for convenience) involving algorithms that detect patterns in data and use those patterns to make predictions.

However, developing an AI-based next-generation NRCS WSF model involved overcoming long-standing roadblocks to transitioning ML from research into a genuine operational river forecasting environment. AI was applied to streamflow prediction over 25 years ago (Hsu et al., 1995; Minns and Hall, 1996). Despite ongoing research demonstrating forecast skill improvements over statistical and process-based hydrologic models, migration to operational river hydrology has been limited, with no openly documented adoption of AI by operational WSF systems in the western US. Reasons include its black-box character and resulting lack of geophysical explainability, lack of emphasis on generating prediction uncertainty estimates, and concerns about overtraining (Abrahart et al., 2012; Fleming et al., 2015). Underlying these specific issues, there may be two broader questions. One is fundamental: some have recently argued that the ongoing evolution of machine learning has yielded substantially new and superior approaches to physically understanding hydrological systems, which the hydrologic community as a whole has not yet acknowledged or adjusted to (Nearing et al., 2021). The other is practical. In particular, it has been the collective qualitative observation of the authors, working with both AI and operational river forecasting over several decades, that these two communities of practice are largely disconnected: operational hydrologists typically have little familiarity with AI and often feel uncomfortable with it, whereas researchers specializing in ML applications to hydrology often focus on exploring the latest AI innovations rather than meeting the needs of operational hydrologists.

The approach taken, therefore, was to determine a holistic set of characteristics required of a next-generation WSF model, and work hand-in-hand with the operational community to craft an integrative solution meeting those specific requirements. Several well-proven techniques from AI, statistical modeling, evolutionary computing, ensemble modeling, and other areas were selected and combined to form this novel hybrid approach (Fleming and Goodbody, 2019), termed the multi-model machine learning metasystem ($M^4$).

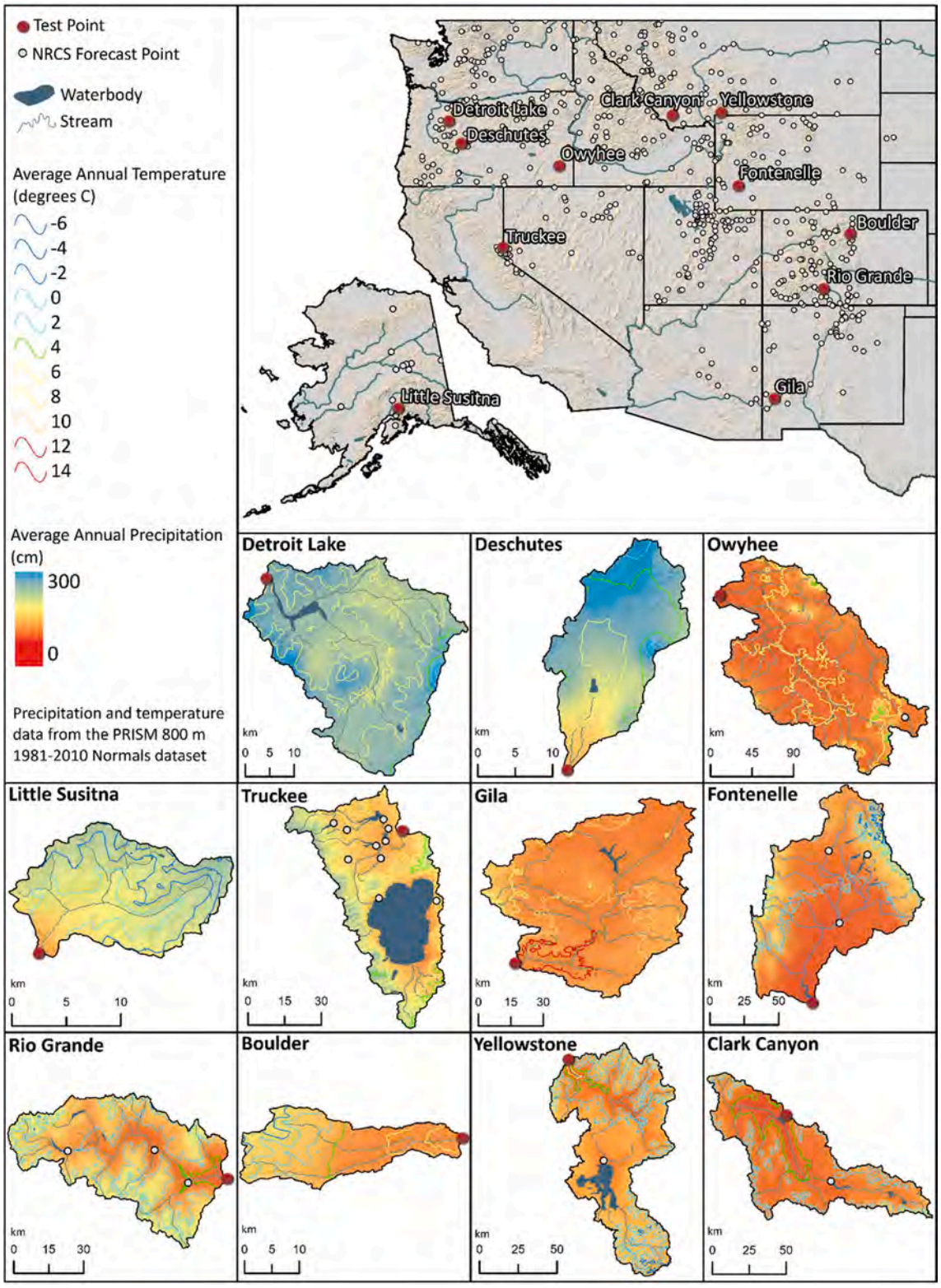


**Fig. 1.** Locations of all NRCS water supply forecast points, and selected test basins.

### 1.3. Study goals

In this study, we apply $M^4$ and evaluate its performance characteristics and suitability for widespread operational implementation. Despite technical vetting in the data science literature and initial demonstrations of hydrologic applicability (Fleming and Goodbody, 2019), WSF testing of $M^4$ so far has been too limited to justify broad operational adoption yet. Any next-generation NRCS operational WSF method must demonstrate applicability and performance advantages relative to existing operational systems over the wide range of geophysical environments encountered across the NRCS WSF system, which spans the deserts of New Mexico to the icefields of Alaska, and the associated diversity of statistical problem characteristics, data availability, and other practical factors. Further, a telling and necessary test of any prototype

prediction system is to run it "live" in the same operational environment it is ultimately intended to serve in. Given the aforementioned lack of uptake of ML by the operational community, such operational testing may also speak more widely (if indirectly) to overall suitability of AI for routine mainstream large-scale river forecasting, and in particular, whether the design philosophy and technical solutions used in developing $M^4$ are effective at bridging that research-applications gap.

To address these questions about the practical capabilities of $M^4$, we performed two sets of testing. First, hindcasting was completed for 20 test cases spanning 11 locations sampling diverse hydroclimatic settings. Second, live operational testing was performed for a subset of 5 of these locations during the 2020 forecast season.

Though accuracy improvements are vital, acceptance by operational forecast hydrologists additionally requires assessment of a broader set of performance characteristics (e.g., Weber et al., 2012; Cunderlik et al., 2013; Whateley et al., 2015; Fleming and Goodbody, 2019; Peñuela et al., 2020; Fleming et al., 2021). Other questions evaluated include effectiveness of efforts to build high levels of robustness, flexibility, and automation into the new approach; logistical feasibility of associated workflows and computations in a time- and resource-constrained setting typical of agencies that run such forecast systems operationally; consistency of performance capabilities across a variety of watershed characteristics, without the need for manual case-by-case local-scale fine-tuning of modeling procedures; and physical plausibility of outcomes, including both generation of physically reasonable predictions, and amenability of the resulting forecasts to interpretation in terms of current seasonal climatic conditions and known hydrologic processes.

The influence these additional considerations have on which technologies are adopted into operational systems should not be underestimated, nor is it unique to NRCS. Forecast software platforms at operational agencies have been recently developed or updated, like VIPER at NRCS (see above), the Hydrological Ensemble Forecast System (HEFS) at National Weather Service River Forecast Centers, and PyForecast at the Bureau of Reclamation (e.g., Demargne et al., 2014; Perkins et al., 2009; https://github.com/usbr/PyForecast). Nevertheless, fundamental geophysical modeling concepts (Fleming and Gupta, 2020) underlying these platforms have not seen major upgrades in decades (Hartmann et al., 2002; Pagano et al., 2004). HEFS uses the 1970s-era Sacramento Soil Moisture Accounting (SAC-SMA) and SNOW-17 process models, and while the modular PyForecast platform has flexibility to easily incorporate innovative modeling techniques, both VIPER and current development versions of PyForecast largely focus on 1990s-era PCR and other linear regression variants. Moreover, model implementation and operation generally remain reliant on an archaic style of subjective manual hydrologist intervention. Research implementing a so-called over-the-loop (OTL) paradigm using automated and objective processes has not, in general, successfully migrated to large-scale mainstream river forecast systems in the region (Seo et al., 2003; Wood et al., 2020). Persistent use of seemingly outdated but proven methods, which has been interpreted by some as technical stagnation (e.g., Hartmann et al., 2002), has occurred largely because new methods, whether physics-based or data-driven, have often failed to match key needs of the operational hydrology community (for detailed discussions see, e.g., Weber et al., 2012; Cunderlik et al., 2013; Whateley et al., 2015; Fleming and Goodbody, 2019; Peñuela et al., 2020). As one example, benefits provided by AI are significant but accompanied by drawbacks restricting operationalization (see Section 1.2).

The $M^4$ approach, and the pragmatic testing regimen for it presented here, are intended to address these roadblocks to migrating OTL and AI-based methods into operational WSF. Accordingly, evaluations are completed here within the context of the existing NRCS WSF system, as described in detail below. Doing so leverages experiential knowledge and established best practices around operational WSF in the US West and enables meaningful comparisons to current operational techniques.

### 1.4. Manuscript organization

The manuscript is organized as follows. Section 2 summarizes the established overall framework for data-driven operational WSF models in western North America, which as noted above we adopt here to ensure apples-to-apples comparisons of $M^4$ against current methods. It then presents specific test cases and datasets considered. Section 3 provides a brief qualitative summary of $M^4$, focusing on linkages to operational forecast community needs. This short synopsis is not intended to be comprehensive, and readers are referred to Fleming and Goodbody (2019) for detailed technical descriptions of $M^4$. Section 4 summarizes the results of hindcast and live operational testing. This includes discussion of performance metrics with comparisons to the existing system, geophysical interpretability of results, and other topics. Broader implications to acceptance of AI in operational hydrology are also briefly discussed. Finally, Section 5 concludes with a summary and outlines research and operationalization plans. Note that to our knowledge, after full roll-out into production systems at NRCS, $M^4$ will be the largest migration of AI into a genuine operational river forecast environment to date.

## 2. Data

### 2.1. Standard WSF problem structure

To evaluate $M^4$ in a realistic operational WSF context, we set up the overall prediction problem in a manner closely resembling the existing NRCS forecast system, which is in turn broadly similar to most other statistically based operational WSF models in western North America. The predictand (target, in ML nomenclature) is spring-summer runoff volume, which is measured at a US Geological Survey streamgage with NRCS adjustments as needed for upstream diversions, or at some other hydrometric monitoring site. Predictors consist of snow water equivalent (SWE) and wintertime-to-date accumulated precipitation measurements at mountain climate monitoring stations, predominantly NRCS SNOTEL or similar sites. Various other datasets, like antecedent streamflow, are sometimes used as supplemental predictors. Note that WSF research has extensively tested additional data types, like remotely sensed SWE, gridded precipitation datasets, seasonal-scale numerical climate model forecasts, and other products, but so far these experimental predictors have experienced limited uptake into operational WSF systems in the western US.

To illustrate, a typical statistically based WSF model might predict, on March 1, the upcoming April 1–July 31 cumulative flow volume at a given point on a given river, using as predictors March 1 SWE and October 1-February 28 total precipitation measured at SNOTEL sites within or near the watershed boundary upstream of the streamgage. The number of such sites varies widely depending on the basin, but about a half-dozen to two dozen is roughly typical. As part of the modeling process, the input datasets are usually amalgamated in some way into an index that serves as the regression predictor, or in the ML nomenclature, a feature that is presented to the supervised learning algorithm.

### 2.2. Test cases

Hindcast testing considered 20 test cases corresponding to January 1 and April 1 forecast dates at 11 existing NRCS forecast locations (Fig. 1). These locations span diverse geophysical environments, including a glacier-fed Alaskan river, several southwestern desert rivers, a watershed with large volcanic aquifer contributions to flow, a comparatively winter rain-dominated Pacific Northwest basin, a Sierra Nevada snowpack-fed endorheic California-Nevada watershed, Missouri subbasins in both the northern and southern Rocky Mountains, Colorado River and Rio Grande headwaters, and so forth (see Table 1 for summaries). These test cases also sample diverse statistical characteristics, like nonlinear functional forms and heteroscedastic and non-normally

**Table 1**
Summary of test-case forecast points drawn from the existing NRCS operational WSF system (see Section 2). Target was April-July volume unless otherwise noted in the table. For hindcasting (see Section 3.3.1 for details), forecast issue dates were January 1 and April 1 for each location unless otherwise specified below; the combination of 11 locations, and two forecast dates for all but two of those locations, gives a total of 20 hindcasting test cases. Five of these forecast points were also used for live operational testing (see Section 3.3.2 for details) in addition to hindcasting, and these are also identified below; for these five locations, models were developed and run operationally on January 1, February 1, March 1, and April 1, 2020.

| Name | USGS ID | Description |
|---|---|---|
| Truckee River at Farad | 1034600 | Endorheic desert watershed within the Great Basin, fed by abundant upstream spring snowmelt from California's moist Sierra Nevada. It forms the outlet of Lake Tahoe and terminates in Pyramid Lake. Peak flows occur on average in May, following peak snow accumulation rates in January and February; there is little rainfall input. Downstream from this gage, the Truckee is the source of drinking water for Reno. It is additionally used for irrigation and hydroelectric power generation, and the US Fish and Wildlife Service uses reservoirs on Truckee tributaries above Farad to manage endangered fish species. Included in both hindcast and live operational testing. |
| Yellowstone River at Corwin Springs | 06191500 | Major tributary of the Upper Missouri River, forming a northwestern headwater basin to the larger Mississippi Basin. At this gage, it is a cold, moderately wet, partially mountainous basin draining parts of Wyoming and Montana; the continental divide forms its western watershed boundary. Upstream it is a centerpiece of Yellowstone National Park, and downstream it is used for irrigation water supplies; tourism and recreation are significant values. It experiences a distinct flow peak, occurring on average in June, driven overall by April-May snowmelt and a May-June peak in rainfall. Included in both hindcast and live operational testing. |
| Owyhee River near Rome | 13181000 | Tributary to the Snake River, eventually contributing flows to the mid-Columbia Basin. At this gage, it is a semi-arid basin covering a mixture of mountainous and plateau areas in the inland Pacific Northwest and parts of the US desert southwest, spanning Nevada, Idaho, and Oregon. Its peak flows occur over a relatively wide freshet spanning February through June, peaking on average in March. The flow regime is largely driven by spring snowmelt and spring rain. The Bureau of Reclamation (BOR) operates a dam on the Owyhee to provide irrigation water for regional agriculture. Included in both hindcast and live operational testing. |
| Rio Grande near Del Norte | 08220000 | At this headwater location in southern Colorado, the Rio Grande is a moderately wet to semi-arid basin in the San Juan Mountains, driven primarily by spring snowmelt with relatively minor summer rain inputs, and peak flow typically occurs in May or June. Downstream, the Rio Grande receives Colorado River water through the BOR San Juan-Chama Project, it forms much of the US-Mexico border, and it provides municipal and agricultural water supplies across Colorado, New Mexico, Texas, and northern Mexico before emptying into the Gulf of Mexico. The target period for the January 1 and April 1 hindcast test cases is April-September, reflecting existing NRCS practice at this location. |
| Deschutes River below Snow Creek | 14050000 | Major tributary to the middle reach of the Columbia River. At this headwater gage, it is a very wet mountain basin draining the summit and east side of Cascade Range. Its water budget is driven by late-spring (May-June) snowmelt derived from a strong winter precipitation peak, but its flows here are strongly modified by unusually strong groundwater-surface water interactions in the extremely porous volcanic aquifers of the Oregon Cascades, leading on average to a late-summer (July-September) discharge peak. Dams and diversions on the Upper Deschutes provide agricultural and municipal water supplies, and the river has significant recreational values. Included in both hindcasting and live operational testing. |
| Gila River near Gila | 094305000 | Tributary to the Lower Colorado River. The continental divide forms its eastern watershed boundary. At this upstream location, it is a semi-arid mountain river draining the Mogollon, Pinos Altos, and Black Ranges of southwestern New Mexico; flows are driven by late winter-early spring mountain snowmelt normally peaking around March, and summer (North American Monsoon) rain typically peaking around July or August. The Upper Gila is relatively pristine, but downstream its flow is heavily diverted for agricultural and municipal water supplies and also supplemented using Colorado River water through the Central Arizona Project. Included in both hindcast and live operational testing. Unlike most other test cases, the target periods were January-May, February-May, March-May, and April-May, respectively, for the January 1, February 1, March 1, and April 1 forecast dates, reflecting existing NRCS operational practice at this location, which in turn reflects established local water management information needs. |
| Beaverhead River inflow to Clark Canyon Reservoir | 06015400 | A cold, moderately wet mountain watershed in southwestern Montana; the continental divide forms its western and southern watershed boundaries. A Missouri River headwater basin, its flows are driven primarily by spring snowmelt with augmentation by early summer rain, and on average, peak discharge occurs around April. BOR operates the Clark Canyon Dam for irrigation and downstream flood control. |
| Little Susitna River near Palmer | 15290000 | A mountainous, subarctic, maritime, glacier- and snow-fed river that flows from the Talkeetna Mountains to the Gulf of Alaska near Anchorage. Largely a wilderness river above this gage, it has significant fisheries and recreation |

(*continued on next page*)

**Table 1** (*continued*)

| Name | USGS ID | Description |
|---|---|---|
| | | values, and it is the only test case to include major upstream glacial cover. Reflecting current NRCS operational practice in Alaska, in turn reflecting region-specific water management needs, there is no January 1 WSF publication date for this location. |
| Boulder Creek near Orodell | 06727000 | A tributary to the St. Vrain River, contributing flow in turn to the South Platte, Platte, Missouri, and Mississippi Rivers. At this gage, it is a moderately wet mountain basin lying within the Front Ranges of the Colorado Rockies; the continental divide forms its western watershed boundary. Its flows show a sharp peak in late spring-early summer, typically reaching a maximum in June driven by May-June snowmelt and summer rain events. It also contains a small amount of glacial ice in its headwaters, augmenting late summer flows, and it is a water supply source for the city of Boulder. The January 1 publication date was not considered in hindcasting due to technical issues with the existing benchmark model (Section 3.3.4) at this location. |
| North Santiam River inflow to Detroit Dam | 14181500 | A tributary to the Santiam River and in turn the Willamette River, a major tributary to the Columbia River at its confluence in Portland. It is a very wet mountain basin running from the west slope of the Cascade Range, and it is dominantly winter rain-fed with secondary spring snowmelt. Detroit Dam is operated by the US Army Corps of Engineers for a variety of uses, including flood control, hydroelectric power generation, irrigation, fisheries, and recreation. Downstream of the dam the North Santiam provides drinking water to a number of communities, including the Oregon state capitol of Salem. The target period is April-June. |
| Green River inflow to Fontenelle Dam | 09211150 | A major headwater tributary to the Upper Colorado Basin. At this gage, it drains the Wind River Range, Wyoming Range, and a large plateau area lying between them, which range from moderately wet to semi-arid. Flows at this location show a broad peak between May and July, resulting from spring snowmelt and rain. BOR operates the Fontenelle Dam as a storage reservoir for the Colorado River Storage Project and for assertion of Wyoming's Colorado River water rights, for hydroelectric power generation, and for other uses. |

distributed errors (e.g., Owyhee River), linear stationary, Gaussian behaviors (e.g., Yellowstone River), and multiple predictive inputs corresponding to both wintertime hydroclimate and complex internal watershed dynamics (e.g., Deschutes River).

In addition, live operational testing was undertaken during the 2020 forecast season for a subset of 5 of these locations (Gila, Deschutes, Yellowstone, Owyhee, and Truckee rivers) trained at multiple consecutive forecast issue dates (January 1, February 1, March 1, and April 1). This forms a total of 20 live operational test cases that partially overlap with the 20 hindcasting test cases.

As noted in Section 2.1, to facilitate meaningful comparisons against current systems, the method was implemented in a fashion, including dataset selection, similar to existing statistical operational WSF models. For each test case, spring-summer runoff volume over the river's established primary management period, usually April-July, was the target. A candidate pool of input variables was assembled to serve as the basis for feature extraction for each test case. Specific variables were closely consistent with existing operational NRCS models for the same combination of location, issue date, and target period (forecast horizon), and include year-to-date precipitation and current snow water equivalent (SWE) at the forecast date at NRCS SNOTEL or related (e.g., California Cooperative Snow Survey program) mountain climate monitoring sites, and for certain basins, antecedent streamflow. We did not capitalize here on emerging alternative predictor types like remotely sensed or climate modeling products, although the method is designed in part with those predictors in mind as discussed below; the result therefore reflects a minimum estimate of potential advantages of the method.

Again following typical procedure for statistically based operational WSF in the western US, we used data over a standard 30-year hydroclimatic normal period (1986–2015). This choice usually reflects a pragmatic attempt to balance longer datasets for better training and testing vs. record length limitations that could restrict the number of sites for which models can be developed. A common secondary motivation is to help mitigate impacts from various climatic and land cover nonstationarties on model development, by restricting the record length used to a recent period which is reasonably representative of current conditions and over which the cumulative impacts of nonstationarities can be reasonably presumed, at most forecast locations, to be sufficiently limited for developing and testing a seasonal-scale prediction model.

## 3. Method

### 3.1. General

As noted in Section 1, this study uses the recently developed $M^4$ prediction analytics engine. Mathematical and computational details of $M^4$ are too lengthy to be repeated here; in the interest of conciseness, readers are referred to Fleming and Goodbody (2019). Instead, Section 3.2 provides a brief qualitative synopsis of the model, focusing on summarizing how specific operational WSF needs were identified and what design steps were taken in an effort to meet those requirements. Section 3.3 then describes application of this method to the test cases outlined in Section 2.

### 3.2. Summary of $M^4$ prediction engine

#### 3.2.1. Design process, concept, and criteria

Preparatory steps during initial $M^4$ development included a thorough inventory and assessment of needs and options before undertaking detailed technical design. The process began with documenting the existing NRCS system, which like many operational WSF systems has evolved organically over the decades. Its capabilities and limitations were then assessed, including documentation of known issues discovered during forecast operations over the years, and completion of extensive statistical diagnostics. Progress in data-driven WSF was reviewed with an eye to identifying the most promising potential directions for a next-generation NRCS model. Implications of global anthropogenic climate change on requirements for WSF were additionally considered. Finally, an initial blueprint and preliminary scoping models were developed to test potential ideas. The overall conclusions were that a new WSF model was warranted, that AI was the best solution pathway for NRCS, and that for AI to be effective and useful in a practical operational hydrology context, it must be deployed in a highly application-specific way (similar conclusions have been reached in other fields, e.g., Meredig, 2018).

The resulting design concept was framed as a convergence of detailed hydrologic process knowledge, a mature understanding of potentially applicable machine learning concepts and tools, and

practical operational water management requirements (Fig. 2, top). This led in turn to identification of 9 specific design criteria (Table 2). Several existing techniques met some of these criteria but none adequately satisfied all. A hybrid approach was therefore developed (Fleming and Goodbody, 2019) specifically to meet these design requirements (Fig. 2, bottom; see Section 3.2.2).

#### 3.2.2. Overview

Fig. 2 sketches the main elements of $M^4$. In summary, the metasystem consists of six semi-independent forecast systems, each centered around a different supervised learning algorithm. A pool of candidate input predictor variables is defined by the hydrologist for a given forecast problem (combination of forecast location, issue date, and target period), similar to current-generation statistical WSF models in western North America (Section 2). Principal component analysis (PCA), an unsupervised statistical learning technique, is used for pattern recognition in the input data matrix, and the extracted features are directed to the supervised learning models. Supervised learners include two substantially different kinds of neural network (a monotone artificial neural network, mANN, and a monotone composite quantile regression neural network, MCQRNN), linear regression (LR), quantile regression with adjustments to ensure non-crossing quantiles (QR), random forests (RF), and a support vector machine (SVM; more specifically, support vector regression). A genetic algorithm (GA) optimizes feature extraction and selection separately for each forecast system, that is, which of the input predictor variables in the candidate pool and corresponding PCA modes to retain. The GA objective function uses a penalty to ensure that multiple predictors are retained, which is operationally important for functional redundancy in the event of sensor failures or other technical issues. The result of each semi-independent system is a forecast distribution, i.e., a probability density function for future seasonal flow volume. In standard operational WSF practice in the western US, this uncertainty information is presented as prediction intervals, corresponding to 0.1, 0.3, 0.7, and 0.9 quantile (respectively, 90, 70, 30, and 10% exceedance probability) flow volumes. $M^4$ generates these intervals using either intrinsic probability models for the two quantile regression techniques (MCQRNN and QR) or a Box-Cox transform space-based heuristic for the other models; both are nonparametric methods that accommodate heteroscedastic and non-Gaussian error distributions. Results are averaged across the models to form an ensemble mean best-estimate with prediction intervals. Algorithmic logic is introduced at various points to automate various processes, including but not limited to hyperparameter optimization, or to ensure certain conditions are met by the solution, like an ensemble-pruning algorithm contributing to strictly nonnegative predictions.

Fleming and Goodbody (2019) provide important methodological details around regularization (essentially, overtraining mitigation), hyperparameter selection and tuning (pre-calibration and optimization of high-level machine learning parameters), prediction interval generation, optimal feature extraction, further information on each of the constituent statistical and machine learning models and how they are implemented within $M^4$, and other key $M^4$ technical elements. In the interest of conciseness readers are referred there for complete descriptions of the data science underlying the $M^4$ prediction analytics engine assessed in this study.

That said, we briefly elaborate below on two aspects that warrant particular attention from a more general operational hydrologic modeling perspective: explainable AI, and autonomous machine learning. These two concepts and the pragmatic approaches used to achieve, or at least adequately approximate, them for our purposes are summarized in Sections 3.2.3 and 3.2.4, respectively. Additionally, all the methods used to collectively form $M^4$ were specifically chosen to help satisfy particular requirements outlined in Table 2; some related points around technique selection are briefly summarized in Section 3.2.5.

#### 3.2.3. Geophysical consistency and explainability

We borrow the term "physics-aware AI" from materials science (Meredig, 2018). It is a broad but useful term that refers to general intersections between ML applications and the underlying process physics, or qualitative domain-specific knowledge more broadly, of the problem to which ML is applied. It is a holistic concept primarily focusing on making AI useful to mainstream science and engineering users through such mechanisms as explainable machine learning, theory-guided data science, and a general alignment of machine learning solutions with existing bodies of physical knowledge of the system being modeled and associated practical considerations like the inherent data-sparseness of certain fields. Many specific technical approaches fall under this rubric; moreover, given the centrality of explainable machine learning to the future of AI, it is an extremely dynamic computer science research topic in which new paradigms emerge on a regular basis, though many of these are far from ready for use in high-stakes operational river forecasting systems at SDOs.

We adopted a comparatively straightforward, strongly pragmatic, and WSF-focused approach that concentrates on balancing two $M^4$ design criteria: generating forecasts that are physically reasonable and explainable (criterion 6 in Table 2) while using well-proven ML algorithms (criterion 7). Three general steps were taken. (1) Automation notwithstanding, features engineering in $M^4$ remains hydrologist-directed through input candidate pool selection and decisions around the maximum number of PCA modes the genetic algorithm is permitted to retain. These choices reflect end-user knowledge around representativeness, reliability, quirks, and capabilities of potential input variables or measurement sites, and geophysical interpretations of PCA modes, which in practice are known to usually correspond to watershed-scale indices of hydroclimatic forcing or aquifer-stream interactions (e.g., Garen, 1992; Fleming and Goodbody, 2019; see also Section 4). It is a key location in the AI development process for domain experts to insert physical hydrologic knowledge. That a hydrologist should select candidate predictors may seem obvious to water resource scientists and engineers but runs contrary to some contemporary AI directions, like certain data-mining and big-data applications. (2) WSF was reframed as a low-dimensional problem with a parsimonious solution. This is not, in itself, physics-aware AI. However, PCA input pre-processing and compact ML architectures enable direct graphical visualization of input–output relationships in most cases, revealing relationships the AI discovered, and facilitating their geophysical interpretation. Criterion 9 therefore supports criterion 6. (3) Monotonicity and nonnegativity constraints are imposed at various locations within $M^4$. This includes selection of specific machine learning methods allowing nonnegativity (MCQRNN) and monotonicity (mANN, MCQRNN) constraints; inclusion of nonlinear supervised learners into the multi-model ensemble that can track nonlinear relationships and thus further contribute to avoidance of negative-valued predictions (mANN, MCQRNN, RF, SVM; see Section 4 for an example of such nonlinearities, and how $M^4$ captures and communicates them); conversely, inclusion of linear supervised learners as a small subset of the multi-model ensemble to further contribute to monotonicity of the final ensemble solution (QR, LR); careful and application-specific regularization-related hyperparameter pre-calibration steps (see also Section 3.2.4) that enable but place limits on nonlinearity; and a final ensemble-pruning algorithm to further enforce non-negativity, as mentioned above (see also Fig. 2). These functional characteristics of monotonicity and nonnegativity correspond to known aspects of hydroclimatic relationships – for example, that runoff volume cannot be negative-valued, or that a heavy snowpack does not, all else held equal, lead to low flow volume. Again, for details see Fleming and Goodbody (2019).

Most of these steps constitute theory-guided AI, that is, a priori alignment of AI algorithms with known geophysical processes (see Karpatne et al., 2017). Doing so in turn encourages geophysical explainability; examples are discussed in Section 4. Additional regularization is an added benefit, as theory-guided a priori constraints limit

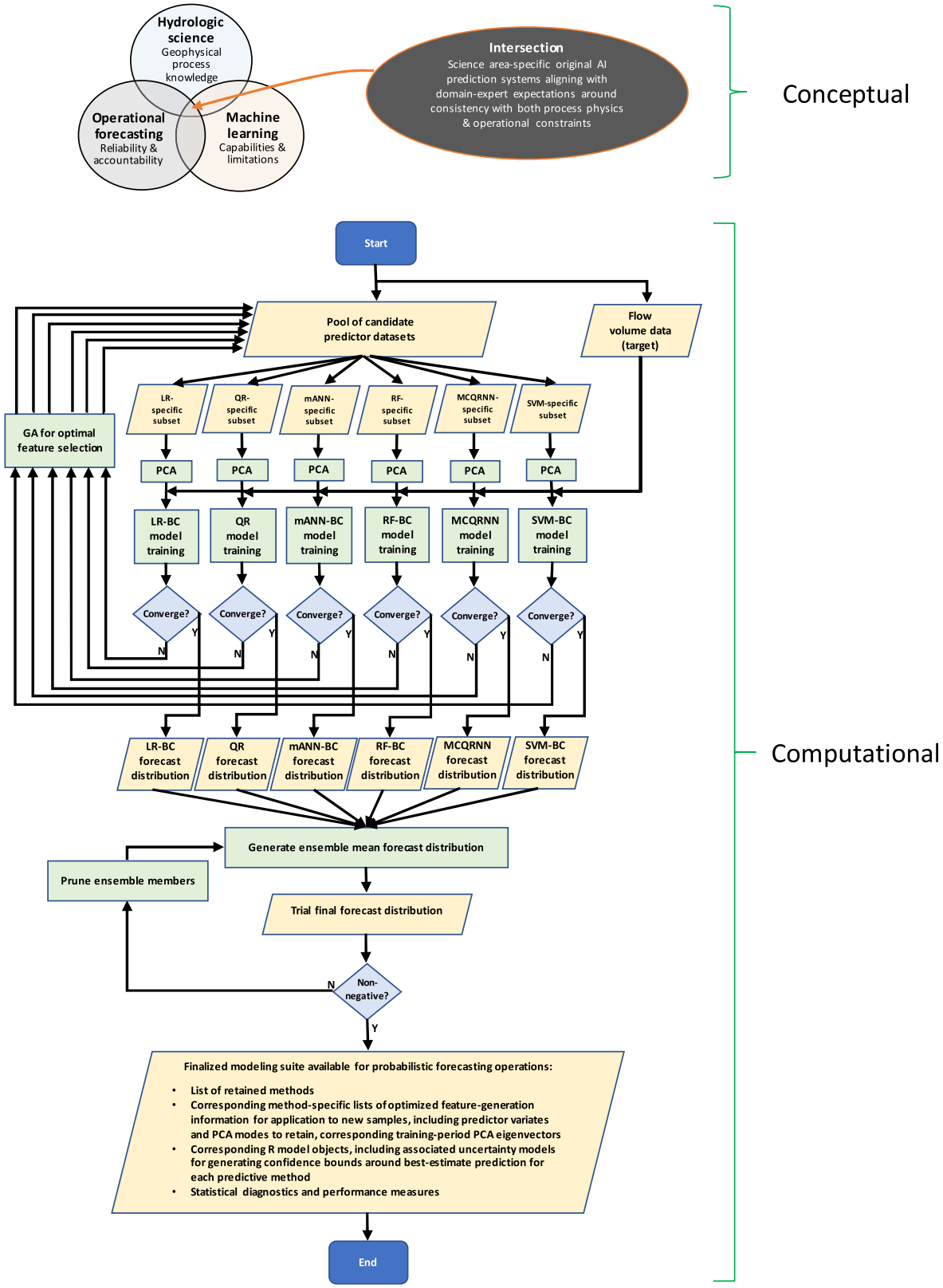


**Fig. 2.** Simplified schematic representation of M[4]. Operational acceptance of AI-based WSF requires a broad three-way convergence (top), giving specific design attributes (see Table 2). Process map (bottom) illustrates main components. PCA: principal component analysis; LR: linear regression; QR: quantile regression; mANN: monotone artificial neural network; RF: random forests; MCQRNN: monotone composite quantile regression neural network; SVM: support vector machine; BC: Box-Cox transform; GA: genetic algorithm. See text and Fleming and Goodbody (2019) for details.

**Table 2**
Design criteria. See text and Fleming and Goodbody (2019) for acronyms and further details. Criteria were determined through sustained dialog between model developers and model users at NRCS. While some of these criteria are specific to the NRCS operational environment or to machine learning applications in hydrology, overall the requirements dovetail closely with factors that have previously been identified as crucial for development and operational adoption of new river prediction modeling technologies. Examples of such intersections include suitability matrix concepts, integrating multifaceted performance measurement suites around both simulation accuracy and operational logistics, as demonstrated for hydrologic modeling by Cunderlik et al. (2013); and the concepts of relative advantage, complexity, compatibility, trialability, and observability within the diffusion-of-innovations framework introduced by Rogers (2003) and adapted to seasonal hydroclimatic forecasting by Whateley et al. (2015).

| Criterion | Explanation |
|---|---|
| 1. Improved forecast accuracy | Improved WSF accuracy has deep social, economic, and environmental value in the region, particularly as population growth and climate change narrow margins between supply and demand |
| 2. Improved potential for automation | Required for an AI-based system operated by non-AI experts; needed for more frequent WSF updating going forward; dovetails with objective "over-the-loop" hydrometeorological prediction concepts; aligns with democratization (Hill et al., 2016) of ML use |
| 3. Relatively low cost & good ease of development, implementation, and operation | Logistical, including computational, feasibility: user- or hardware-intensive systems present practical hurdles with transition to, and operation of, a new WSF system; reliability is a key operational need facilitated by relatively straightforward, robust designs |
| 4. Seamlessly address known technical issues | Predictions must accommodate nonlinear functional forms, uncertainty intervals must accommodate heteroscedastic and non-Gaussian residuals, and predicted volumes must be nonnegative, without slow and subjective user intervention (e.g., transforms) |
| 5. Modular & expandable | Crucial for avoiding obsolescence of, and therefore protecting investments in, any operational forecasting platform, particularly for a modeling system using AI, which is a rapidly evolving field |
| 6. Geophysical consistency & explainability | Must overcome nominal black-box limitation of AI: forecasts and models must be guided by hydrologic theory and easily interpreted in terms of known hydroclimatic processes; a relatable hydrologic 'storyline' around the forecast is mandatory for operational WSF |
| 7. Balance innovation & performance gains vs. established building blocks & proven tools | Transitioning OTL AI-based technology into operational WSF requires bridging distinct professional cultures, and balancing new and old; development process therefore adopted a MAYA (most advanced yet acceptable) design principle (e.g., Hekkert et al., 2003) |
| 8. Multi-model ensemble framework | Address equifinality and model selection uncertainty present in all hydrologic modeling; ensemble of methods having substantially different properties may improve reliable metasystem function for diverse geophysical environments across the US West |
| 9. Dimensionality reduction & extraction of multiple independent input signals | High-dimensional collinear input data matrices are currently common in operational WSF and will only grow more prevalent in the future, with spatially distributed inputs like remote sensing, seasonal numerical climate model, and snow model products |

the solution space available to the machine learning algorithm and therefore reduce fitting to noise (e.g., Karpatne et al., 2017; Zhang and Zhang, 1999). More broadly, these elements of the $M^4$ metasystem may intersect with another high-level evolutionary trajectory in machine learning: the transition from first-wave (handcrafted knowledge), to second-wave (statistical learning), to now-emerging third-wave (contextual learning) AI (for a synopsis see Launchberry, 2021). Kratzert et al. (2018) and Fleming and Goodbody (2019) might be considered early attempts at integrating third-wave AI philosophies and capabilities into hydrologic prediction but approach this challenge differently. Kratzert et al. (2018) explore potential abilities of particular deep learning architectures to capture and reveal certain hydrologic processes with minimal or no application of a priori geophysical knowledge; that is, early experiments suggest this approach can learn some geophysical context without significant a priori guidance. $M^4$ instead moves more incrementally toward third-wave approaches, by leveraging second-wave and certain aspects of still-relevant (Launchberry, 2021) first-wave AI approaches to achieve specific operational goals in practical water resource science that necessarily include the application of machine learning within a predefined, but broad, geophysically relevant solution space, which is then refined by $M^4$ on a case-by-case basis. That is, it learns within a broad set of theory-guided constraints (monotonicity, nonnegativity, etc.) and capabilities (subject matter expert-guided features engineering, genetic algorithm-based feature selection, etc.) that reflect overall geophysical context of the general problem class (western North American WSF), and in doing so, it refines that geophysical context through the hydroclimatically interpretable solutions it learns and communicates for each river (see Section 4.2.2 for examples).

#### 3.2.4. *AutoML and pre-calibration*

Improved automation (criterion 2 of Table 2) was achieved by judicious and application-specific use of autonomous machine learning (AutoML in the data-science nomenclature, e.g., Thornton et al., 2013; Guyon et al., 2015) and pre-calibration. Algorithms were developed to automate most optimization and decision points, including setting ML hyperparameters. One example is automated determination of optimal ANN hidden-layer size on the basis of cross-validated goodness-of-fit and information theoretic metrics (Fleming and Goodbody, 2019). In other cases, hyperparameters were set to robust default pre-calibrated values based on extensive experimentation (Section 3.2.1) using a subset of WSF test cases. One example was completing, during initial prototype testing, hundreds of $M^4$ training runs to locate generally usable defaults for the population size and number of generations in the genetic algorithm, and defining a user protocol for diverging from those defaults if desired.

In general, this operational WSF-specific combination of AutoML algorithms and default hyperparameters involved establishing reasonable trade-offs. For instance, increasing the population size and the number of generations in the genetic algorithm improves out-of-sample prediction skill, but the relationship is nonlinear and quickly reaches a point of diminishing returns; balancing this observation against minimization of run times helped set default values. Similarly, in creating AutoML algorithms to automate optimal ANN topology selection, the better (up to a point) pattern recognition capabilities enabled by additional hidden-layer neurons were balanced against associated training complications, like longer run times and greater susceptibility to both overtraining and, conversely, trapping in local minima in the nonlinear neural network cost function (e.g., Hsieh, 2009).

AutoML is additionally facilitated in $M^4$ by selecting methods that match the statistical and physical requirements of western US WSF, i.e., criterion 4 supports criterion 2 (Table 2). Specifically, nonlinear AI techniques, heteroscedastic and non-Gaussian prediction intervals, and enforcement of non-negativity constraints reduce the amount of manual hydrologist intervention (in the form of selecting and applying predictand transforms, for instance) required to develop and operate WSF

models (Sections 1.3 and 3.2; see also examples in Section 4).

#### 3.2.5. Some relationships to other AI philosophies

As noted above, the goal of $M^4$ is to fuse existing methods into a hybrid that meets the specific design criteria we defined for a next-generation AI-based US West-wide operational WSF model at NRCS. This practical, applications-focused design philosophy is not generally typical of AI research in water resource science and engineering (see again Section 1), and to help illustrate how it motivates certain methodological choices in $M^4$ development, it is helpful to consider a few examples.

One example is how design criterion 8 was approached. Model equifinality, and using multi-model ensembles to address it, are well-established hydrologic concepts but are usually reserved for process-simulation models (e.g., Beven and Binley, 1992; Bourdin et al., 2014). Similar concepts have emerged independently in the statistical and AI communities but usually involve ensembles of nearly identical models, such as committees formed from bootstrap aggregated (bagged) neural networks (e.g., Breiman, 1996; Breiman, 2001; Wolpert, 1996; Burnham and Anderson, 2002). In contrast, strong methodological diversity within a multi-model ensemble is desirable, reducing error correlation across constituent models and increasing noise suppression within the ensemble (e.g., Monteleoni et al., 2011). A complementary concept is that, at a geophysical level, we might expect certain models to work better in certain environments, yet the final WSF system must function well across a wide range of hydroclimatic settings and terrestrial hydrologic processes (Sections 1 and 2). This implies that model diversity within the multi-model ensemble might improve consistency of prediction accuracy across the NRCS system, insofar as poor performance of one or more of the models in some particular location may be compensated by good performance of other models in the ensemble, and vice versa at other locations. This is intimately related to the underlying reasons why multi-model ensembles are effective in geophysical modeling (for details see Hagedorn et al., 2005). We therefore chose several fundamentally different supervised AI methods to include in the multi-model ensemble (Fig. 2). For instance, ANNs imitate the brain's biological network of neurons and synapses, RFs reflect the decision-tree framework of human choice, and SVMs are abstract hyperdimensional mathematical constructs. (As a corollary, $M^4$ constitutes a super-ensemble with respect to some of its constituent AI methods that are in turn ensemble learners, i.e., random forests and bagged neural networks.)

As another example, that a method is not currently chosen for $M^4$ does not imply we are critical of it, only that it is not presently judged to adequately satisfy the multi-faceted and operations-oriented design criteria (Table 2) or other basic fitness-for-purpose (see, e.g., Cunderlik et al., 2013) considerations in NRCS WSF. Consider, for instance, deep learning, sequential online learning, and transfer learning. These three branches of AI show substantial promise in hydrologic testing (e.g., Lima et al., 2017; Jiang et al., 2018; Kratzert et al., 2018). However, to our knowledge none has experienced even initial WSF testing (failing criterion 7 at this time). Additionally, deep learning is often computationally expensive to the point of imposing restrictive hardware requirements (potentially failing criterion 3), and while deep learning appears to offer strong knowledge-discovery capabilities in some geoscientific applications (e.g., Reichstein et al., 2019; Nearing et al., 2021), it remains unclear whether deep-learning architectures, which by definition are complex, are amenable to fast and easy geophysical interpretation of the type needed in our operational forecasting application (potentially failing criterion 6; see also Section 4.2.2). Transfer learning is potentially advantageous to building models for many forecast sites and dates, but $M^4$ can be trained rapidly due to AutoML/pre-calibration (Section 3.2.4) and some simple parallelization across processor cores on a standard personal computer (see above and Fleming and Goodbody, 2019). Further, hydrologic experimentation so far with these three classes of AI has largely focused on hourly or daily flood prediction, which has different requirements from seasonal volume forecasting. For instance, deep learning was developed largely for big data, whereas standard operational WSF problems generally involve modestly sized datasets (see Section 2) that may not effectively capitalize on, or even be adequate for, deep learning. Similarly, online sequential learning has advantages in cases of rapid new data acquisition, but in WSF only one new sample appears per year (Section 2).

By the same token, however, AI and its hydrologic applications are fast-evolving fields encompassing many existing, and emerging, techniques. This rapid development pace motivates design criterion 5: a modular, expandable structure that facilitates integrating new AI methods into $M^4$ (Table 2).

### 3.3. Application to test cases

#### 3.3.1. Retrospective analysis

Hindcast testing was performed for each of the 20 test cases described in Section 2. Five metrics, described immediately below, were used to measure hindcast performance. Several additional, more qualitative, evaluation criteria were also considered as discussed in Section 3.3.3.

Root mean square error (RMSE) and coefficient of determination ($R^2$) quantify deterministic prediction accuracy. RMSE provides an intuitive sense of typical predictive error and is closely related to regression standard error, and $R^2$ gives the proportion of target variance explained by the model. We also consider the ranked probability skill score (RPSS), a measure of the probabilistic skill of the model, framed in terms of its ability, relative to a naïve climatology forecast, to predict the probability of dry, normal, or wet years as defined by terciles of the observed flow volumes (e.g., Wiegel et al., 2007; Guihan, 2014; Fleming and Goodbody, 2019). RMSE, $R^2$, and RPSS were assessed using cross-validated predictions calculated by the general method of Garen (1992), which is widely used for WSF applications of PCR in the western US, except for RPSS in the case of the two quantile regression methods, which was estimated using their intrinsic probability models (see also, e.g., Pagano et al., 2004; Rosenberg et al., 2011; Lehner et al., 2017; Fleming and Goodbody, 2019). These conventional accuracy metrics do not penalize the negative-valued and therefore non-physical predictions made by standard statistical models in some cases (see criterion 4 in Table 2; examples are provided in Section 4) or reward the physical acceptability of predictions made by $M^4$. A more comprehensive portrayal of model performance is therefore provided by additionally tracking binary metrics that flag whether or not the model-predicted best estimates, and the lowest of the associated prediction intervals considered in standard WSF applications (Section 3.2.2), meet the physicality requirement of being nonnegative for all available sample times.

#### 3.3.2. Live operations

Live operational testing was undertaken during the 2020 forecast season for a subset of five locations at multiple forecast issue dates (Section 2). For the 1 February and 1 March forecast dates, predictor candidate pools were similar to those used in hindcasting; for the 1 January and 1 April operational forecasts, the same models were used as in hindcasting (see Section 2). Operational testing was completed alongside, but separately from, the existing PCR-based NRCS forecast system. The primary goals of the live testing were to confirm compliance of $M^4$ with two core requirements of an operational WSF system: related workflows must be logistically feasible in a time- and resource-constrained operational environment, and any given forecast must be readily and succinctly explainable in terms of known geophysical processes and current climatic conditions.

#### 3.3.3. Evaluation criteria

The evaluation criteria approximately reflect the practical design criteria outlined in Section 3.2 and Table 2. For hindcast testing, assessment included the five quantitative metrics described in Section

3.3.1, and in particular, relative performance improvements for these metrics compared to the linear PCR benchmark of the existing NRCS system (see Section 3.3.4 below). Additionally, robustness, ease of use, and automation potential were qualitatively assessed during both hindcast and operational testing. This included ability to train and operate $M^4$ model suites across a diverse set of test cases quickly with no manual tuning. We also assessed amenability of the resulting forecasts to interpretation, focusing on live operational tests and ability to infer a relatable 'storyline' around the current forecast. This is a requirement for achieving user and client buy-in for an operational WSF system but is conventionally regarded as a challenge for ML (Section 1).

#### 3.3.4. Benchmark

Current NRCS operational PCR models as developed and implemented in the VIPER operational forecasting software platform (Garen, 1992; Perkins et al., 2009; see also Sections 1.2 and 1.3) for the test cases (Section 2) provide a challenging and broadly relevant performance benchmark for $M^4$. PCR models are the basis of NRCS and other operational WSF systems, represent a general reference point as a standard linear Gaussian statistical regression approach, and are widely used in hydroclimatic research (Section 1.2). Using PCR-based WSFs as a point of comparison for $M^4$ is further reinforced by studies suggesting, relative to extended streamflow prediction (ESP)-based process-simulation models currently in western North American operational use, similar accuracy as well as better prediction uncertainty intervals due to underdispersion common in the ensemble spreads of most operational ESP systems (e.g., Gobena and Gan, 2010; Harpold et al., 2020). Prediction intervals for the benchmark model are generated using a heuristic typical of current operational PCR-based WSF (including NRCS) systems and widely employed in other regression applications, that is, error is assumed to follow a stationary normal distribution centered at the best-estimate prediction with a standard deviation equal to the regression standard error (e.g., Garen, 1992; Hyndman and Athanasopoulos, 2013). Benchmark model performance was tracked using the same criteria and procedures described in Sections 3.3.1 and 3.3.3.

## 4. Results and discussion

### 4.1. Retrospective testing

#### 4.1.1. An introductory example: The Owyhee River

Preliminary comparison of outcomes for the Owyhee River April 1 publication date (Fig. 1; see Section 2.2) against conventional PCR techniques provides a sense of the practical benefits of $M^4$ and a useful starting point for more detailed examination of test-case results. Note that this river is known from operational NRCS experience to be one of the more challenging forecast points in the western US for reasons that will become apparent below and is therefore an instructive litmus test.

Fig. 3(a) gives hindcast predictions from conventional PCR as used by NRCS and others. For several samples, the best-estimate runoff predictions are negative-valued and therefore physically impossible. Additionally, it does not generate required time-varying and asymmetric prediction bounds, which are obviously too wide in low-flow years and too narrow in high-flow years, further contributing to negative-valued prediction intervals. These issues are routinely addressed successfully in PCR-based WSF using predictand transforms, a common approach to applying classical statistical models to nonlinear, non-stationary, and non-Gaussian prediction problems. However, such procedures require manual user intervention to determine whether a transform is needed and if so which one. This is a slow non-OTL procedure (Sections 1 and 3);

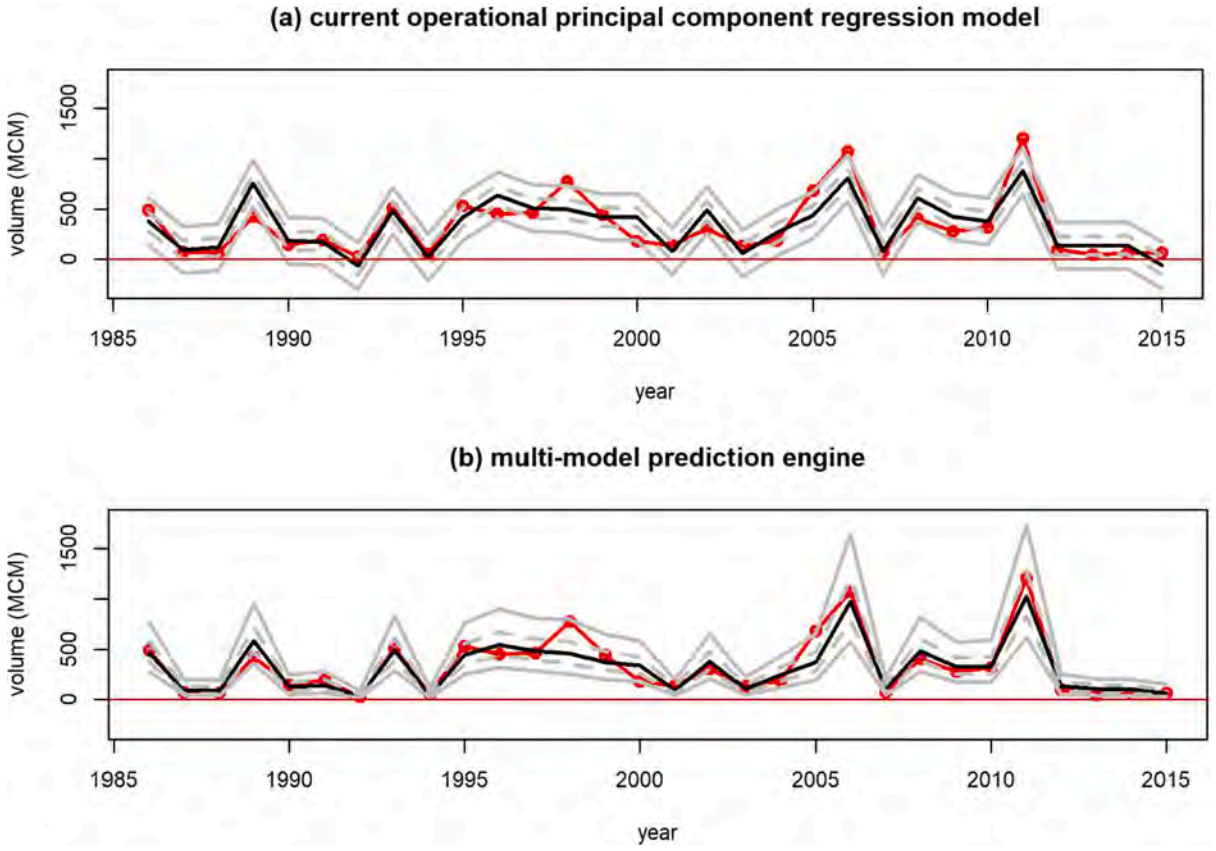


**Fig. 3.** April 1 Owyhee spring-summer volume forecasts from (a) existing PCR-based WSF system implemented in the VIPER software platform, (b) $M^4$. Units are millions of cubic meters (MCM). Red dots connected by thick red line: observations; thick black line: best-estimate forecast; solid gray lines: 0.10 and 0.90 quantile prediction intervals, taken to be minimum and maximum reasonable estimates in operational NRCS practice; dashed gray lines: 0.30 and 0.70 quantile prediction intervals. Red horizontal line gives zero volume for reference.

depends on expert opinion, undermining objectivity, reproducibility, and defensibility; and does not separate distributional modifications from functional form modifications (see Fleming and Goodbody, 2019). Moreover, in operational practice at a few locations, multiple models are maintained in the current NRCS system, such as a non-transformed primary model giving the best overall performance, and a transformed model used on an as-needed basis to avoid negative-valued predictions during dry years or other complications; if done well (cf. Wood et al., 2020 vs. Weber et al., 2012) these subjective, ad hoc, time-consuming model development and selection choices can improve the accuracy and flexibility of traditional linear models but are even further removed from an ideal OTL workflow. In contrast, Fig. 3(b) shows $M^4$ generates strictly non-negative best-estimate forecasts and associated prediction intervals. The prediction bounds can, when needed, vary in width from year to year and be asymmetric about the best estimate. No user input or intervention was required, apart from specifying the input variable candidate pool. $M^4$ additionally gives better forecast skill than linear PCR, as discussed in subsequent sections.

#### 4.1.2. Performance within multi-model ensemble

The benefit of extending multi-model ensembles to a diverse set of supervised machine learning methods for WSF (Section 3.2.5; see also Fleming and Goodbody, 2019) is confirmed by comparative results on test sites across the western US (Table 3). Equifinality is pervasive, as commonly seen for hydrologic and other models (Beven and Binley, 1992; Wolpert, 1996; Burnham and Anderson, 2002; Hagedorn et al., 2005). Relative to other techniques within the metasystem, a certain individual modeling method may, for a given test case, perform best on one metric but worst on another; or for a given forecast date, perform best at one location but not at others; or for a given location, perform well for one forecast date but not for others (Table 3).

Multi-model ensembles provide an established means for addressing such model selection uncertainty and give tangible performance improvements (Table 3). The ensemble mean forecast distribution frequently (one or more performance metrics in each of nine test cases) beats the performance of any of its constituent ensemble members through mutual error cancellation, a known advantage of ensemble modeling. Perhaps more importantly, metasystem mean performance almost without exception beats, matches, or is second-best to all of its constituent systems. This is far more consistent performance across performance metrics, locations, and forecast dates than any of the six individual modeling techniques within the metasystem. Such consistency is a fundamental but occasionally overlooked advantage of multi-model ensemble means (Hagedorn et al., 2005) and is important for reliable, efficient, and effective application across a large region with diverse geophysical and statistical characteristics (see ensemble modeling discussion in Section 3.2.5).

#### 4.1.3. Performance relative to existing system

Fig. 4 and Table 3 reveal the $M^4$ ensemble mean forecast distribution also meets or, in most cases, beats the current NRCS WSF model for every quantitative performance measure. Recall from Section 3.3.4 that this current PCR model is a meaningful general benchmark for operational WSF skill in the western US. On average, $R^2$ and RPSS improve by over 50% and RMSE is reduced by 13%. These accuracy improvements on 20 diverse test cases appear to mainly reflect a combination of flexibility provided by nonlinear machine learning, robustness provided by methodologically diverse multi-model ensembles, and embedding of geophysical process constraints (Section 3.2). For rivers like the Deschutes where preliminary diagnostics (Section 3.2.1) suggested no significant departures from linear stationary Gaussian processes, improvements were comparatively modest; in such cases, the ensemble AI essentially retrieves a linear stationary Gaussian model, as expected if regularization is properly implemented (Hsieh et al., 2003; Fleming, 2007). By the same token, benefits relative to the existing system were strongest for test cases like Owyhee, Gila, Clark Canyon, and Truckee, which are nonlinear, non-Gaussian, or heteroscedastic and/or where the existing system produced occasional non-physical negative predictions. Note the final ensemble mean forecast distribution was always non-negative, a key advantage over the current statistical model, which in 6 hindcast test cases provided a forecast distribution containing non-physical values in at least one year. These advantages are also qualitatively obvious for individual cases (Section 4.1.1).

### 4.2. Live operational testing

#### 4.2.1. General performance

Live testing at a subset of 5 forecast locations during the 2020 forecast season (Section 3.3) alongside the current operational system was undertaken primarily to evaluate certain practical aspects of actually using $M^4$ in a genuinely operational setting. The testing confirmed logistical feasibility of associated near-real time workflows, and $M^4$ was found reliable and simple to use, with no need for manual intervention. This OTL approach stands in contrast to existing operational WSF models, which in practice often require manual subjective choices during operations around rebuilding using different predictor sets or transforms (statistical models) or adjustments of parameters, internal states, or input data values (process models) (see Sections 1 and 3). Further, during part of the 2020 forecast season, manual snow survey sites could not be monitored due to the COVID-19 pandemic and impacts of the associated quarantine on field surveys. It was found that $M^4$ modeling suites previously developed for certain forecast points and dates that made particularly heavy use of manual snow survey data were easily and quickly retrained during routine forecast operations to use only telemetered SNOTEL data. This example illustrates that, in practice, the metasystem can provide needed flexibility and convenience when an unexpected operational condition arises.

Out-of-sample forecast accuracy comparisons to other methods were performed in the hindcasting of Section 4.1.3 and obviously were not a goal for live operational testing – little can be gleaned in this respect from a single forecast season, which effectively amounts to a sample of one in WSF. It is nevertheless encouraging to preliminarily note that, considering all 4 forecast dates at all 5 locations, mean RMSE improvement over the benchmark model in live operations was 10%, roughly comparable to improvements seen in hindcasting. Additionally, in operations, the range in the best-estimate 2020 spring-summer volume forecast across all 4 forecast issue dates for a given river decreased relative to the current NRCS model (by 21%, averaged across the 5 sites). If this apparent increase in the stability of the best-estimate prediction value from one forecast date to the next, for a given river during a given forecast season, is rigorously confirmed in further operational testing, it may reflect the inherent performance consistency advantages provided by multi-model ensemble averaging (see Section 4.1.2). Provided it is not at the expense of decreased accuracy (and the opposite is seen here, as noted above), such steadiness in the forecast is in general viewed as a desirable operational characteristic by SDOs because it simplifies hydroclimatic interpretations and client communications.

#### 4.2.2. Geophysical interpretation of two live AI-based forecasts

Ability to readily determine how model behaviors relate to physical hydrologic processes is necessary for meeting professional responsibilities around assessing the reliability of forecasts used for high-impact water management decisions, and for verification diagnostics. Physical interpretability is also vital for communicating forecasts to clients, who often include the general public, such as why a river volume prediction increased or decreased and by how much since the last forecast date. Recall from Section 1 that for some western US rivers experiencing complex or contentious water management issues, WSFs are legal requirements specified by legislation, court decisions, or international treaties. Such forecasts are routinely subject to intense public, and even political, scrutiny. Amenability to physical explanation is therefore a prerequisite for operational WSF systems in the region,

**Table 3**
Metasystem performance on 20 hindcast test cases at 11 sites (see Fig. 1, Table 1, and text). Outcomes shown for constituent machine learning and statistical models, and final multi-model ensemble mean forecast distribution derived from them. As a benchmark, results are provided for the established official NRCS forecast model (VIPER) using a linear stationary Gaussian PCR approach known to generally perform at least as well as other operational WSF methods widely used in the US West, including ESPs (see text). Operational VIPER models in some cases included predictand transforms. Only April forecasts are considered here for Boulder Creek and the Little Susitna River (see Table 1 for details). Performance metrics are coefficient of determination ($R^2$), root mean square error (RMSE), ranked probability skill score (RPSS), and flags for whether negative-valued best estimates (BE) or 0.10 quantile prediction bounds (PB) occurred for any sample (see Section 3.3.1 for details). RMSE is in millions of cubic meters (MCM). Asterisk for a given model denotes the automated negativity-check algorithm in $M^4$ (see Fig. 2, Section 3.2, and Fleming and Goodbody, 2019) removed it from the ensemble. For the final ensemble mean forecast, comparative performance is summarized by superscripts on a metric-by-metric basis: [a] outperforms all retained constituent models for that metric, [b] matches the performance of the best-performing of its retained constituent models, [c] outperforms VIPER, [d] matches VIPER.

| | | $M^4$ prediction analytics engine | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | *VIPER* | LR | QR | mANN | RF | MCQRNN | SVM | Ensemble |
| Truckee Apr 1 | | | | | | | | |
| $R^2$ | *0.89* | 0.93 | 0.93 | 0.94 | 0.91 | 0.94 | 0.93 | **0.94**[b,c] |
| RMSE | *61.4* | 47.6 | 46.9 | 43.5 | 54.5 | 44.2 | 50.2 | **43.3** [a,c] |
| RPSS | *0.75* | 0.81 | 0.80 | 0.80 | 0.82 | 0.80 | 0.71 | **0.81**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *Y* | N | Y | N | N | N | N | **N** [b,c] |
| Truckee Jan 1 | | | | | | | | |
| $R^2$ | *0.21* | 0.25 | 0.20 | 0.14 | 0.21 | 0.15 | 0.41 | **0.27**[c] |
| RMSE | *162* | 158 | 173 | 176 | 162 | 170 | 142 | **156**[c] |
| RPSS | *0.05* | 0.05 | 0.23 | 0.02 | 0.10 | 0.13 | 0.27 | **0.17**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *Y* | N | Y | N | N | N | N | **N** [b,c] |
| Yellowstone Apr 1 | | | | | | | | |
| $R^2$ | *0.79* | 0.81 | 0.82 | 0.82 | 0.72 | 0.82 | 0.83 | **0.82**[c] |
| RMSE | *252* | 240 | 244 | 236 | 305 | 236 | 239 | **235** [a,c] |
| RPSS | *0.61* | 0.62 | 0.65 | 0.60 | 0.55 | 0.61 | 0.62 | **0.63**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Yellowstone Jan 1 | | | | | | | | |
| $R^2$ | *0.58* | 0.58 | 0.60 | 0.57 | 0.61 | 0.58 | 0.57 | **0.62** [a,c] |
| RMSE | *357* | 356 | 351 | 362 | 351 | 358 | 361 | **342** [a,c] |
| RPSS | *0.20* | 0.20 | 0.25 | 0.19 | 0.30 | 0.25 | 0.21 | **0.26**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Owyhee Apr 1 | | | | | | | | |
| $R^2$ | *0.67* | 0.67* | 0.65* | 0.70 | 0.85 | 0.64 | 0.81 | **0.83**[c] |
| RMSE | *170* | 169* | 180* | 160 | 125 | 181 | 133 | **130**[c] |
| RPSS | *0.49* | 0.53* | 0.54* | 0.45 | 0.43 | 0.54 | 0.59 | **0.56**[c] |
| BE < 0? | *Y* | Y* | Y* | N | N | N | N | **N** [b,c] |
| PB < 0? | *Y* | Y* | Y* | N | N | N | N | **N** [b,c] |
| Owyhee Jan 1 | | | | | | | | |
| $R^2$ | *0.14* | 0.26 | 0.25* | 0.50 | 0.28 | 0.17 | 0.49 | **0.43**[c] |
| RMSE | *276* | 255 | 263* | 207 | 251 | 280 | 211 | **225**[c] |
| RPSS | *0.10* | 0.13 | 0.11* | 0.03 | 0.05 | 0.21 | 0.09 | **0.14**[c] |
| BE < 0? | *N* | N | N* | N | N | N | N | **N** [b,d] |
| PB < 0? | *Y* | Y | Y* | N | N | N | N | **N** [b,c] |
| Clark Canyon Apr 1 | | | | | | | | |
| $R^2$ | *0.47* | 0.54* | 0.61* | 0.57 | 0.56 | 0.59 | 0.44 | **0.60** [a,c] |
| RMSE | *59.4* | 55.8* | 52.1* | 53.8 | 55.7 | 52.6 | 62.0 | **52.7**[c] |
| RPSS | *0.28* | 0.32* | 0.39* | 0.24 | 0.23 | 0.37 | 0.17 | **0.30**[c] |
| BE < 0? | *Y* | Y* | Y* | N | N | N | N | **N** [b,c] |
| PB < 0? | *Y* | Y* | Y* | N | N | N | N | **N** [b,c] |
| Clark Canyon Jan 1 | | | | | | | | |
| $R^2$ | *0.18* | 0.21 | 0.27* | 0.25 | 0.34 | 0.18 | 0.41 | **0.41**[b,c] |
| RMSE | *74.8* | 72.9 | 70.9* | 71.9 | 66.8 | 75.5 | 64.0 | **64.0**[b,c] |
| RPSS | *0.06* | 0.06 | 0.25* | −0.01 | 0.06 | 0.17 | 0.01 | **0.13**[c] |
| BE < 0? | *N* | N | N* | N | N | N | N | **N** [b,d] |
| PB < 0? | *Y* | Y | Y* | N | N | N | N | **N** [b,c] |
| Gila Apr 1 | | | | | | | | |
| $R^2$ | *0.62* | 0.69 | 0.71 | 0.80 | 0.73 | 0.71 | 0.74 | **0.76**[c] |
| RMSE | *12.2* | 11.1 | 11.3 | 9.0 | 10.5 | 10.9 | 10.4 | **9.9**[c] |
| RPSS | *0.53* | 0.59 | 0.64 | 0.62 | 0.64 | 0.63 | 0.64 | **0.66** [a,c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | Y | N | N | N | N | **N** [b,d] |
| Gila Jan 1 | | | | | | | | |
| $R^2$ | *0.14* | 0.25 | 0.33 | 0.36 | 0.46 | 0.24 | 0.25 | **0.37**[c] |
| RMSE | *70.6* | 64.9 | 62.8 | 60.3 | 56.0 | 65.6 | 66.0 | **60.0**[c] |
| RPSS | *0.17* | 0.23 | 0.35 | 0.31 | 0.16 | 0.29 | 0.25 | **0.34**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Boulder Apr 1 | | | | | | | | |
| $R^2$ | *0.28* | 0.29 | 0.33 | 0.22 | 0.57 | 0.33 | 0.59 | **0.47**[c] |

(*continued on next page*)

**Table 3** (*continued*)

| | | $M^4$ prediction analytics engine | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Metric | *VIPER* | LR | QR | mANN | RF | MCQRNN | SVM | Ensemble |
| RMSE | *14.7* | 14.6 | 14.5 | 15.5 | 11.4 | 14.1 | 11.1 | **12.7**[c] |
| RPSS | *0.06* | 0.01 | 0.12 | 0.01 | 0.35 | 0.10 | 0.20 | **0.19**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | Y | N | N | **N** [b,d] |
| Detroit Lake Apr 1 | | | | | | | | |
| $R^2$ | *0.55* | 0.62 | 0.63 | 0.59 | 0.52 | 0.62 | 0.52 | **0.62**[c] |
| RMSE | *118* | 108 | 106 | 112 | 122 | 109 | 123 | **108**[c] |
| RPSS | *0.32* | 0.29 | 0.34 | 0.29 | 0.37 | 0.37 | 0.29 | **0.36**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Detroit Lake Jan 1 | | | | | | | | |
| $R^2$ | *0.11* | 0.24 | 0.14 | 0.33 | 0.26 | 0.18 | 0.29 | **0.31**[c] |
| RMSE | *168* | 153 | 165 | 145 | 152 | 160 | 148 | **147**[c] |
| RPSS | *0.12* | 0.11 | 0.11 | 0.23 | 0.17 | 0.21 | 0.19 | **0.21**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Fontenelle Apr 1 | | | | | | | | |
| $R^2$ | *0.62* | 0.73 | 0.72 | 0.70 | 0.70 | 0.72 | 0.75 | **0.75**[b,c] |
| RMSE | *233* | 227 | 235 | 236 | 239 | 231 | 219 | **218** [a,c] |
| RPSS | *0.41* | 0.53 | 0.58 | 0.60 | 0.53 | 0.52 | 0.58 | **0.59**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *Y* | N | Y | N | N | N | N | **N** [b,c] |
| Fontenelle Jan 1 | | | | | | | | |
| $R^2$ | *0.29* | 0.28 | 0.34 | 0.37 | 0.43 | 0.27 | 0.47 | **0.41**[c] |
| RMSE | *370* | 370 | 354 | 347 | 334 | 373 | 320 | **335**[c] |
| RPSS | *0.11* | 0.08 | 0.14 | 0.10 | 0.19 | 0.12 | 0.11 | **0.16**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Deschutes Apr 1 | | | | | | | | |
| $R^2$ | *0.78* | 0.81 | 0.82 | 0.75 | 0.74 | 0.81 | 0.81 | **0.83** [a,c] |
| RMSE | *6.7* | 6.3 | 5.9 | 7.0 | 7.4 | 6.2 | 6.3 | **5.9**[b,c] |
| RPSS | *0.61* | 0.59 | 0.66 | 0.56 | 0.56 | 0.60 | 0.52 | **0.61** [d] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Deschutes Jan 1 | | | | | | | | |
| $R^2$ | *0.55* | 0.55 | 0.57 | 0.50 | 0.43 | 0.57 | 0.56 | **0.57**[b,c] |
| RMSE | *9.5* | 9.5 | 9.5 | 10.2 | 10.8 | 9.3 | 10.0 | **9.3**[b,c] |
| RPSS | *0.15* | 0.07 | 0.17 | 0.05 | 0.17 | 0.19 | 0.11 | **0.17**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Little Susitna Apr 1 | | | | | | | | |
| $R^2$ | *0.26* | 0.54 | 0.57 | 0.44 | 0.42 | 0.51 | 0.60 | **0.59**[c] |
| RMSE | *24.8* | 19.6 | 20.2 | 21.5 | 22.3 | 20.1 | 18.7 | **18.7**[b,c] |
| RPSS | *0.18* | 0.37 | 0.42 | 0.39 | 0.41 | 0.40 | 0.31 | **0.43** [a,c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Rio Grande Apr 1 | | | | | | | | |
| $R^2$ | *0.57* | 0.62 | 0.64 | 0.54 | 0.58 | 0.59 | 0.59 | **0.64**[b,c] |
| RMSE | *151* | 142 | 140 | 155 | 151 | 147 | 148 | **138** [a,c] |
| RPSS | *0.43* | 0.41 | 0.48 | 0.37 | 0.36 | 0.46 | 0.38 | **0.45**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| Rio Grande Jan 1 | | | | | | | | |
| $R^2$ | *0.47* | 0.56 | 0.61 | 0.59 | 0.60 | 0.58 | 0.65 | **0.64**[c] |
| RMSE | *168* | 153 | 144 | 147 | 146 | 149 | 137 | **138**[c] |
| RPSS | *0.32* | 0.37 | 0.44 | 0.36 | 0.35 | 0.39 | 0.39 | **0.41**[c] |
| BE < 0? | *N* | N | N | N | N | N | N | **N** [b,d] |
| PB < 0? | *N* | Y | N | N | N | N | N | **N** [b,d] |

including those based on statistical and machine learning methods (e.g., Garen, 1992; Weber et al., 2012; Fleming and Goodbody, 2019; Fleming et al., 2021). As discussed in Sections 1 and 3, such relatable hydroclimatic 'storylines' are widely perceived to run contrary to the nominally black-box nature of machine learning, in turn slowing the migration of AI into operational hydrology. Some successful early attempts notwithstanding (e.g., Cannon and McKendry, 2002; Fleming, 2007), explainable or 'glass-box' AI largely remains at the cutting edge of geoscience (Kratzert et al., 2018; Reichstein et al., 2019; McGovern et al., 2019; Nearing et al., 2021). Our live operational forecasting provided an opportunity to test the pragmatic, WSF-specific approach to explainable machine learning implemented in $M^4$ (Section 3.2.3).

Two examples are summarized below. These were the most complicated interpretive scenarios encountered during live operational testing and, as such, illustrate ability of the $M^4$ metasystem to extract geophysical process reasoning from an AI-based predictive approach.

We first consider the February 1, 2020 operational forecast of 101% normal for February-May 2020 Gila River flow volume. During $M^4$ training, candidate predictors for this operational test case were forecast-date SWE and water year-to-date precipitation at a few sites in this remote mountain tributary to the Lower Colorado River (Table 1, Fig. 1), similar to the current WSF system (Section 2). In this particular test case, for all models in $M^4$, the genetic algorithm retained only the leading PCA mode, which (given the candidate predictors used) is a de

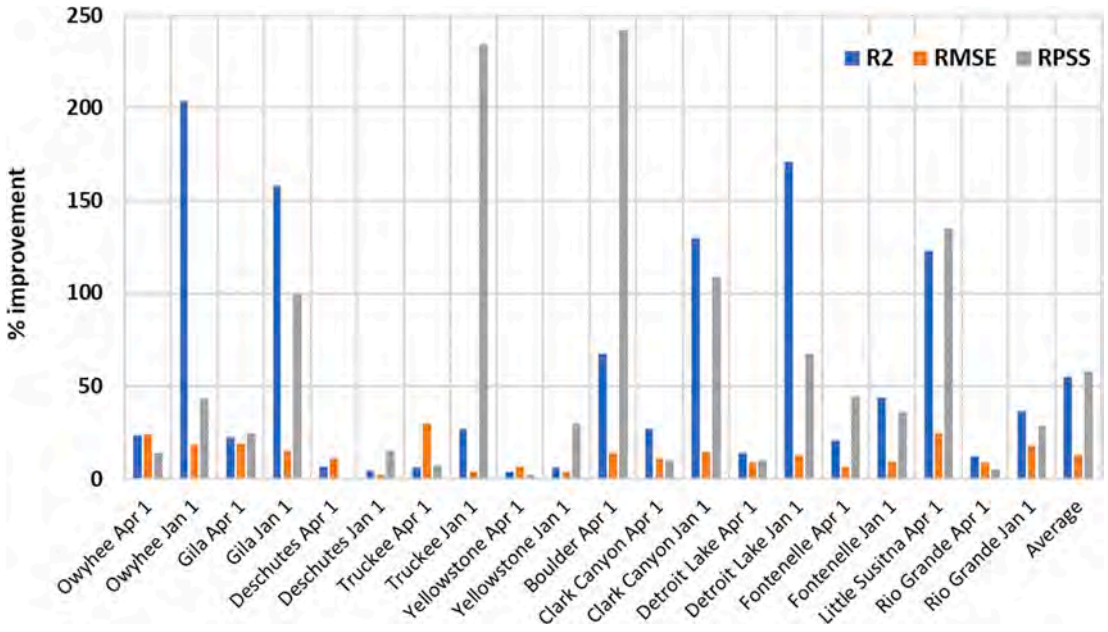


**Fig. 4.** Percentage improvements provided by $M^4$ ensemble mean forecast on three common performance metrics for 20 test cases, relative to existing NRCS operational WSF system. Average improvement across all hindcast test cases is also shown. Other key performance aspects were additionally considered in the overall evaluation (see text).

facto watershed-scale index of wintertime climate conditions, and in particular, basin water inputs. The genetic algorithm effectively casts votes for which geophysical drivers, among the candidates in the pool, it thinks most important. Thus, the resulting 'popular vote' across models, along with the PCA eigenvector, illuminate which inputs most influence the forecast. We see (Table 4) that some predictors (e.g., Silver Creek Divide SWE) were more popular than others (e.g., Signal Peak SWE), whereas eigenvector weights across retained predictors were roughly uniform for a given model in this case.

Further, dimensionality reduction (Section 3) allows direct visualization of the ensemble mean empirical input–output map detected by $M^4$ (Fig. 5). In this example, such a graphical representation illustrates the relationship between the sole retained feature in each model (the leading PCA mode, which as noted previously is an index of wintertime precipitation inputs) and the sole target (February-May volume), averaged across constituent models. The relationship exhibits a shallower slope in dry years (Fig. 5). This decrease in the first derivative of the input–output map during drought conditions is a functional form that is known, from both $M^4$ testing and NRCS operational experience with current WSF models, to be especially widespread in semi-arid rivers like the Gila, and capturing it in a WSF model is needed to avoid negative-valued predictions sometimes generated by linear models over the relevant state space (blue dashed line in Fig. 5; see also Section 4.1).

This nonlinearity reflects several geophysical causes. Wet-year flow is closely coupled with, and therefore sensitive to, variations in winter precipitation and snowpack, giving a steeper curve; in contrast, during dry years, a higher proportion of springtime snowmelt goes to refilling soil moisture and aquifer storage before producing a flow response in this desert river. That is, the phenomenon can be viewed as an approximate seasonal-scale analogy to the well-known nonlinearity of daily or hourly rainfall-runoff relationships: infiltration limits reached during storms reduce the mitigating impact of soil moisture storage and more directly couple surface runoff to rainfall fluctuations, increasing surface runoff generation per unit precipitation, whereas stable baseflow contributions from soil, aquifer, channel, wetland, and other natural storage mechanisms partly flatten the rainfall-runoff relationship during dry spells. Additionally, wet-year runoff efficiency improves due to proportionately lower evapotranspiration losses, reflecting cooler temperatures and greater cloud cover associated with wet conditions here, giving a greater runoff increase per unit precipitation increase (Lukas and Harding, 2020). These factors may also dovetail with climate elasticity. If flow volume dependence on precipitation inputs follows a power-law for a given river, its precipitation elasticity of runoff is fixed and equal to the power-law exponent (Sankasubramanian et al., 2001). Though additional work would be required to formally tie the curve in Fig. 5 to climate elasticity, preliminary power-law fits to this input–output relationship non-parametrically estimated by $M^4$, following simple linear rescaling to ensure positive-valued PC1 scores,

**Table 4**
Leading-mode eigenvector for each model, adjusted to common polarity across models, and model-voting results for each predictor in the candidate pool. Results are for February 1 Gila River $M^4$ forecast models. P is Oct 1-to-Jan 31 accumulated precipitation, SWE is Feb 1 start-of-day SWE. - indicates predictor was not selected for retention by genetic algorithm for the corresponding model. Popular vote for each variable across models, and PCA loadings for each retained variable for a given model, give some indication of the relative influences of various inputs on the forecast. Current values during the February 1, 2020 operational forecast are provided for each candidate predictor as a percentage of its mean over the 30-year normal period used in model development.

| Candidate predictor | Leading-mode PCA eigenvector entries | | | | | | % models voting for variable | % normal |
|---|---|---|---|---|---|---|---|---|
| | LR | QR | mANN | RF | MCQRNN | SVM | | |
| Lookout Mountain P | – | 0.55 | – | 0.54 | – | 0.51 | 50 | 107 |
| Lookout Mountain SWE | – | 0.59 | 0.58 | – | – | 0.49 | 50 | 15 |
| Signal Peak P | 0.71 | – | 0.55 | – | 0.71 | 0.51 | 67 | 117 |
| Signal Peak SWE | – | – | – | 0.60 | – | – | 17 | 0 |
| Silver Creek Divide P | – | – | – | – | – | – | 0 | 113 |
| Silver Creek Divide SWE | 0.71 | 0.59 | 0.60 | 0.60 | 0.71 | 0.50 | 100 | 156 |

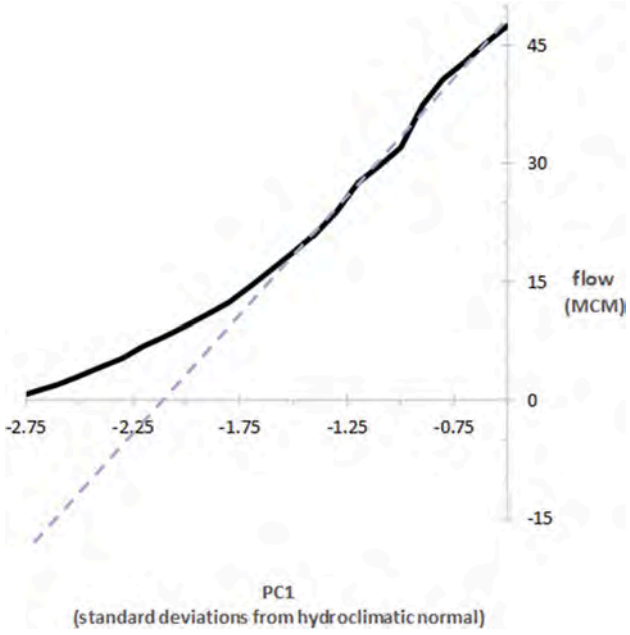


**Fig. 5.** Nonlinear ensemble mean relationship (thick black line) between AI-based Gila River volume prediction and leading-mode PCA scores time series (PC1), which indexes watershed-scale winter climatic inputs, illustrating partial flattening during dry years. For reference, blue dashed line continues linear relationship to low-flow conditions.

implies an elasticity of roughly 2 ($R^2 > 0.95$), consistent with Colorado Basin elasticity estimates from standard methods (e.g., Vano et al., 2012). That this nonlinearity is strongest in semi-arid rivers also seems consistent with Vano et al. (2012), who found higher elasticities in drier basins. All things considered, it seems clear that nonlinear relationships between climate forcing and watershed response, empirically detected by $M^4$, are explainable in terms of known hydrological processes, connect data-driven WSF to broader concepts in watershed hydrology and climate science, and suggest specific future research directions in physical hydrology, collectively belying the conventional black-box view of machine learning.

Moreover, with the foregoing interpretive tools and context in mind, the forecast of near-normal runoff issued operationally by $M^4$ on February 1, 2020 is easily diagnosed to form a relatable and compact storyline for clients. From Table 4, precipitation overall was somewhat above-normal throughout the watershed. That in turn increased soil moisture basin-wide, as well as snowpack at Silver Creek Divide, which among the SNOTEL sites considered in this test case is the most northerly, highest-elevation, on average highest-SWE, and generally most informative (given its selection by all six models) for spring runoff prediction. These factors pushed up the forecast. However, snowpack was very low in the south (Signal Peak) and east (Lookout Mountain), presumably reflecting temperature-controlled variations in winter precipitation phase across this southwestern New Mexico watershed, pulling spring-summer runoff projections back down to near-normal. Additionally, in part because of nonlinear relationships between wintertime weather and runoff volume (Fig. 5), which occur because snowpack and precipitation are not the sole environmental processes affecting streamflow, percentages of normal match only approximately between observed inputs and predicted flow.

Our second example is the January 1, 2020 operational forecast of 71% normal for April-July 2020 Deschutes River flow volume. During $M^4$ training, candidate predictors for this operational test case were forecast-date SWE and water year-to-date precipitation at a few sites in the remote mountain headwaters of this mid-Columbia River tributary (Table 1, Fig. 1), and antecedent streamflow. This is similar to the current WSF system (Section 2). Use of antecedent streamflow reflects the large known impact of volcanic aquifers in generating and stabilizing Deschutes River flows; surface water-groundwater interactions are unusually pronounced here, leading to muted seasonality in flow and strong memory in streamflow time series (Table 1; e.g., O'Connor et al., 2003; Risley et al., 2005).

Resulting 'popular votes' (see above) and eigenvectors are given in Table 5. The popular vote cast by the genetic algorithm across the six constituent $M^4$ models favors Irish Taylor precipitation, and in particular antecedent Deschutes streamflow, the only candidate variable retained by all models. In this test case, the genetic algorithm selected only the leading mode for half the models and both the leading and second modes for the remainder. For models retaining only the leading mode (QR, RF, and SVM), two of the five candidate predictors were

**Table 5**
As in Table 4, but for the January 1 forecast of Deschutes River April-July flow volume, and reformatted to give eigenvectors corresponding to both the leading PCA mode, and to the second PCA mode for constituent $M^4$ models that retain it. Eigenvector entries for models that do not retain the second PCA mode are marked not applicable (n/a), and as in Table 4, – indicates predictor was not selected for retention by the genetic algorithm for the corresponding model. P is Oct 1-to-Dec 31 accumulated precipitation, SWE is Jan 1 start-of-day SWE, and Q is antecedent (December) total flow volume.

| | Candidate predictor | | | | |
|---|---|---|---|---|---|
| | Irish Taylor SWE | Irish Taylor P | Three Creeks Meadow SWE | Three Creeks Meadow P | Deschutes below Benham Falls Q |
| *PCA eigenvector entries, leading mode* | | | | | |
| LR | 0.59 | 0.69 | – | – | 0.42 |
| QR | – | – | – | 0.71 | 0.71 |
| mANN | 0.44 | 0.50 | 0.49 | 0.48 | 0.29 |
| RF | – | 0.71 | – | – | 0.71 |
| MCQRNN | 0.59 | 0.69 | – | – | 0.42 |
| SVM | – | 0.71 | – | – | 0.71 |
| *PCA eigenvector entries, second mode* | | | | | |
| LR | −0.56 | −0.03 | – | – | 0.83 |
| QR | n/a | n/a | n/a | n/a | n/a |
| mANN | −0.51 | 0.06 | −0.33 | 0.29 | 0.74 |
| RF | n/a | n/a | n/a | n/a | n/a |
| MCQRNN | −0.56 | −0.03 | – | – | 0.83 |
| SVM | n/a | n/a | n/a | n/a | n/a |
| *% of models voting for candidate variable* | | | | | |
| | 50 | 83 | 17 | 33 | 100 |
| *Current % normal for candidate variable* | | | | | |
| | 39 | 49 | 71 | 57 | 74 |

selected: accumulated precipitation at either Three Creeks Meadow or Irish Taylor, and antecedent flow. The corresponding eigenvector weights were equal across the two predictors for a given model. For models retaining both the leading and second PCA modes (LR, mANN, MCQRNN), the leading-mode eigenvector predominantly weighted snow or precipitation, whereas the second-mode loading pattern predominantly weighted antecedent flow. Recall that PCA mode order is determined by relative ability to explain variance in the input matrix, not relative ability to explain variance in the target when those modes are used as predictive features in a supervised learner. Overall, then, various GA-optimized models address the combination of two distinct forcing mechanisms – wintertime hydroclimate, and groundwater conditions – in one of two ways. QR, RF, and SVM retain a single PCA mode which aggregates the effects of both forcing mechanisms. For LR, mANN, and MCQRNN, the leading PCA mode gives a watershed-scale index of seasonal climate to date, as in other test cases (e.g., foregoing Gila River example), whereas the second mode indexes groundwater contributions.

For all six constituent $M^4$ models, functional forms in this test case were approximately linear, as expected based on hindcast testing and preliminary diagnostics (see Section 4.1.3). Graphical representations cannot be compactly provided for the ensemble mean response as in Fig. 5, because different models retain different numbers of features for the Deschutes, but we can easily examine the input–output maps for each of the six models individually. A representative example is provided in Fig. 6, illustrating a nearly planar response surface extracted by one of the artificial neural networks. Mechanisms for the apparent absence of substantial nonlinearity here requires further study, but comparisons of functional forms across several test cases from different hydroclimatic settings (not shown here for conciseness) strongly suggest that it reflects water abundance in some way. A possible explanation is that the aforementioned plentiful aquifer contributions to streamflow, plus heavy snowmelt and precipitation inputs to this very wet Pacific Northwest basin near the crest of the Cascades Range, are such that basin water balance does not become sufficiently depleted, even in drought years, for the form of the input–output mapping to change as a function of wetness as seen in the Gila and other semi-arid or arid basins, where flow volumes can approach zero in dry years and the functional form must therefore level off at low flows as in Fig. 5.

As for the Gila River above, with these interpretive tools and context in mind the January 1, 2020 operational April-July volume forecast for the Deschutes River is easily diagnosed to form a relatable and compact storyline for clients. Very early-winter (January 1) SWE and accumulated precipitation data are typically poor indicators of the total snowpack that will be eventually available for spring-summer melt in the Upper Deschutes Basin. They therefore offer limited WSF skill at this January publication date and are collectively given less influence on the ensemble forecast by $M^4$. In contrast, the only candidate predictor retained in all six $M^4$ methods is antecedent flow, consistent with the unique hydrogeologic characteristics of the Deschutes River, stabilizing its flows and generating extensive time series memory that facilitates forecasting in an autoregressive model-like fashion. (Note that this relationship between catchment storage and streamflow memory, and the resulting streamflow forecasting capability, can also be explicitly tied to a finite-difference approximation to the linear reservoir model of watershed hydrology; for details see Fleming (2007) and references therein). The 74% normal value of antecedent flow, with some additional support from the 71% normal value of SWE at Three Creeks Meadow for the mANN model, therefore brings up the ensemble mean predicted volume to 71% of normal despite dry wintertime conditions to January 1, 2020.

Model structures, optimized feature sets, and associated hydroclimatic explanations varied significantly across test cases, as would be expected given their geophysical diversity. However, it was found that straightforward physical interpretations of the models, and of their operational forecasts in light of currently observed conditions, were readily apparent in all cases. In general, these explanations were roughly similar to or simpler than those for the two operational test cases detailed above.

### 4.3. *Mainstreaming AI in hydrology: Some implications of $M^4$ operational viability*

Much effort has been invested over recent decades in improving physics-oriented process-simulation models of river hydrology. In theory, these models have certain advantages over data-driven volume-prediction methods, such as improved physical process diagnostics, better suitability for nonstationary environments, and ability to

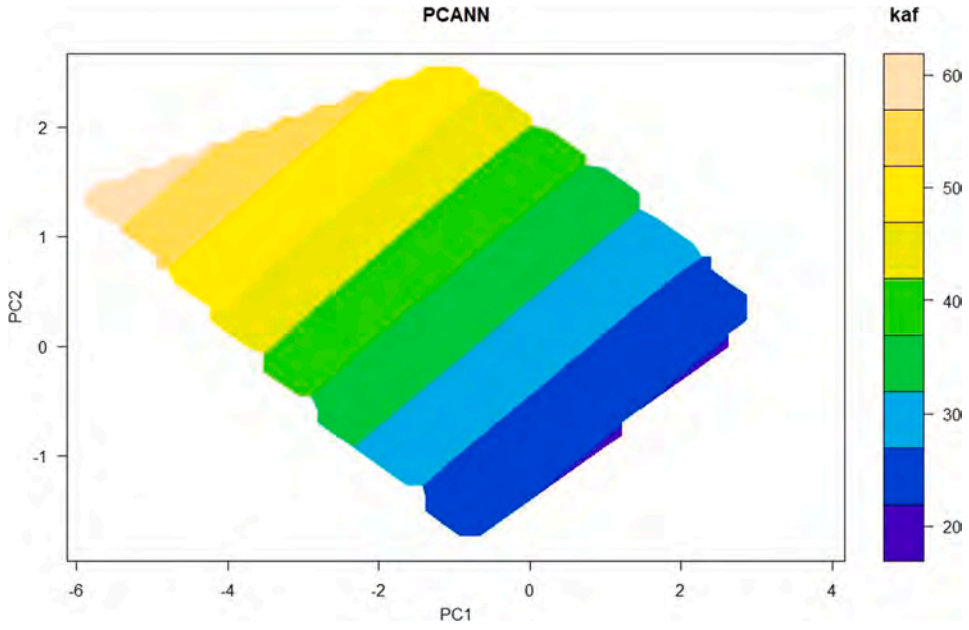


**Fig. 6.** Approximately linear (near-planar) response surface for the monotone artificial neural network (mANN) relating the January 1 prediction of April-July Deschutes River flow volume to the leading and second PCA modes. Both modes were retained by genetic algorithm feature-optimization for this particular constituent model within the $M^4$ metasystem as trained and tested for this combination of forecast date, target period, and location.

generate daily streamflow traces that are useful in some applications. Moreover, the nominally black-box nature of machine learning exacerbates, relative even to classical statistical methods, the shortcomings of data-driven prediction frameworks relative to more explicitly physics-oriented approaches. Given these considerations, is AI a viable alternative at all for operational WSF? The results of this study provide some insight into that question, which is increasingly pressing as, on the one hand, water scarcity and the associated need for better hydrometeorological forecasts increase, and on the other hand, AI progressively permeates science and society in what has been termed the fourth paradigm of science (Hey et al., 2009) and subsequently the fourth industrial revolution (Schwab, 2017).

It is useful to begin by identifying some general advantages of data-driven models for operational WSF. Statistical volume modeling is a proven method that remains the backbone of most WSF systems in western North America, often serving as either the sole prediction technology or as a complement to ESPs. Some organizations currently running statistical WSF models include NRCS, California Department of Water Resources, Colorado Basin River Forecast Center, BC Hydro, Bureau of Reclamation and Army Corps of Engineers forecasts in the Columbia Basin, British Columbia River Forecast Centre, and Alberta Environment. Reasons for continued interest in data-driven models include intrinsically lower development and operation costs, similar forecast skill, greater amenability to new predictive data types like multiple climate indices, operational simplicity and robustness, and easier and more accurate estimates of forecast uncertainty, relative to ESPs (e.g., Gobena et al., 2013; Risley et al., 2005; Fleming and Dahlke, 2014; Hsieh et al., 2003; Grantz et al., 2005; Harpold et al., 2016; Regonda et al., 2006a; Rosenberg et al., 2011; Pagano et al., 2014; Minxue et al., 2016; Mendoza et al., 2017; Robertson et al., 2013). There is also accumulating evidence that certain carefully implemented machine learning algorithms can make accurate hydroclimatic predictions under conditions not sampled during a historical training period, that is, extrapolate successfully (Schnorbus and Cannon, 2014; Shrestha et al., 2017; Kratzert et al., 2019), facilitating their use in a changing climate for instance. Conversely, some theoretical advantages of process-based models over data-driven approaches are incompletely realized in practice. For instance, diagnosing forecast failures in complex process models is not always straightforward, and process-simulation streamflow models sufficiently accurate for widespread operational use by the applied water resource community normally contain parameters requiring de facto statistical calibration to historical records, potentially undermining applicability to nonstationary environments.

More fundamental considerations also motivate temporally coarse-grained prediction techniques like data-driven seasonal volume models. Temporal aggregation often simplifies the underlying statistical physics of any system, and the optimal level of model detail and complexity required to describe and predict that system therefore depends on problem timescale. Specifically, temporal data aggregation typically increases the signal-to-noise ratio of lower-frequency geophysical processes and can (at least partially) linearize functional relationships and attenuate statistical complications like autocorrelation (if the aggregation interval exceeds the decorrelation timescale) and non-Gaussian distributions (reflecting the central limit theorem). For details around these general principles, and some specific hydrology and climate examples, see, e.g., Packard et al. (1980); Finney et al. (1998); Daw et al. (2003); Penland (1996); Hsieh (2009); Newman (2007); Newman (2012); Micovic and Quick (2009); and Fleming and Barton (2015). Hydrologic process-simulation models usually contain strongly nonlinear, serially correlated, and non-Gaussian physics relevant to the hourly to daily timescales at which they typically operate. This is obviously needed for short-term flood forecasting but represents a multiple (2 to nearly 4) order-of-magnitude mismatch to the task of predicting, on an annual basis, accumulated spring-summer flow volume a few months ahead based mainly on snowpack data. Most of the additional information generated by a process-simulation model is, therefore, effectively discarded when outcomes are integrated to form the seasonal volume predictions that are the primary basis for western US water management.

Despite these advantages, and a quarter-century of research applications of AI to hydrologic prediction that consistently demonstrate better accuracy than both process-simulation and conventional statistical methods (e.g., Nearing et al., 2021), ML has largely failed to penetrate operational hydrology in general and WSF in particular. Reasons were summarized in Section 1 and primarily relate to lack of alignment of AI-based hydrologic models with the specific practical needs of operational WSF, including but not limited to geophysical explainability. Also as briefly summarized in Sections 1 and 3, $M^4$ was therefore designed to satisfy those specific practical criteria (Fleming and Goodbody, 2019). Verifying whether $M^4$ can actually accomplish that task in practice was a major goal of this study.

The retrospective and live operational testing conducted here demonstrate that these NRCS design criteria appear to have been met by $M^4$, with broader potential implications for transitioning AI into operational systems. Relative to similarly configured conventional PCR-based NRCS WSF models, which as noted above have accuracies approximately typical of operational WSF systems in the western US including process-based ESP models (Section 3.3.4), the metasystem provides better forecast skill and more realistic prediction intervals, is more robust and automated, more consistently yields physically reasonable outcomes, both integrates and is interpretable in terms of physical hydrologic process knowledge, is applicable across strongly heterogeneous geophysical environments spanning the western US and Alaska, and functions well in a genuine operational setting. Though technically more complex than current-generation statistically based operational WSF methods, it requires fewer resources to implement and operate relative to many process-simulation models. Considered collectively, these testing results show that – with careful and application-specific design and implementation – AI has capacity to bridge the gap from research to operations in a large operational WSF setting at a major service-delivery organization. Given that water resource science, engineering, and management is ultimately a practical field, and that viability in applied operational settings is therefore a key test for the overall relevance of hydrologic modeling technologies, this demonstration establishes a positive general precedent, and perhaps an implementation template, for operational hydrologic applications of ML.

That said, certain $M^4$ characteristics may additionally suggest a path for combining process-based and AI-based models. For both theoretical and practical reasons, physics-oriented process-simulation hydrological models will continue playing a major role in operational WSF in the western US. Improvements to such physics-oriented approaches may, in turn, prove valuable to improving the accuracy and comprehensiveness of WSF-related information generated for water managers and the general public. The multi-model ensemble philosophy underlying $M^4$ may enable a means for integrating those advances with concurrent ML advances; this is discussed in Section 5.

## 5. Conclusions

To summarize, in a fundamental departure from legacy WSF systems and philosophies, we investigated an OTL approach based on a recently developed, multi-model ensemble prediction analytics engine, $M^4$, that incorporates automated and explainable AI. In this study, $M^4$ was tested in both retrospective applications and live forecast operations for a relatively large and hydroclimatically diverse sample of test cases drawn from the current NRCS forecast system. This testing suggests $M^4$ meets the theoretical and practical needs that NRCS defined for its operational WSF environment, including geophysical interpretability, objectivity, efficiency, predictive performance, robustness, ease of implementation and use by an operational team, and other key attributes. Ability to generate clear hydroclimatic ‘storylines’ is particularly notable, given

the black-box reputation of machine learning and the need for such geophysical explanations in operational WSF practice. Satisfying these criteria is in turn a required step in $M^4$ adoption by NRCS as the basis for the next generation of the largest stand-alone operational WSF system in the western US, which to our knowledge will be the first successful migration of AI into a genuinely operational large-scale river prediction system. The result demonstrates that past roadblocks to operationalization of machine learning in hydrology can be overcome with careful and collaborative multi-disciplinary design, and in particular by a development philosophy that focuses on first identifying the practical operational needs of SDOs and then working hand-in-hand with the operational community to develop suitably purpose-specific machine learning solutions that meet those needs. This may set a positive precedent for transitioning AI from research to practice in water resource science, engineering, and management.

Going forward, at least three general research and development directions are apparent. The first is a production software environment to facilitate large-scale operational deployment of $M^4$ at NRCS by serving as a platform and interface for the prediction engine. The prototype platform, currently under development in collaboration with research partners, is based on a Modeling-as-a-Service (MaaS) construct implemented on a private cloud (David et al., 2014). Functionalities span database linkages, a graphical user interface, multi-user server-based implementation amenable to distributed and remote computing, straightforward prediction engine version updating, and interactive capabilities around graphics, mapping, data pre-processing, and forecast distribution post-processing.

Second, the combination of improved accuracy from nonlinear machine learning methods, a robust process for applying these AI algorithms in a physics-aware and operational WSF-specific way, dimensionality reduction, and the flexibility of a modular ensemble framework, may collectively engender faster and more widespread and successful integration of $M^4$ with other methods and products than often feasible in current-generation operational river prediction platforms. For example, numerical climate models offer some seasonal-scale prediction skill, but the results require somewhat elaborate downscaling procedures to use as input to hydrologic process-simulation models, and overall, present operational process-based and statistical hydrology models alike do not appear to be sufficiently accurate and generalizable to capitalize effectively on the additional information these climate models may provide (Gobena and Gan, 2010; Yuan et al., 2013; Mendoza et al., 2017). In contrast, significant improvements in predictive capacity of the nonlinear AI metasystem relative to existing operational WSF technologies, together with its data compression steps and a relatively high level of data-agnosticism, create a platform that should in principle be intrinsically more amenable to quickly and effectively leveraging emerging high-dimensional operational WSF inputs. These predictors include process-based mountain snow (e.g., iSNOBAL; Hedrick et al., 2019) and seasonal-to-subseasonal (S2S) climate (e.g., CFSv2; Saha et al., 2014) model predictions, and snow remote sensing (e.g., MODIS; Tran et al., 2019) and data assimilation (e.g., ASO; Painter et al., 2016) products. This in turn provides potential opportunities for further WSF skill improvements. We are also collaborating with research partners to complement PCA with a new form of non-negative matrix factorization (Vesselinov et al., 2019) to further improve physical interpretability of forecasts (Fleming et al., 2021), and additional supervised learning systems may be integrated beyond the six used here.

Third, the multi-threaded philosophy underlying this collection of semi-independent forecast systems can be easily extended to ingest forecasts from outside sources into its ensemble, including process simulation model-based operational WSFs. Though rare, precedent exists for fusing data-driven and physics-based hydrologic predictions into model-agnostic ensembles (Najafi and Moradkhani, 2016). This creates opportunities to integrate $M^4$ forecasts with, for example, ESPs from US Geological Survey PRMS process-simulation models that NRCS operates at selected locations (Leavesley et al., 2010), leverage innovations in physics-based hydrologic prediction models being developed elsewhere (e.g., WRF-Hydro and the NOAA National Water Model; Cohen et al., 2018), and reinstate formal multi-agency forecast coordination between NRCS and National Weather Service River Forecast Centers that predominantly use ESPs (Pagano et al., 2014). Doing so could improve diversity within the multi-model ensemble and capitalize on the advantages of both AI-based and process-based models (see Sections 3.2.5, 4.1.2, and 4.3). WSFs from alternative statistical models operated by other SDOs, like the Bureau of Reclamation and California Department of Water Resources, could similarly be included where available. Note that while operation of several WSF models across several government agencies may appear inefficient, the multiple governance goals and technical approaches associated with that diversity is known to provide long-term adaptability, robustness, and much-needed operational redundancy for western US water management (see extensive reviews by Doyle, 2012; Hrachowitz and Clark, 2017). The primary drawback for water managers is determining how, in practice, to use these multiple, sometimes partially conflicting, sources of WSF information. Blending multiple hydrologic modeling paradigms into the multi-model framework, as suggested above, would provide a mechanism for addressing this current-practices gap. Combined with aforementioned likely improvements in ability to use emerging predictor sources, this implies the metasystem has potential to grow into a rigorous and nimble integration platform for bringing together multiple data sources and prediction modeling technologies, in turn helping promote more accurate, robust, and usable WSFs across the American West.

## CRediT authorship contribution statement

**Sean W. Fleming:** Conceptualization, Formal analysis, Investigation, Methodology, Project administration, Software, Validation, Visualization, Resources, Writing – original draft, Writing – review & editing, Funding acquisition. **David C. Garen:** Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing. **Angus G. Goodbody:** Conceptualization, Methodology, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing. **Cara S. McCarthy:** Conceptualization, Methodology, Investigation, Resources, Writing – original draft, Writing – review & editing, Project administration, Funding acquisition. **Lexi C. Landers:** Writing – original draft, Writing – review & editing, Visualization.

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements and data availability

All data used in this study are publicly available from wcc.sc.egov.usda.gov/reportGenerator.

## References

Abrahart, R.J., Anctil, F., Coulibaly, P., Dawson, C.W., Mount, N.J., See, L.M., Shamseldin, A.Y., Solomatine, D.P., Toth, E., Wilby, R.L., 2012. Two decades of anarchy? Emerging themes and outstanding challenges for neural network river forecasting. Prog. Phys. Geogr. 36, 480–513.

Barnett, T.P., Adam, J.C., Lettenmaier, D.P., 2005. Potential impacts of a warming climate on water availability in snow-dominated regions. Nature 438, 303–309.

Beckers, J.V.L., Weerts, A.H., Tijdeman, E., Welles, E., 2016. ENSO-conditioned weather resampling method for seasonal ensemble streamflow prediction. Hydrol. Earth Syst. Sci. 20, 3277–3287.

Beven, K., Binley, A., 1992. The future of distributed models: model calibration and uncertainty prediction. Hydrol. Process. 6, 279–298.

Bourdin, D.R., Fleming, S.W., Stull, R.B., 2012. Streamflow modelling: a primer on applications, approaches, and challenges. Atmos. Ocean 50, 507–536.

Bourdin, D.R., Nipen, T.N., Stull, R.B., 2014. Reliable probabilistic forecasts from an ensemble reservoir inflow forecasting system. Water Resour. Res. 50, 3108–3130.
Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140.
Breiman, L., 2001. Random forests. Mach. Learn. 45, 5–32.
Burnham, K.P., Anderson, D.R., 2002. Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach, 2nd ed. Springer, New York, NY.
Bureau of Reclamation, 2016. SECURE Water Act Section 9503(c) – Reclamation Climate Change and Water 2016. Bureau of Reclamation, Policy and Administration, Denver, CO.
Cannon, A.J., McKendry, I.G., 2002. 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. 22, 1687–1708.
Clarke, G.K.C., Jarosch, A.H., Anslow, F.S., Radic, V., Menounos, B., 2015. Projected deglaciation of western Canada in the twenty-first century. Nat. Geosci. 8, 372–377.
Cohen, S., Prashievicz, S., Maidment, D.R., 2018. National water model. J. Am. Water Resour. Assoc. 54, 767–770.
Cunderlik, J.M., Fleming, S.W., Jenkinson, R.W., Thiemann, M., Kouwen, N., Quick, M., 2013. Integrating logistical and technical criteria into a multiteam, competitive watershed model ranking procedure. ASCE J. Hydrol. Eng. 18, 641–654.
David, O., Lloyd, W., Rojas, K., Arabi, M., Geter, F., Ascough, J., Green, T., Leavesley, G., Carlson, J., 2014. Modeling-as-a-Service (MaaS) using the Cloud Services Innovation Platform (CSIP). In: Ames, D.P., Quinn, N.W.T., Rizzoli, A.E. (Eds.), Proceedings, International Congress on Environmental Modelling and Software. San Diego, CA, pp. 243–250.
Daw, C.S., Finney, C.E.A., Tracy, E.R., 2003. A review of symbolic analysis of experimental data. Rev. Sci. Instrum. 74, 915–930.
Demargne, J., Wu, L., Regonda, S.K., Brown, J.D., Lee, H., He, M., Seo, D.-J., Hartman, R., Herr, H.D., Fresch, M., Schaake, J., Zhu, Y., January 2014. The science of NOAA's operational hydrologic ensemble forecast service. Bull. Am. Meteorol. Soc. 80–98.
Doyle, M.W., 2012. America's rivers and the American experiment. J. Am. Water Resour. Assoc. 48, 820–837.
Eldaw, A.K., Salas, J.D., Garcia, L.A., 2003. Long-range forecasting of the Nile River flows using climatic forcing. J. Appl. Meteorol. 42, 890–904.
Finney, C.E.A., Green Jr, J.B., Daw, C.W., 1998. Symbolic Time-Series Analysis of Engine Combustion Measurements, SAE Technical Paper 980624. SAE International, Warrendale PA. https://doi.org/10.4271/980624.
Fleming, S.W., 2007. Artificial neural network forecasting of nonlinear Markov processes. Can. J. Phys. 85, 279–294.
Fleming, S.W., Barton, M., 2015. Climate trends but little net water supply shifts in one of Canada's most water-stressed regions over the last century. J. Am. Water Resour. Assoc. 51, 833–841.
Fleming, S.W., Bourdin, D.R., Campbell, D., Stull, R.B., Gardner, T., 2015. Development and operational testing of a super-ensemble artificial intelligence flood-forecast model for a Pacific Northwest river. J. Am. Water Resour. Assoc. 51, 502–512.
Fleming, S.W., Dahlke, H.E., 2014. Parabolic Northern-Hemisphere river flow teleconnections to El Niño-Southern Oscillation and the Arctic Oscillation. Environ. Res. Lett. 9 https://doi.org/10.1088/1748-9326/9/10/104007.
Fleming, S.W., Goodbody, A.G., 2019. A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. IEEE Access 7, 119943–119964.
Fleming, S.W., Gupta, H.V., 2020. The physics of river prediction. Phys. Today 73, 46–52.
Fleming, S.W., Vesselinov, V.V., Goodbody, A.G., 2021. Augmenting geophysical interpretation of data-driven operational water supply forecast modeling for a western US river using a hybrid machine learning approach. J. Hydrol. 597, 126327.
Garen, D.C., 1992. Improved techniques in regression-based streamflow volume forecasting. J. Water Resour. Plann. Manage. 118, 654–669.
Garen, D.C., 1998. ENSO indicators and long-range climate forecasts: usage in seasonal streamflow volume forecasting in the western United States, American Geophysical Union Fall Conference, San Francisco, CA.
Gelfan, A.N., Motovilov, Y.G., 2009. Long-term hydrological forecasting in cold regions: retrospect, current status, and prospect. Geogr. Compass 3 (5), 1841–1864.
Glabau, B., Nielsen, E., Mylvahanan, A., Stephan, N., Frans, C., Duffy, K., Giovando, J., Johnson, J., 2020. Climate and Hydrology Datasets for RMJOC Long-Term Planning Studies, Second Edition, Part II: Columbia River Reservoir Regulation and Operations – Modeling and Analyses. River Management Joint Operating Committee. Available at www.bpa.gov/p/Generation/Hydro/Documents/RMJOC-II_Part_II.PDF.
Glantz, M.H., 1982. Consequences and responsibilities in drought forecasting: the case of Yakima, 1977. Water Resour. Res. 18, 3–13.
Gobena, A.K., Gan, T.Y., 2009. Statistical ensemble seasonal streamflow forecasting in the South Saskatchewan River Basin by a modified nearest neighbors resampling. ASCE J. Hydrol. Eng. 14, 628–639.
Gobena, A.K., Gan, T.Y., 2010. Incorporation of seasonal climate forecasts in the ensemble streamflow prediction system. J. Hydrol. 385, 336–352.
Gobena, A.K., Weber, F.A., Fleming, S.W., 2013. The role of large-scale climate modes in regional streamflow variability and implications for water supply forecasting: a case study of the Canadian Columbia Basin. Atmos. Ocean 51, 380–391.
Grantz, K., Rajagopalan, B., Clark, M., Zagona, E., 2005. A technique for incorporating large-scale climate information in basin-scale ensemble streamflow forecasts. Water Resour. Res. 41, W10410. https://doi.org/10.1029/2004WR003467.
Guihan, R., 2014. Integrating Emerging River Forecast Center Streamflow Products into the Salt Lake City Parley's Drinking Water System. University of Massachusetts-Amherst, Masters Degree Project.
Guyon I, Bennett K, Cawley G, Escalante HJ, Escalera S, Ho TK, Macià N, Ray B, Saeed M, Statnikov A, Viegas E. 2015. Design of the 2015 ChaLearn AutoML challenge. Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12-17 July 2015, pp. 1-8.
Hagedorn, R., Doblas-Reyes, F.J., Palmer, T.N., 2005. The rationale behind the success of multi-model ensembles in seasonal forecasting – I. basic concept. Tellus 57A, 219–233.
Hamlet, A.F., Huppert, D., Lettenmaier, D.P., 2002. Economic value of long-lead streamflow forecasts for Columbia River hydropower. J. Water Resour. Plann. Manage. 128, 91–101.
Harpold, A., Dettinger, M., McAfee, S., Rajagopoal, S., Sturtevant, J., 2020. Seasonal water supply forecasting in the western US under declining snowpack. Southwest Climate Adaptation Center Stakeholder Meeting, May 6, 2020, Reno, NV.
Harpold, A.A., Sutcliffe, K., Clayton, J., Goodbody, A., Vazquez, S., 2016. Does including soil moisture observations improve operational streamflow forecasts in snow-dominated watersheds? J. Am. Water Resour. Assoc. 53, 179–196.
Harrison, B., Bales, R., 2016. Skill assessment of water supply forecasts for western Sierra Nevada watersheds. J. Hydrol. Eng. 21 https://doi.org/10.1061/(ASCE)HE.1943-5584.0001327.
Hartmann, H.C., Bales, R., Sorooshian, S., 2002. Weather, climate, and hydrologic forecasting for the US Southwest: a survey. Clim. Res. 21, 239–258.
Hedrick, A.R., Marks, D., Marshall, H.P., McNamara, J., Havens, S., Trujillo, E., Sandusky, M., Robertson, M., Johnson, M., Bormann, K.J., Painter, T.H., 2019. From drought to flood: a water balance analysis of the Tuolumne River basin during extreme conditions (2015–2017). Hydrol. Process. 34, 2560–2574.
Hekkert, P., Snelders, D., van Wieringen, P.C.W., 2003. 'Most advanced, yet acceptable': typicality and novelty as joint predictors of aesthetic preference in industrial design. Br. J. Psychol. 94, 111–124.
Hey, T., Tansley, S., Tolle, K., 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, WA.
Hill, J., Mulholland, G., Persson, K., Seshadri, R., Wolverton, C., Meredig, B., 2016. Materials science with large-scale data and informatics: unlocking new opportunities. Mater. Res. Soc. Bull. 41, 399–409.
Hoekema, D.J., Ryu, J.H., 2013. Evaluating economic impacts of water conservation and hydrological forecasts in the Salmon Tract, southern Idaho. Trans. Am. Soc. Agric. Biol. Eng. 56, 1399–1410.
Hrachowitz, M., Clark, M.P., 2017. The complementary merits of competing modelling philosophies in hydrology. Hydrol. Earth Syst. Sci. 21, 3953–3973.
Hsieh, W.W., 2009. Machine Learning Methods in the Environmental Sciences. Cambridge University Press, Cambridge, UK.
Hsieh, W.W., Yuval, Li J., Shabbar, A., Smith, S., 2003. Seasonal prediction with error estimation of Columbia River streamflow in British Columbia. J. Water Resour. Plann. Manag. 129, 146–149.
Hsu, K.-L., Gupta, H.V., Sorooshian, S., 1995. Artificial neural network modeling of the rainfall-runoff process. Water Resour. Res. 31, 2517–2530.
Hyndman, R.J., Athanasopoulos, G., 2013. Forecasting: principles and practice. OTexts, Melbourne, Australia. http://otexts.org/fpp/. Accessed on 22 September 2017.
James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning. Springer, New York, NY.
Jiang, S., Zheng, Y., Babovic, V., Tian, Y., Han, F., 2018. A computer vision approach to fusing spatiotemporal data for hydrological modeling. J. Hydrol. 567, 25–40.
Kalra, A., Miller, W.P., Lamb, K.W., Ahmad, S., Piechota, T., 2013. Using large-scale climate patterns for improving long lead time streamflow forecasts for Gunnison and San Juan river basins. Hydrol. Process. 27, 1543–1559.
Karpatne, A., Atluri, G., Faghmous, J.H., Steinback, M., Banerjee, A., Ganguly, A., Shekhar, S., Samatova, N., Kumar, V., 2017. Theory-guided data science: a new paradigm for scientific discover from data. IEEE Trans. Knowl. Data Eng. 29, 2318–2331.
Kennedy, A.M., Garen, D.C., Koch, R.W., 2009. The association between climate teleconnection indices and Upper Klamath seasonal streamflow: Trans-Niño index. Hydrol. Process. 23, 973–984.
Kratzert, F., Klotz, D., Brenner, C., Schulz, K., Herrnegger, M., 2018. Rainfall-runoff modelling using Long Short-Term Memory (LSTM) networks. Hydrol. Earth Syst. Sci. 22, 6005–6022.
Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A.K., Hochreiter, S., Nearing, G.S., 2019. Toward improved predictions in ungauged basins: exploiting the power of machine learning. Water Resour. Res. 55, 11344–11354.
Launchberry, J., 2021. A DARPA Perspective on Artificial Intelligence. Defense Advanced Research Projects Agency, www.darpa.mil/about-us/darpa-perspective-on-ai, accessed 21 May 2021.
Leavesley, G., David, O., Garen, D.C., Goodbody, A.G., Lea, J., Marron, J., Perkins, T., Strobel, M., Tama, R., 2010. A modeling framework for improved agricultural water-supply forecasting. In: Proceedings, Joint Federal Interagency Hydrologic Modeling Conference, Las Vegas, NV, June 28-July 1, 2010, 12 p.
Lehner, F., Wood, A.W., Llewellyn, D., Blatchford, D.B., Goodbody, A.G., Pappenberger, F., 2017. Mitigating the impacts of climate nonstationarity on seasonal streamflow predictability in the US southwest. Geophys. Res. Lett. 44, 12208–12217.
Lima, A.R., Hsieh, W.W., Cannon, A.J., 2017. Variable complexity online sequential extreme learning machine, with applications to streamflow prediction. J. Hydrol. 555, 983–994.
Lukas, J., Harding, B., 2020. Current Understanding of Colorado River Basin Climate and Hydrology, Chap. 2 in Colorado River Basin Climate and Hydrology: State of the Science, edited by J. Lukas and E. Payton, p. 42-81. Western Water Assessment, University of Colorado Boulder, Boulder CO.

Mahabir, C., Hicks, F.E., Fayek, A.R., 2003. Application of fuzzy logic to forecast seasonal runoff. Hydrol. Process. 17, 3749–3762.
McGovern, A., Lagerquist, R., Gagne II, D.J., Jergensen, G.E., Elmore, K.L., Homeyer, C. F., Smith, T., 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. (November), 2175–2199.
McGuire, M., Wood, A.W., Hamlet, A.F., Lettenmaier, D.P., 2006. Use of satellite data for streamflow and reservoir storage forecasts in the Snake River Basin. ASCE J. Water Resour. Plann. Manag. 132, 97–110.
Mendoza, P.A., Wood, A.W., Clark, E., Rothwell, E., Clark, M.P., Nijssen, B., Brekke, L.D., Arnold, J.R., 2017. An intercomparison of approaches for improving operational seasonal streamflow forecasts. Hydrol. Earth Syst. Sci. 21, 3915–3935.
Meredig, B., 2018. Solving industrial materials problems with machine learning. Presentation at the American Physical Society March Meeting, Los Angeles, CA.
Micovic, Z., Quick, M.C., 2009. Investigation of the model complexity required in runoff simulation at different time scales. Hydrol. Sci. J. 54, 872–885.
Minns, A.W., Hall, M.J., 1996. Artificial neural networks as rainfall-runoff models. Hydrol. Sci. J. 41, 399–417.
Minxue, H., Whitin, B., Hartman, R., Henkel, A., Fickenschers, P., Staggs, S., Morin, A., Imgarten, M., Haynes, A., Russo, M., 2016. Verification of ensemble water supply forecasts for Sierra Nevada watersheds. Hydrology 3, 35. https://doi.org/10.3390/hydrology3040035.
Monteleoni, C., Schmidt, G.A., Saroha, S., Asplund, 2011. Tracking climate models. Journal of Statistical Analysis and Data Mining 4, 372–392.
Moradkhani, H., Meier, M., 2010. Long-lead water supply forecast using large-scale climate predictors and independent component analysis. J. Hydrol. Eng. 15, 744–762.
Najafi, M.R., Moradkhani, H., 2016. Ensemble combination of seasonal streamflow forecasts. J. Hydrol. Eng. 21 https://doi.org/10.1061/(ASCE)HE.1943-5584.0001250.
Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, G.V., 2021. What role does hydrological science play in the age of machine learning? Water Resour. Res. 57 e2020WR028091.
Newman, M., 2007. Interannual to decadal predictability of tropical and north Pacific sea surface temperatures. J. Clim. 20, 2333–2356.
Newman, M., 2012. An empirical benchmark for Pacific Ocean variability and predictability. Canadian Centre for Climate Modeling and Analysis–Pacific Climate Impacts Consortium Joint Seminar, 11 September 2012, Victoria, BC.
O'Connor, J.E., Grant, G.E., Haluska, T.L., 2003. Overview of geology, hydrology, geomorphology, and sediment budget of the Deschutes River basin, Oregon. In: O'Connor, J.E., Grant, G.E. (Eds.), A Peculiar River: Geology, Geomorphology, and Hydrology of the Deschutes River, Oregon, Water Science and Application 7. American Geophysical Union, Washington DC.
Oubeidillah, A.A., Tootle, G.A., Moser, C., Piechota, T., Lamb, K., 2011. Upper Colorado River and Great Basin streamflow and snowpack forecasting using Pacific oceanic–atmospheric variability. J. Hydrol. 410, 169–177.
Packard, N.H., Crutchfield, J.P., Farmer, J.D., Shaw, R.S., 1980. Geometry from a time series. Phys. Rev. Lett. 45, 712–716.
Pagano, T.C., Garen, D.C., Sorooshian, S., 2004. Evaluation of official western US seasonal water supply outlooks, 1922–2002. J. Hydrometeorol. 5, 896–909.
Pagano, T.C., Garen, D.C., Perkins, T.R., Pasteris, P.A., 2009. Daily updating of operational statistical seasonal water supply forecasts for the western US. J. Am. Water Resour. Assoc. 45, 767–778.
Pagano, T.C., Wood, A.W., Werner, K., Tama-Sweet, R., 2014. Western US Water Supply Forecasting: a tradition evolves. Eos, Trans., AGU 95, 28–29.
Painter, T.H., Berisford, D.F., Boardman, J.W., Bormann, K.J., Deems, J.S., Gehrke, F., Hedrick, A., Joyce, M., Laidlaw, R., Marks, D., Mattmann, C., McGurk, B., Ramirez, P., Richardson, M., Skiles, S.M., Seidel, F.C., Winstral, A., 2016. The Airborne Snow Observatory: fusion of scanning lidar, imaging spectrometer, and physically-based modeling for mapping snow water equivalent and snow albedo. Rem. Sens. Environ. 184, 139–152.
Penland, C., 1996. A stochastic model of IndoPacific sea surface temperature anomalies. Physica D 98, 534–558.
Peñuela, A., Hutton, C., Pianosi, F., 2020. Assessing the value of seasonal hydrological forecasts for improving water resource management: insights from a pilot application in the UK. Hydrol. Earth Syst. Sci. 24, 6059–6073.
Perkins, T.R., Pagano, T.C., Garen, D.C., 2009. Innovative operational seasonal water supply forecasting technologies. J. Soil Water Conserv. 64, 15–17.
Regonda, S.K., Rajagopalan, B., Clark, M., Zagon, E., 2006a. A multimodel ensemble forecast framework: application to spring seasonal flows in the Gunnison River Basin. Water Resour. Res. 42 https://doi.org/10.1029/2005WR004653.
Regonda, S.K., Rajagopalan, B., Clark, M., 2006b. A new method to produce categorical streamflow forecasts. Water Resour. Res. 42 https://doi.org/10.1029/2006WR004984.
Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566, 195–204.
Reisner, M., 1986. Cadillac Desert. Viking, New York, NY.
Risley JC, Gannett MW, Lea JK, Roehl EA Jr. 2005. An Analysis of Statistical Methods for Seasonal Flow Forecasting in the Upper Klamath River Basin of Oregon and California. Scientific Investigations Report 2005-5177, US Geological Survey, Reston, VA.
Robertson, D.E., Pokhrel, P., Wang, Q.J., 2013. Improving statistical forecasts of seasonal streamflows using hydrological model output. Hydrol. Earth Syst. Sci. 17, 579–593.
Rogers, E., 2003. Diffusion of Innovations. Free Press, New York, NY.
Rosenberg, E.A., Wood, A.W., Steinemann, A.C., 2011. Statistical applications of physically based hydrologic models to seasonal streamflow forecasts. Water Resour. Res. 47 https://doi.org/10.1029/2010WR010101.
Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., Behringer, D., Hou, Y.T., Chuang, H.Y., Iredell, M., Ek, M., Meng, J., Yang, R., Mendez, M.P., Van den Dool, H., Zhang, Q., Wang, W., Chen, M., Becker, E., 2014. The NCEP Climate Forecast System Version 2. J. Clim. 27, 2185–2208.
Sankasubramanian, A., Vogel, R.M., Limbrunner, J.F., 2001. Climate elasticity of streamflow in the United States. Water Resour. Res. 37, 1771–1781.
Schnorbus, M.A., Cannon, A.J., 2014. Statistical emulation of streamflow predictions from a distributed hydrological model: application to CMIP3 and CMIP5 climate projections for British Columbia, Canada. Water Resour. Res. 50, 8907–8926.
Seo, D.-J., Koren, V., Cajina, N., 2003. Real-time variational assimilation of hydrologic and hydrometeorological data into operational hydrologic forecasting. J. Hydrometeorol. 4, 627–641.
Serafin F, David O, Carlson JR, Green TR, Rigon R. Bridging technology transfer boundaries: integrated cloud services deliver results of nonlinear process models as surrogate model ensembles. In preparation for submission to Environmental Modelling and Software.
Schwab, K., 2017. The Fourth Industrial Revolution. Penguin Random House, New York, NY.
Shrestha, R.R., Cannon, A.J., Schnorbus, M.A., Zwiers, F.W., 2017. Projecting future nonstationary extreme streamflow for the Fraser River, Canada. Clim. Change 145, 289–303.
Singh, V.P., Woolhiser, D.A., 2002. Mathematical modeling of watershed hydrology. ASCE J. Hydrol. Eng. 7, 270–292.
Thornton C, Hutter F, Hoos HH, Leyton-Brown K. 2013. Auto-weka: Combined selection and hyperparameter optimization of classification algorithms. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pages 847–855, New York, NY, USA, 2013, doi:10.1145/2487575.2487629.
Tran, H., Nguyen, P., Ombadi, M., Hsu, K.-l., Sorooshian, S., Qing, X., 2019. A cloud-free MODIS snow cover dataset for the contiguous United States from 2000 to 2017. Scientific Data 6. https://doi.org/10.1038/sdata.2018.300.
Trubilowicz JW, Chorlton E, Déry SJ, Fleming SW. 2015. Satellite remote sensing for water resource applications in British Columbia. Innovation, Journal of the Association of Professional Engineers and Geoscientists of British Columbia, April/May, 18-20.
Vano, J.A., Das, T., Lettenmaier, D.P., 2012. Hydrologic sensitivities of Colorado River runoff to changes in precipitation and temperature. J. Hydrometeorol. 13, 932–949.
Vesselinov, V.V., Mudunuru, M.K., Karra, S., O'Malley, D., Alexandrov, B.S., 2019. Unsupervised machine learning based on non-negative tensor factorization for analyzing reactive-mixing. J. Comput. Phys. 15, 85–104.
Weber, F., Garen, D., Gobena, A., 2012. Invited commentary: themes and issues from the workshop 'Operational River Flow and Water Supply Forecasting'. Canad. Water Resour. J./Revue canadienne des resources hydriques 37, 151–161.
Whateley, S., Palmer, R.N., Brown, C., 2015. Seasonal hydroclimatic forecasts as innovations and the challenges of adoption by water managers. ASCE J. Water Resour. Plann. Manag. 141, 04014072.
Wiegel, A.P., Liniger, M.A., Appenzeller, C., 2007. The discrete Brier and ranked probability skill scores. Mon. Weather Rev. 135, 118–124.
Wolpert, D.H., 1996. The lack of a priori distinctions between learning algorithms. Neural Comput. 8, 1341–1390.
Wood AW, Lettenmaier DP. 2006. A test bed for new seasonal hydrologic forecasting approaches in the western United States. Bulletin of the American Meteorological Society, December, 1699-1712.
Wood AW, Woelders L, Lukas J. 2020. Streamflow Forecasting, Chap. 8 in Colorado River Basin Climate and Hydrology: State of the Science, edited by J. Lukas and E. Payton, 287-333. Western Water Assessment, University of Colorado Boulder, Boulder, CO.
Yao, H., Georgakakos, A., 2001. Assessment of Folsom Lake response to historical and potential future climate scenarios, 2, reservoir management. J. Hydrol. 249, 176–196.
Yuan, X., Wood, E.F., Roundy, J.K., Pan, M., 2013. CFSv2-based seasonal hydroclimatic forecasts over the conterminous United States. J. Clim. 26, 4828–4847.
Zhang H, Zhang Z. Feedforward networks with monotone constraints, in Proc. IEEE Int. Joint Conf. Neural Netw., Washington, DC, USA, vol. 3, Jul. 1999, pp. 1820-1823.

# APPENDIX B

Appendix B contains the following:

> Fleming SW, Goodbody AG. 2019. A machine learning metasystem for robust probabilistic nonlinear regression-based forecasting of seasonal water availability in the US West. *IEEE Access*, 7, 119943-119964.

This article provides a machine learning-focused description of M$^4$, including detailed technical descriptions of what's under the hood, some initial testing to WSF problems, and some key performance traits around the multi-model aspect of the metasystem.

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 $\Delta t$ 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

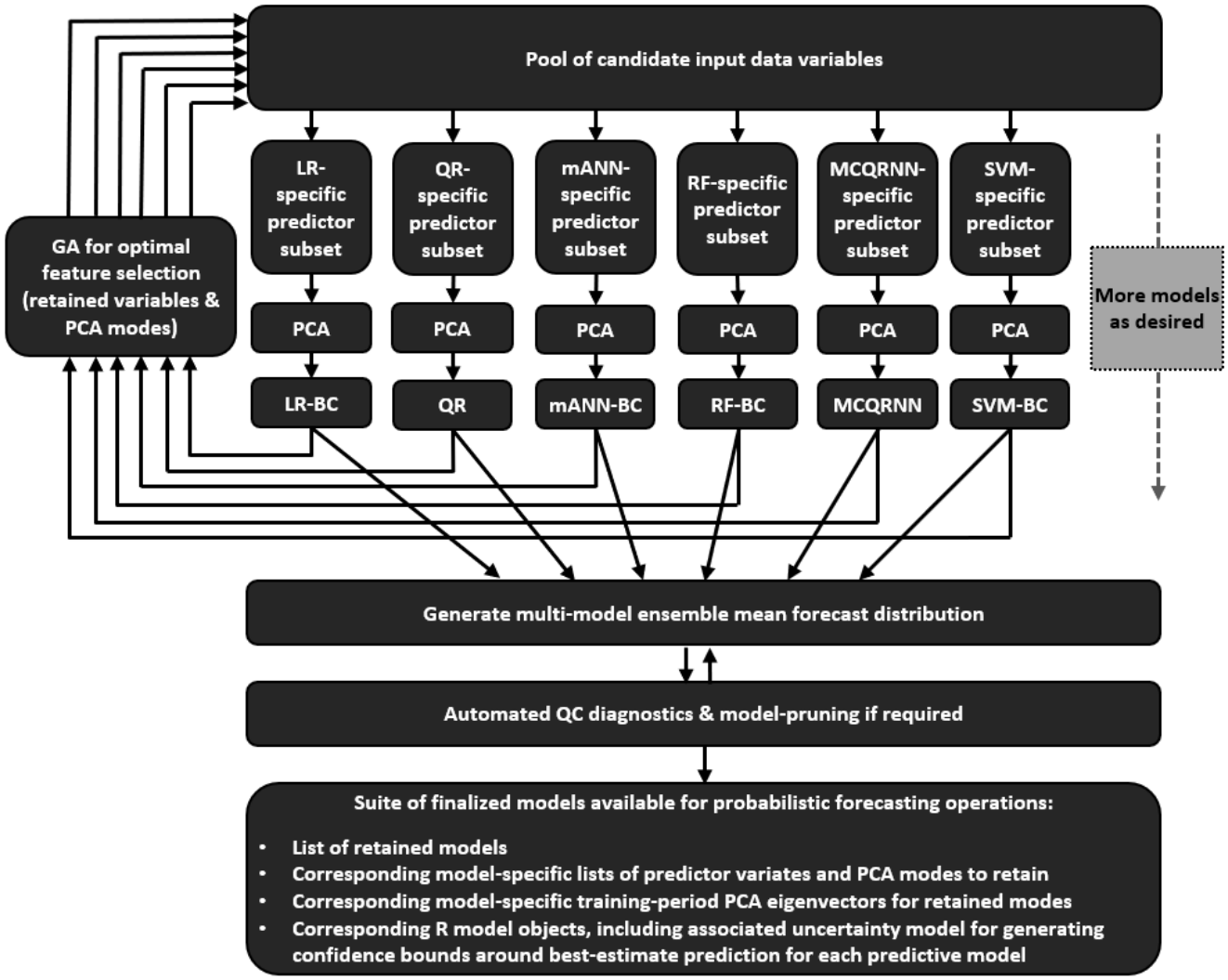


**FIGURE 1.** Approximate representation of some major components and workflows in the model-building step. Items listed in the bottom box are the main products of the model-building process and are subsequently used in forecast operations. Not all system components and workflows are illustrated here. 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, $\lambda$:

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

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

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$^{\text{th}}$ 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 $m^{\text{th}}$ 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 $m^{\text{th}}$ 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 $\psi$ 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 $\alpha^{\text{th}}$ 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 $\alpha^{\text{th}}$ quantile prediction bounds. Forward and inverse Box-Cox transforms were performed using the `forecast` package [37], which also finds optimal $\psi$. 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 $\varepsilon$-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 $\beta$ 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 $\alpha$ 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); $\varepsilon$, $\gamma$, 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 $\gamma$ 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 $\varepsilon$ and $C$. The search is inefficient, so removing $\gamma$ 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

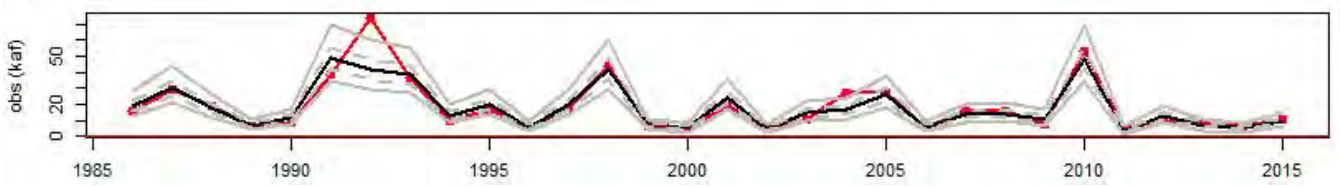


**FIGURE 2.** Observations (red dashed line with dots), best-estimate predictions (black line), and associated 0.10 and 0.90 quantile (solid gray lines) and 0.30 and 0.70 quantile (dashed gray lines) prediction intervals for spring-summer flow volume in kaf of the Gila River. Horizontal red line denotes 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/S0218001416550020X.

[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.

# APPENDIX C

Appendix C contains the following:

> Fleming SW, Garen DC. 2022. Simplified cross-validation in principal component regression (PCR) and PCR-like machine learning for water supply forecasting. *Journal of the American Water Resources Association*, 58, 517-524.

This article describes some detailed technical aspects and comparative testing around the LOOCV techniques used in M$^4$.

JOURNAL OF THE AMERICAN WATER RESOURCES ASSOCIATION

Vol. 58, No. 4 AMERICAN WATER RESOURCES ASSOCIATION August 2022

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

*Sean W. Fleming, and David C. Garen*

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

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

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

## INTRODUCTION

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

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

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

*Citation*: Fleming, S.W., and D.C. Garen. 2022. "Simplified Cross-Validation in Principal Component Regression (PCR) and PCR-Like Machine Learning for Water Supply Forecasting." *Journal of the American Water Resources Association* 58 (4): 517–524. https://doi.org/10.1111/1752-1688.13007.

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

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

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

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

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

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

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

## DATA AND METHOD

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

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

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

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

**Algorithm 1:** Scenario 1 (full) LOOCV procedure

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

**Step 2** Perform PCA on input variables

**Step 2a** Standardize each length ($N$) input time series to zero mean and unit variance on the basis of its sample mean and variance and combine into data matrix, $\mathbf{X}$

**Step 2b** Calculate correlation matrix, $\mathbf{C} = N^{-1}\,\mathbf{X}\,\mathbf{X}^{T}$

**Step 2c** Find eigenvectors $\mathbf{E}$ of $\mathbf{C}$, and scores matrix $\mathbf{A} = \mathbf{E}^{T}\mathbf{X}$ for $N$ samples and $M$ modes

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

**Step 3a** Select subset of PCA modes in $\mathbf{A}$ to use as predictive features, $\mathbf{A'} = a_{m,n}$, $m \in (1,2,\ldots,M)$, $n = 1,N$

**Step 3b** Train a supervised learning algorithm, $f$, to predict expectation values of target, $<\boldsymbol{q}>$ , from $\mathbf{A'}$

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

**Step 4a** For each $t$=1,$N$ calculate expectation value of target, $<q_t>$ , using a supervised model retrained on all data except those from $t^{th}$ sample

**Step 4a(i)** Create data subset leaving out one sample, ${}^{LOO}q_n = q_n \;\forall\; n \neq t$ and ${}^{LOO}x_{m,n} = x_{m,n} \;\forall\; n \neq t$

**Step 4a(ii)** Standardize each length ($N$-1) input variable time series to zero mean and unit variance on the basis of its sample mean and variance, and combine into ${}^{LOO}\mathbf{X}$

**Step 4a(iii)** Calculate correlation matrix, ${}^{LOO}\mathbf{C} = (N\text{-}1)^{-1}\,{}^{LOO}\mathbf{X}\,{}^{LOO}\mathbf{X}^{T}$

**Step 4a(iv)** Find eigenvectors ${}^{LOO}\mathbf{E}$ of ${}^{LOO}\mathbf{C}$, and scores matrix ${}^{LOO}\mathbf{A} = {}^{LOO}\mathbf{E}^{T}\,{}^{LOO}\mathbf{X}$

**Step 4a(v)** Create features matrix by selecting the same PCA modes from ${}^{LOO}\mathbf{A}$ as retained in Step 3a, ${}^{LOO}\mathbf{A'} = {}^{LOO}a_{m,n}$, $m \in (1,2,\ldots,M)$, $n = 1,N\text{-}1$

**Step 4a(vi)** Train predictive model, ${}^{LOO}f$, on ${}^{LOO}\mathbf{A'}$ and ${}^{LOO}\boldsymbol{q}$

**Step 4a(vii)** Retrieve left-out observational data, ${}^{CV}q_t = q_{n=t}$ and ${}^{CV}x_{m,t} = x_{m,n=t}$

**Step 4a(viii)** Standardize the left-out sample for each of the $M$ input variables, ${}^{CV}x_{m,t}$ , using sample means and variances of ${}^{LOO}x_{m,n}$ found in Step 4a(ii), and combine into ${}^{CV}\mathbf{X}$

**Step 4a(ix)** ${}^{CV}\mathbf{A'} = {}^{LOO}\mathbf{E}^{T}\;{}^{CV}\mathbf{X}$ using the same modes $m \in (1,2,\ldots,M)$ retained in Steps 3a and 4a(v) and the same ${}^{LOO}\mathbf{E}$ calculated in Step 4a(iv)

**Step 4a(x)** Estimate $<{}^{CV}q_{n=t}>$ using model, ${}^{LOO}f$ , forced by the features in ${}^{CV}\mathbf{A'}$

**Step 4b** Obtain performance measures (RMSE, $R^2$) by comparing $<{}^{CV}\boldsymbol{q}>$ and $\boldsymbol{q}$ time series

**Algorithm 2:** Scenario 2 (partial) LOOCV procedure

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

**Step 2** Perform PCA on input variables

**Step 2a** Standardize each length ($N$) predictor time series to zero mean and unit variance on the basis of its sample mean and variance and combine into data matrix, $\mathbf{X}$

**Step 2b** Calculate correlation matrix, $\mathbf{C} = N^{-1}\,\mathbf{X}\,\mathbf{X}^{T}$

**Step 2c** Find eigenvectors $\mathbf{E}$ of $\mathbf{C}$, and scores matrix $\mathbf{A} = \mathbf{E}^{T}\mathbf{X}$ for $N$ samples and $M$ modes

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

**Step 3a** Select subset of PCA modes in $\mathbf{A}$ to use as predictive features, $\mathbf{A'} = a_{m,n}$, $m \in (1,2,\ldots,M)$, $n$=1,$N$

**Step 3b** Train a supervised learning algorithm, $f$, to predict expectation values of target, $<\boldsymbol{q}>$ , from $\mathbf{A'}$

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

**Step 4a** For each $t$=1,$N$ calculate expectation value of target, $<q_t>$ , using a supervised model retrained on all data except those from the $t^{th}$ sample

**Step 4a(i)** Create subset of target and selected features leaving out one sample, ${}^{LOO}q_n = q_n \;\forall\; n \neq t$ and ${}^{LOO}\mathbf{A'} = \mathbf{A'} \;\forall\; n \neq t$, where $\mathbf{A'}$ is the original scores matrix found in Step 3a

**Step 4a(ii)** Train predictive model, ${}^{LOO}f$, on ${}^{LOO}\mathbf{A'}$ and ${}^{LOO}\boldsymbol{q}$

**Step 4a(iii)** Retrieve observational data for target and selected features at the left-out sample time, ${}^{CV}q_{n=t}$ and ${}^{CV}\mathbf{A'} = \mathbf{A'}$ for $n = t$ only

**Step 4a(iv)** Estimate $<{}^{CV}q_{n=t}>$ using model, ${}^{LOO}f$ , forced by features, ${}^{CV}\mathbf{A'}$

**Step 4b** Obtain performance measures (RMSE, $R^2$) by comparing $<{}^{CV}\boldsymbol{q}>$ and $\boldsymbol{q}$ time series

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

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

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

## RESULTS AND DISCUSSION

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

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

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

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


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

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

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

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

## CONCLUSIONS

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

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

### *DATA AVAILABILITY STATEMENT*

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

## AUTHOR CONTRIBUTIONS

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

## SUPPORTING INFORMATION

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

## LITERATURE CITED

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

# Supplementary Information

**Fleming and Garen, JAWRA, 2022**

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

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

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

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


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

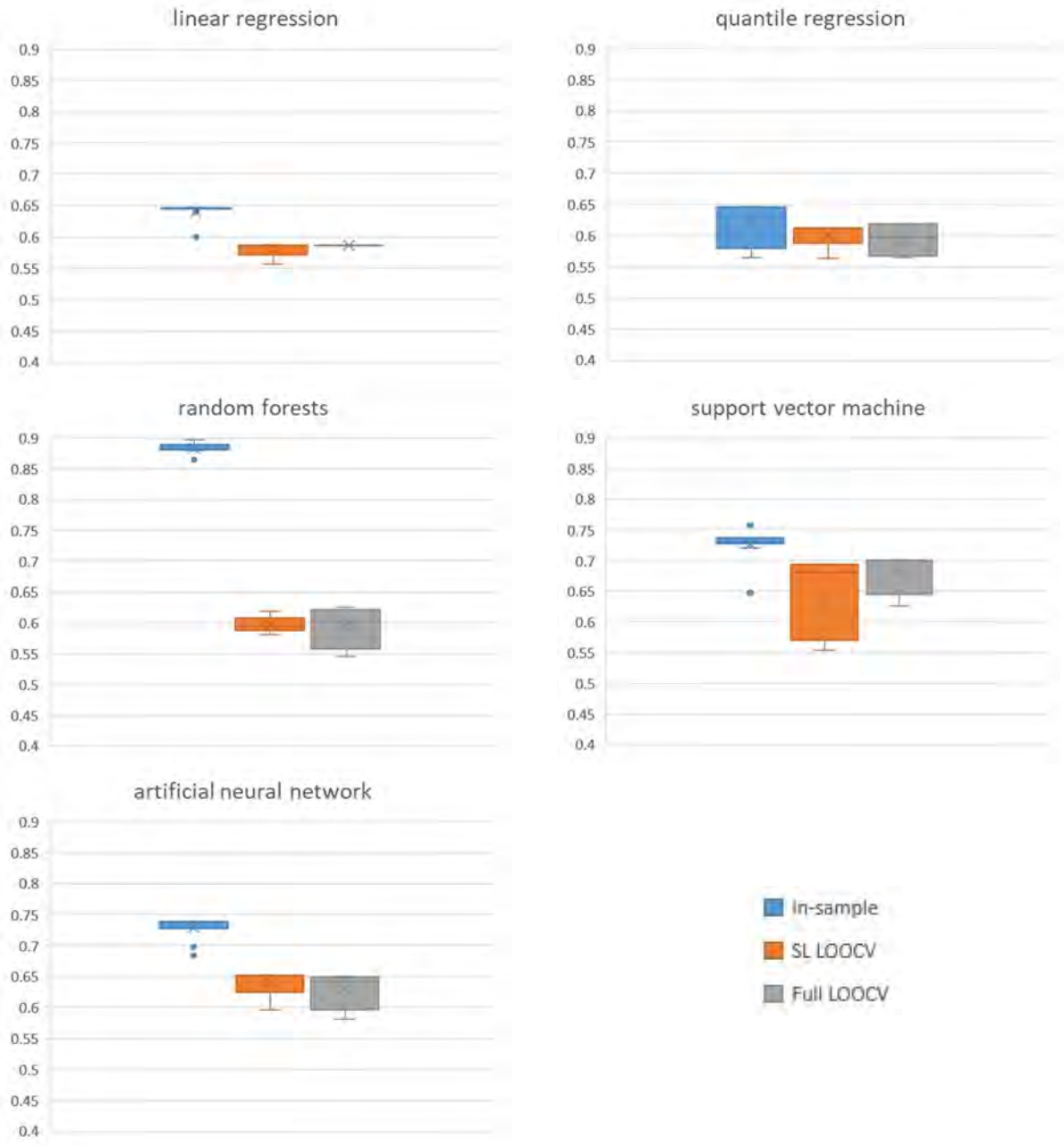


**Figure S2.** As in Figure S1 but for $R^2$

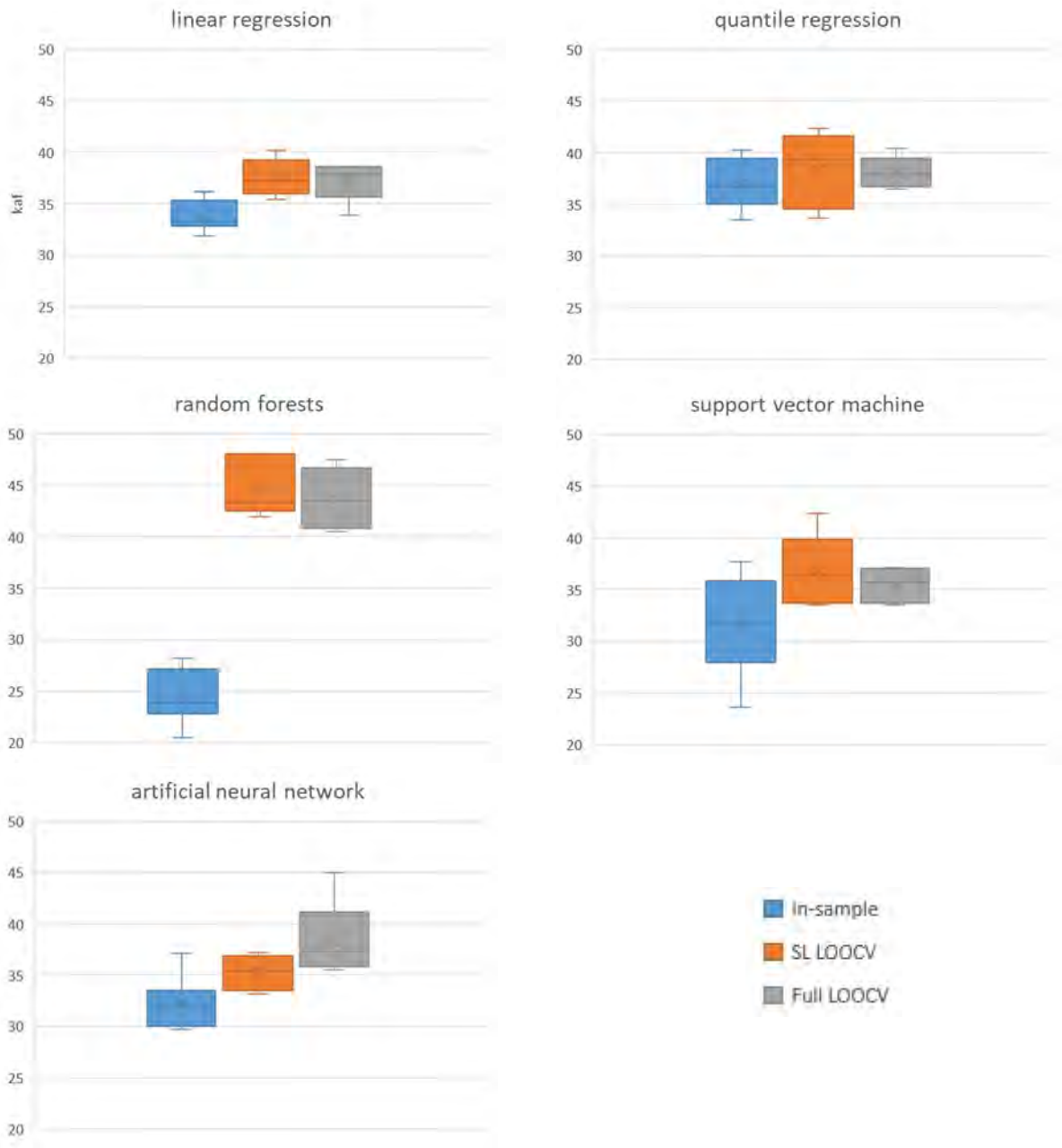


**Figure S3.** As in Figure S1 but for Truckee River April 1 WSF

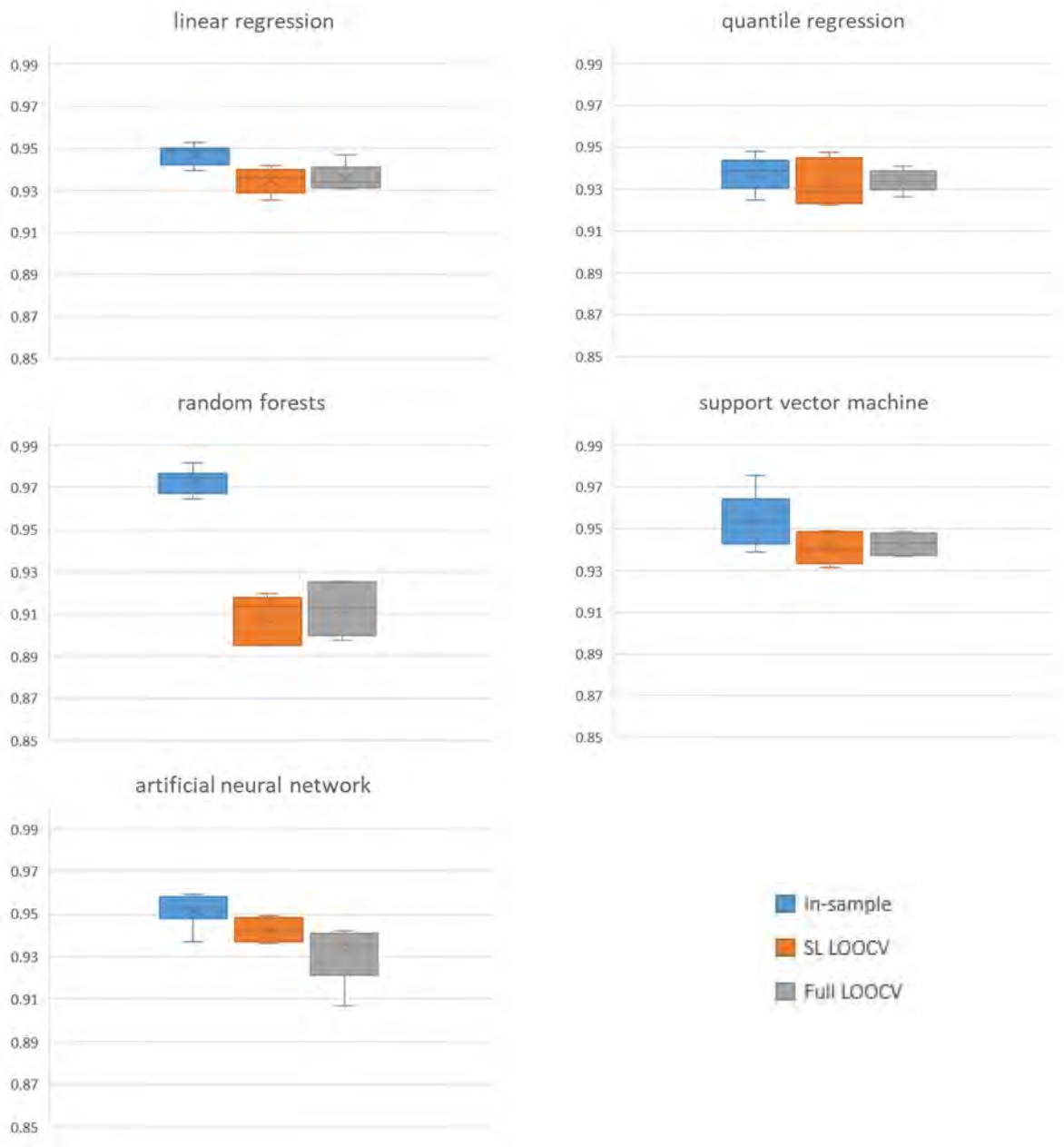


**Figure S4.** As in Figure S3 but for $R^2$

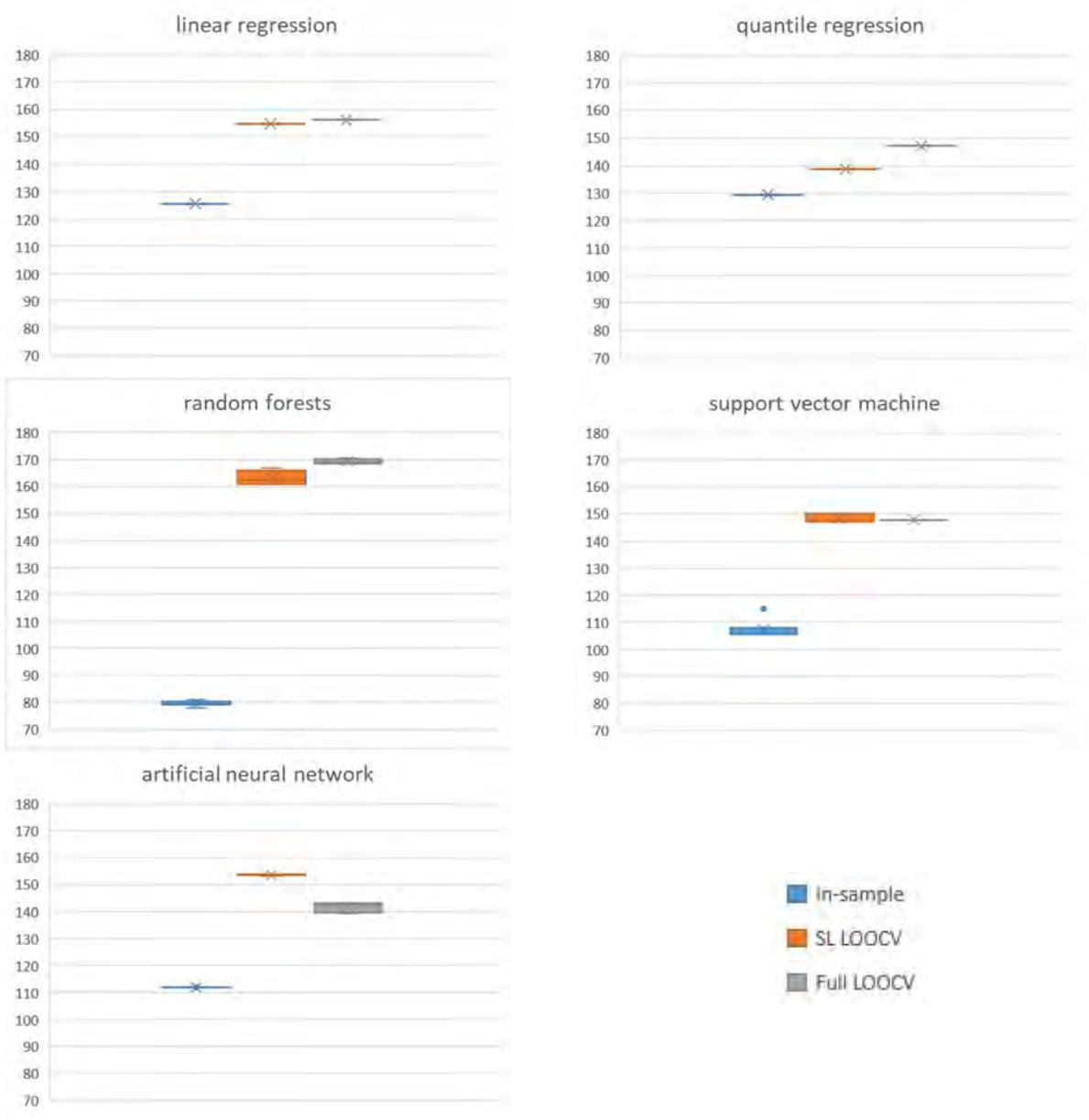


**Figure S5.** As in Figure S1 but using PCA modes 1 through 4 inclusive as the features delivered to the supervised learner, with no automated feature selection. As such, stochasticity inherent to genetic algorithm-based feature optimization is by construction absent in these modeling scenarios, and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


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

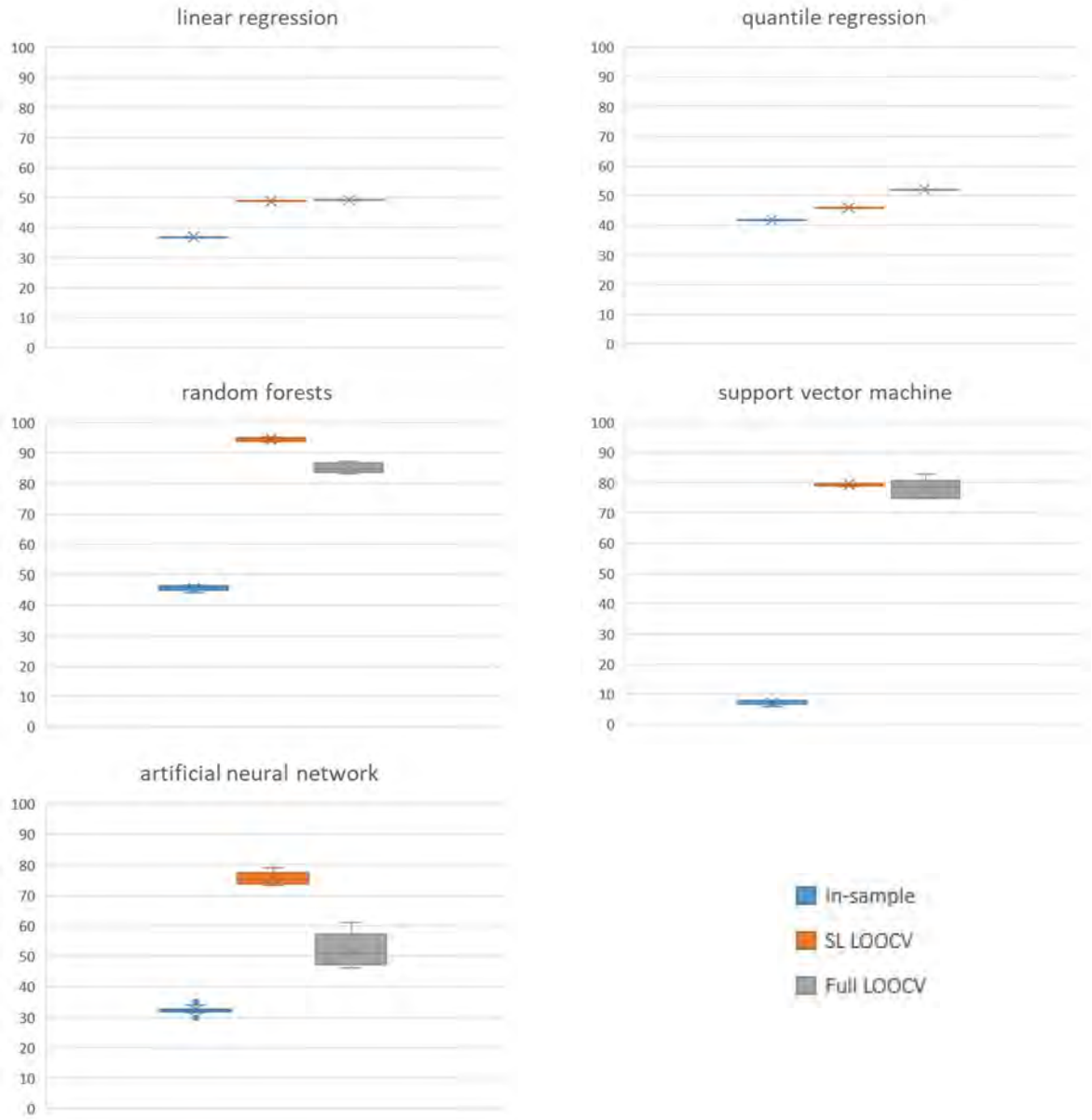


**Figure S7.** As in Figure S5 but for Truckee River April 1 WSF.

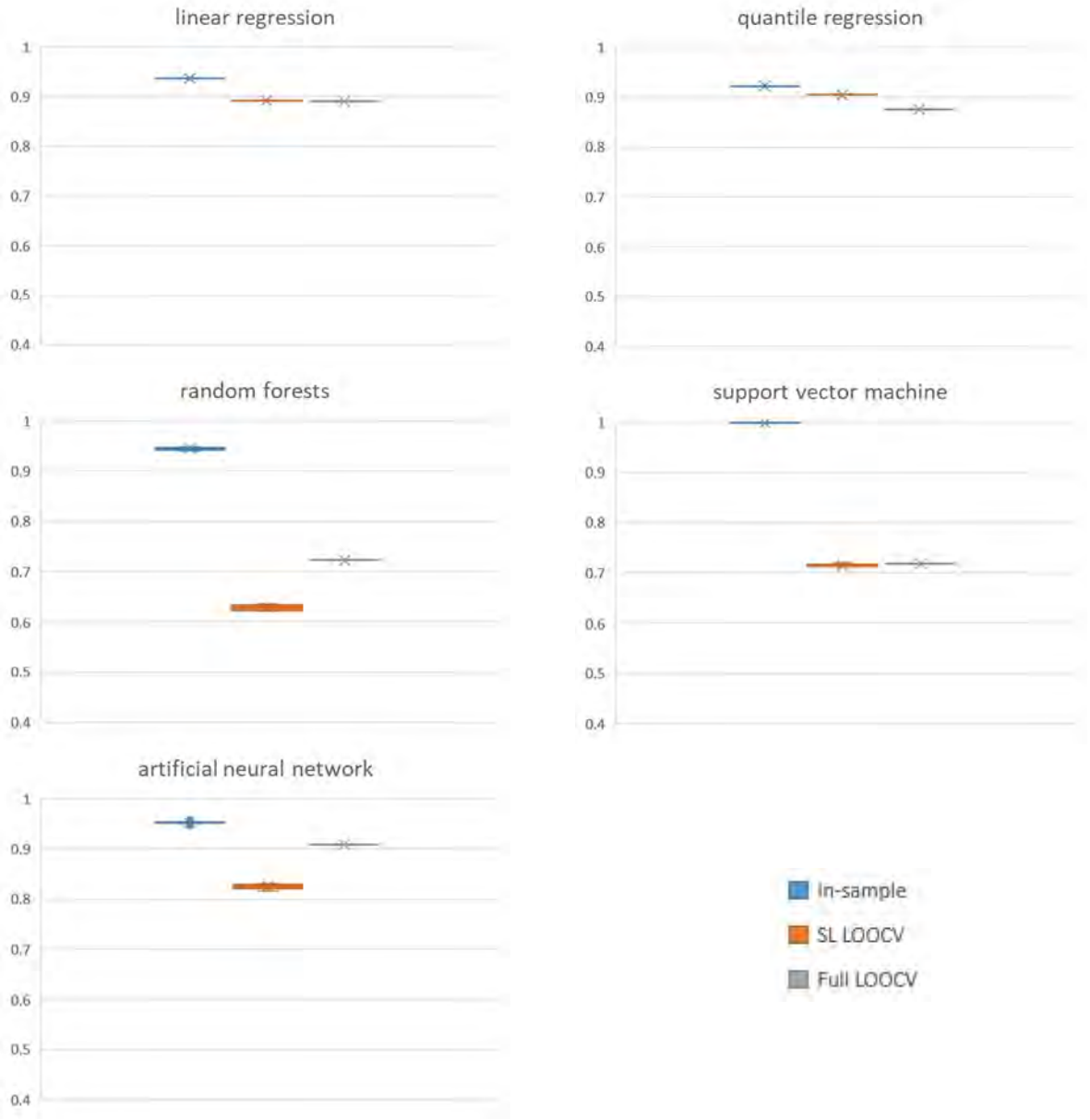


**Figure S8.** As in Figure S7 but for $R^2$.


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

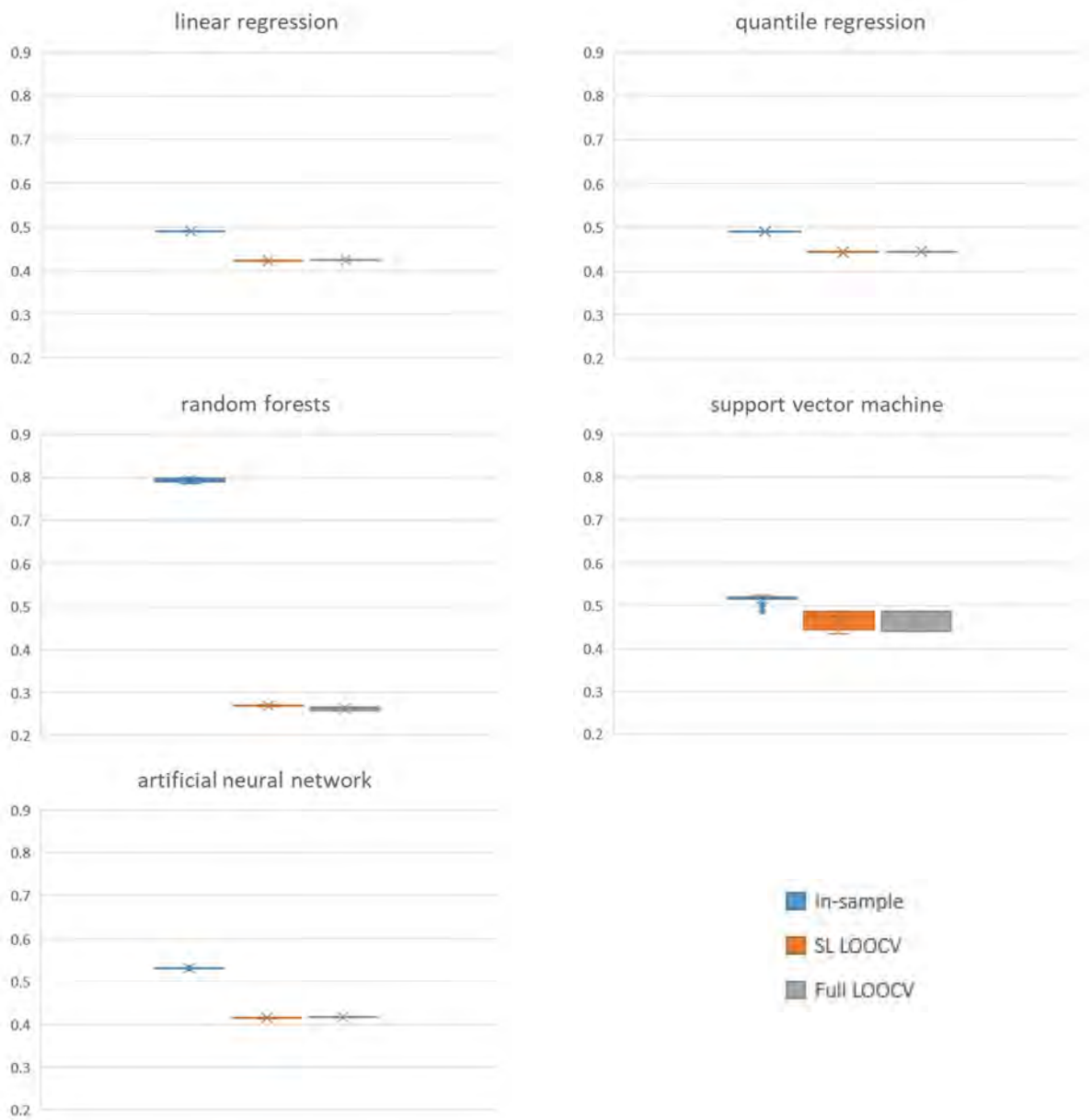


**Figure S10.** As in Figure S9 but for $R^2$.

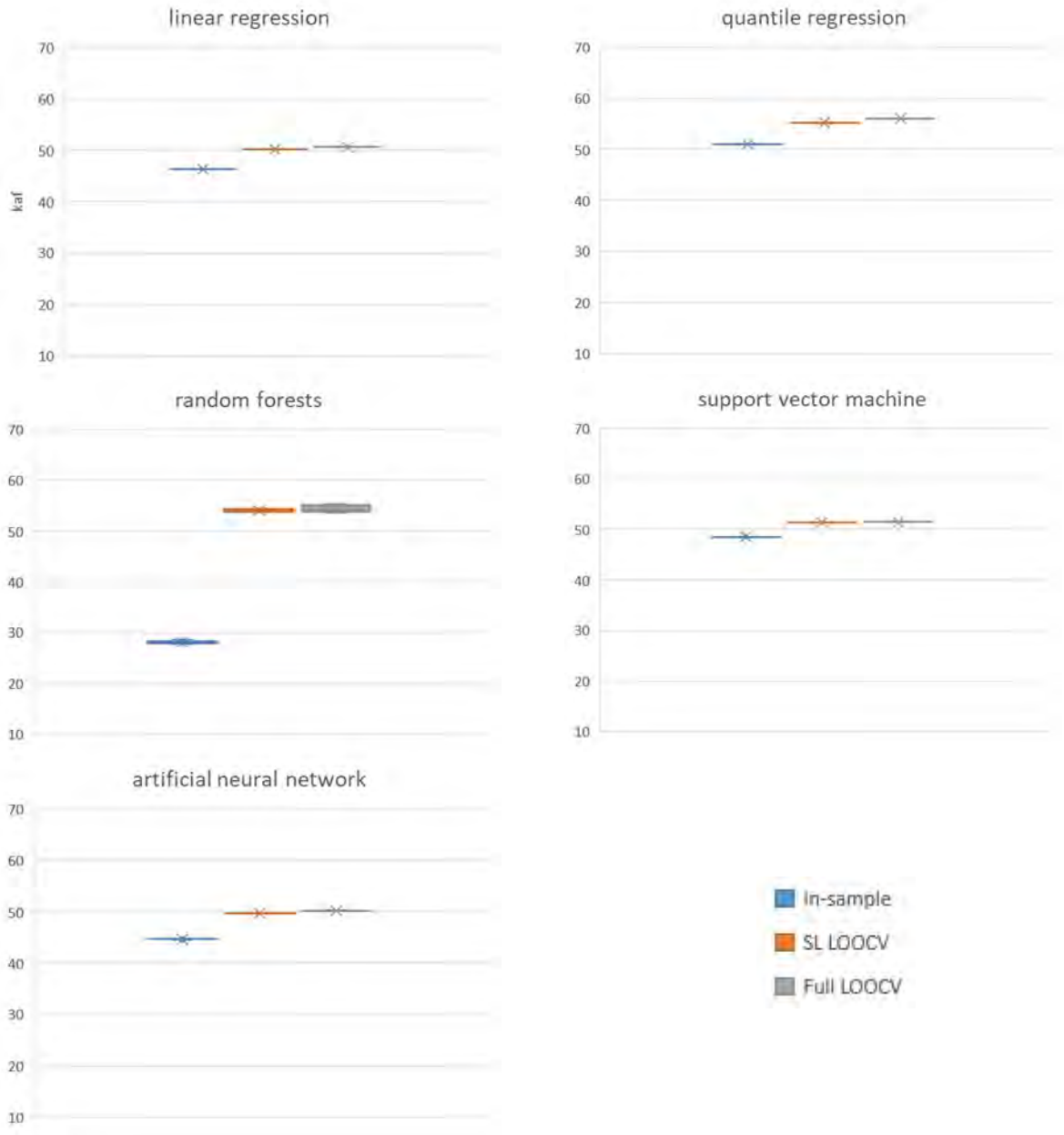


**Figure S11.** As in Figure S9 but for Truckee River April 1 WSF.


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