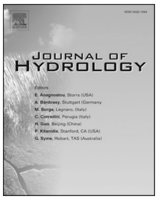
Contents lists available at ScienceDirect

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).

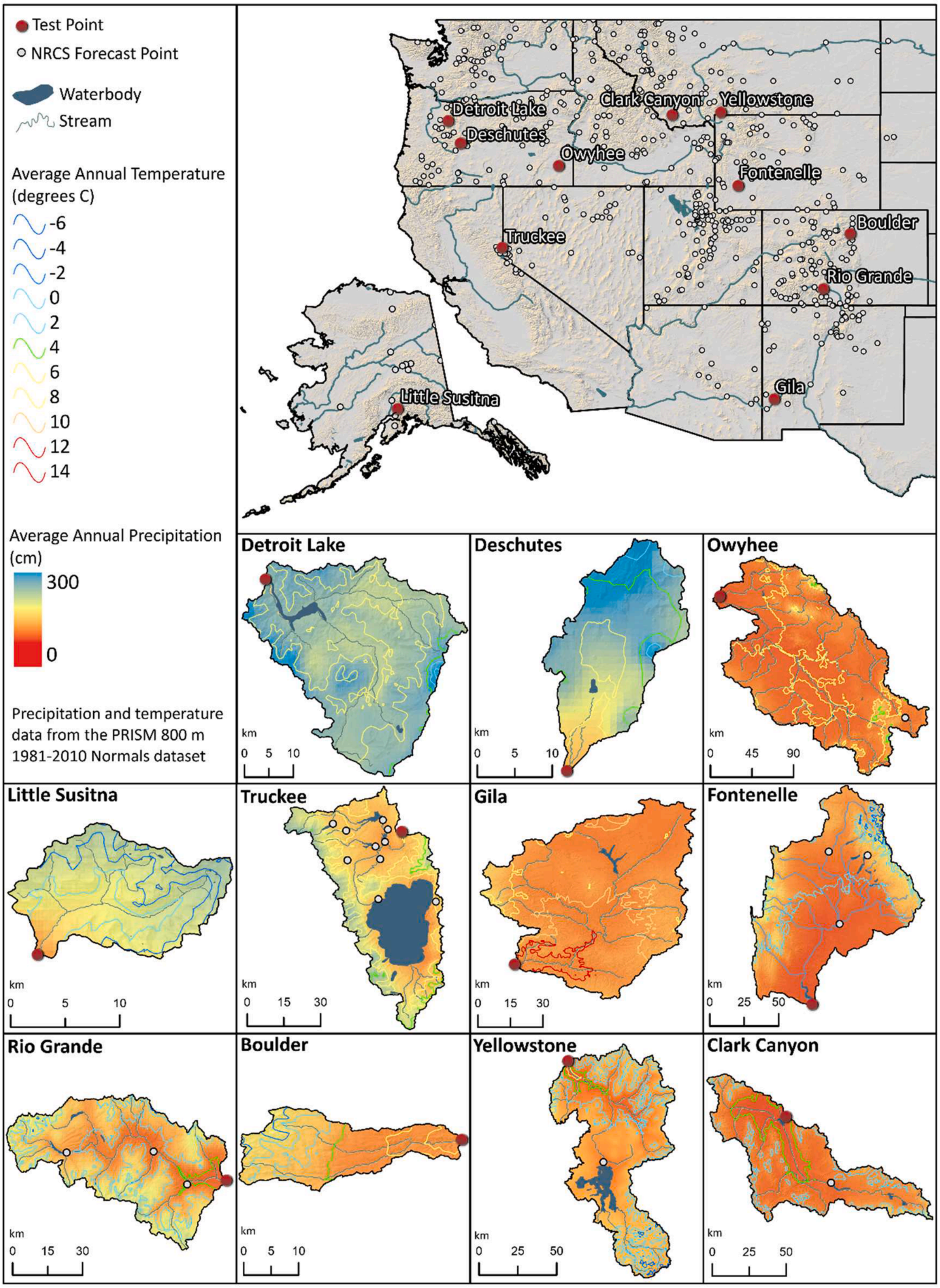


Fig. 1. Locations of all NRCS water supply forecast points, and selected test basins.

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

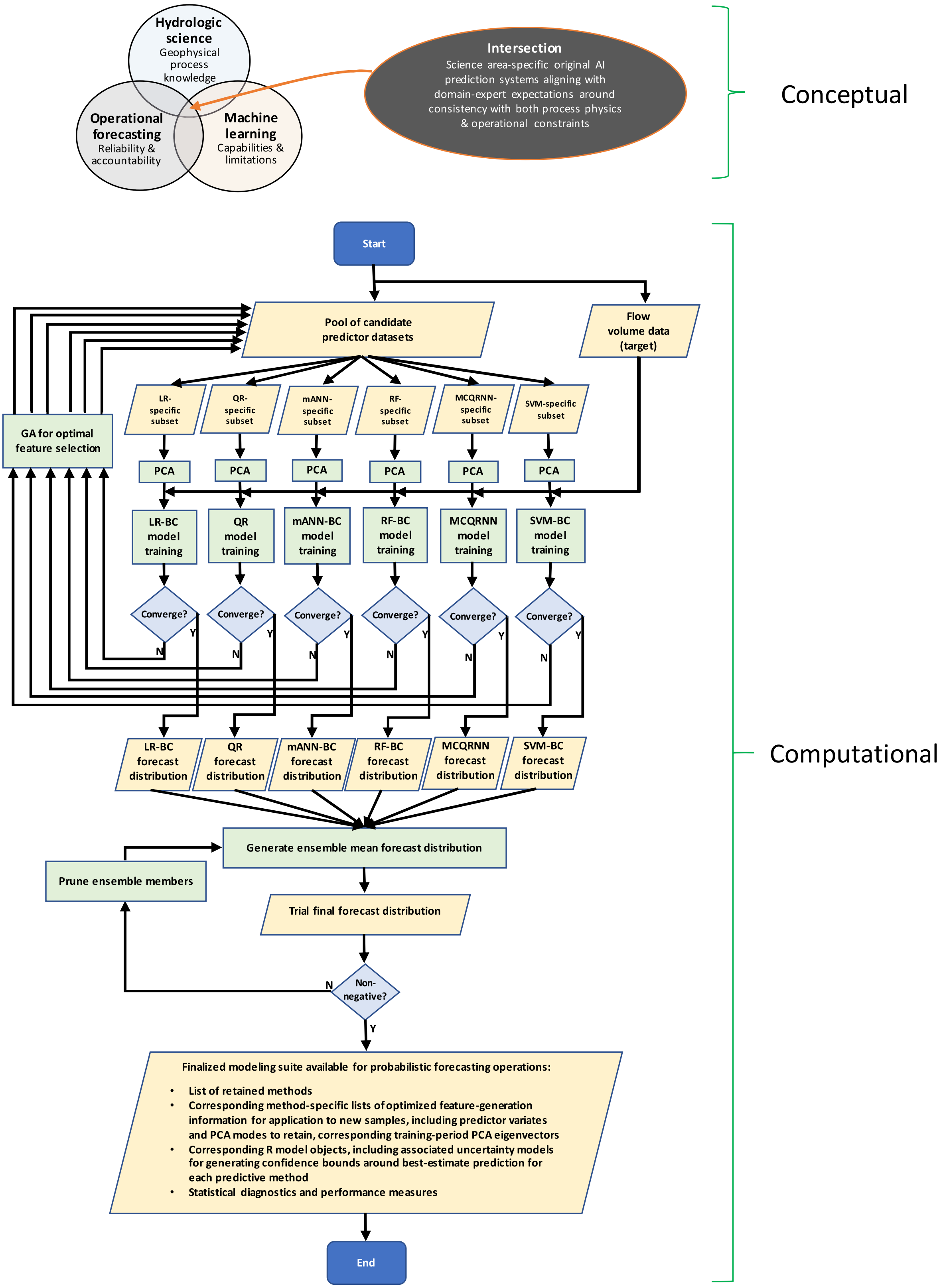

Hydrologic science
Geophysical process knowledge
Operational forecasting
Reliability & accountability
Machine learning
Capabilities & limitations
Intersection
Science area-specific original AI prediction systems aligning with domain-expert expectations around consistency with both process physics & operational constraints
Conceptual
Start
Pool of candidate predictor datasets
Flow volume data (target)
GA for optimal feature selection
LR-specific subset
QR-specific subset
mANN-specific subset
RF-specific subset
MCQRNN-specific subset
SVM-specific subset
PCA
LR-BC model training
QR model training
mANN-BC model training
RF-BC model training
MCQRNN model training
SVM-BC model training
Converge?
Y
N
LR-BC forecast distribution
QR forecast distribution
mANN-BC forecast distribution
RF-BC forecast distribution
MCQRNN forecast distribution
SVM-BC forecast distribution
Generate ensemble mean forecast distribution
Prune ensemble members
Trial final forecast distribution
Non-negative?
N
Y
Finalized modeling suite available for probabilistic forecasting operations:
• List of retained methods
• Corresponding method-specific lists of optimized feature-generation information for application to new samples, including predictor variates and PCA modes to retain, corresponding training-period PCA eigenvectors
• Corresponding R model objects, including associated uncertainty models for generating confidence bounds around best-estimate prediction for each predictive method
• Statistical diagnostics and performance measures
End
Computational

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);

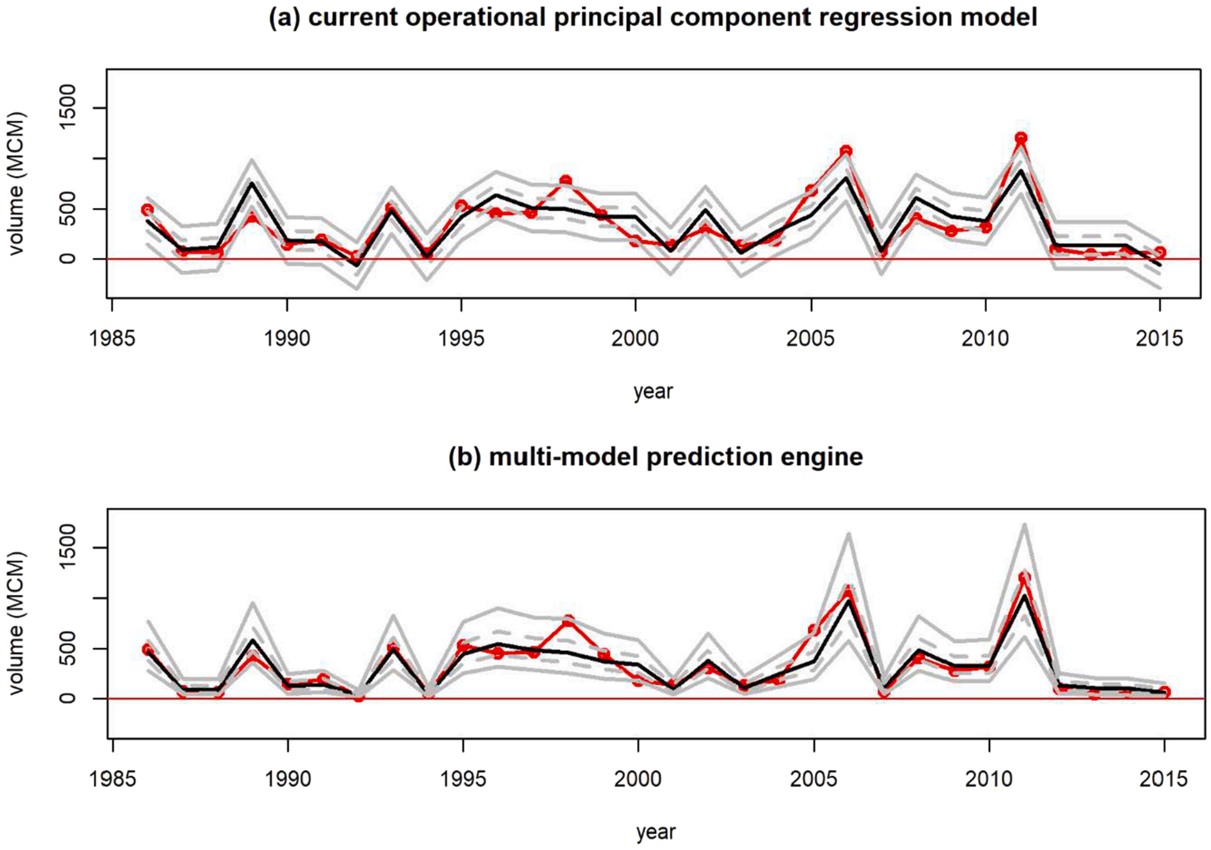


Fig. 3. April 1 Owyhee spring-summer volume forecasts from (a) existing PCR-based WSF system implemented in the VIPER software platform, (b) M^4. Units are millions of cubic meters (MCM). Red dots connected by thick red line: observations; thick black line: best-estimate forecast; solid gray lines: 0.10 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

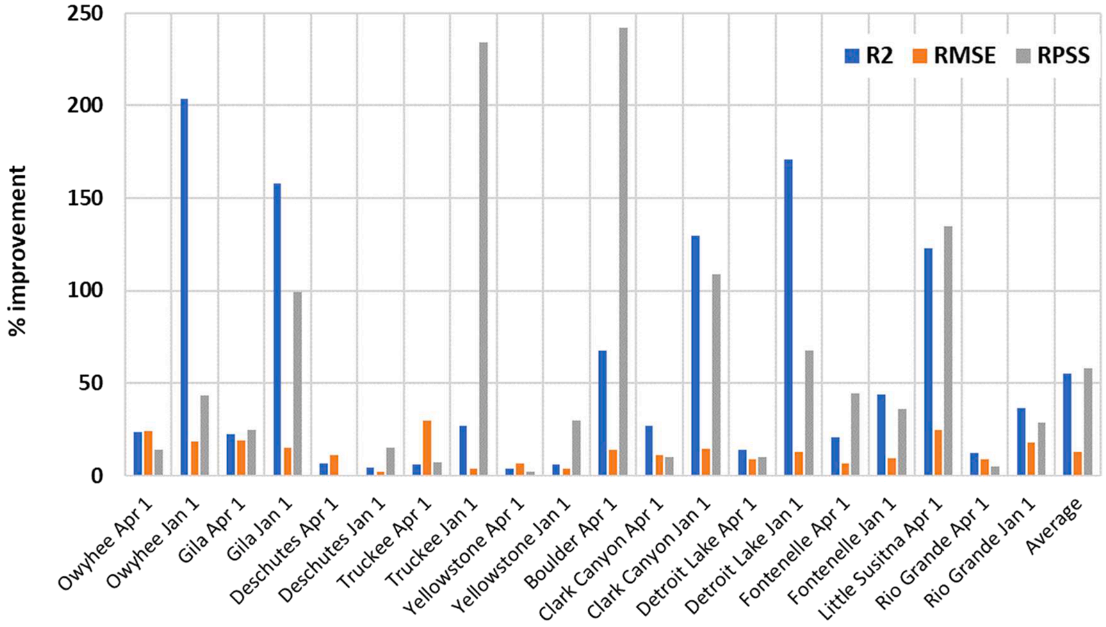


Fig. 4. Percentage improvements provided by M^4 ensemble mean forecast on three common performance metrics for 20 test cases, relative to existing NRCS operational WSF system. Average improvement across all hindcast test cases is also shown. Other key performance aspects were additionally considered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

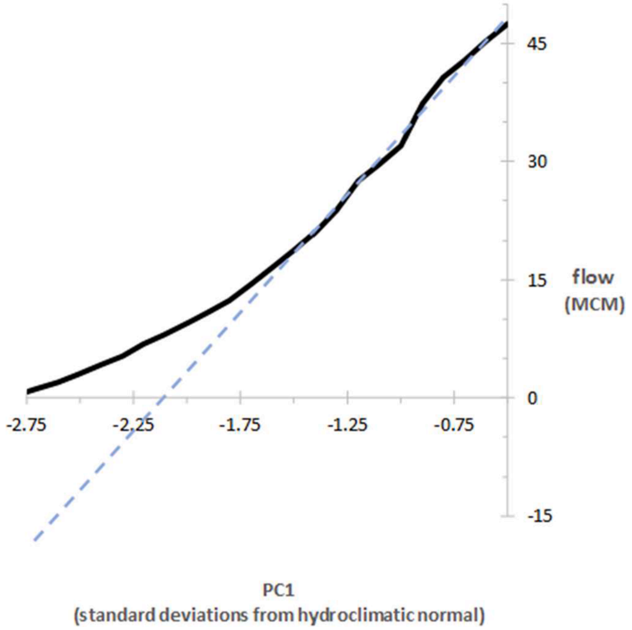


Fig. 5. Nonlinear ensemble mean relationship (thick black line) between AI-based Gila River volume prediction and leading-mode PCA scores time series (PC1), which indexes watershed-scale winter climatic inputs, illustrating partial flattening during dry years. For reference, blue dashed line continues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

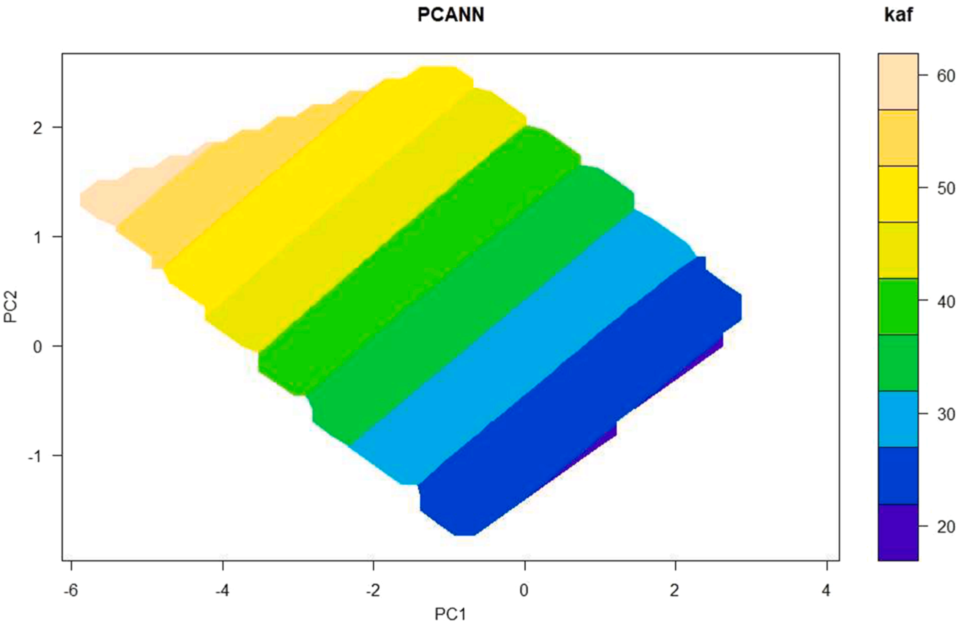


Fig. 6. Approximately linear (near-planar) response surface for the monotone artificial neural network (mANN) relating the January 1 prediction of April-July Deschutes River flow volume to the leading and second PCA modes. Both modes were retained by genetic algorithm feature-optimization for this particular 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.