Interrogating process deficiencies in large-scale hydrologic models with interpretable machine learning

Admin Husic[1], John Hammond[2], Adam N. Price[3], Joshua K. Roundy[4]

[1]Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA
5   [2]U.S. Geological Survey, Maryland-Delaware-D.C. Water Science Center, Catonsville, Maryland, USA
[3]USDA Forest Service, Pacific Northwest Research Station, La Grande, Oregon, USA
[4]Department of Civil, Environmental and Architectural Engineering, University of Kansas, Lawrence, Kansas, USA

Correspondence to: Admin Husic (husic@vt.edu)

10   **Abstract.** Large-scale hydrologic models are increasingly being developed for operational use in the forecasting and planning of water resources. However, the predictive strength of such models depends on how well they resolve various functions of catchment hydrology, which are influenced by gradients in climate, topography, soils, and land use. Most ~~assessments~~assessment of ~~these~~ hydrologic ~~models~~model uncertainty has been limited to traditional statistical ~~approaches. The rise~~methods. Here, we present a proof-of-~~concept approach that uses interpretable~~ machine learning techniques ~~can~~to
15   provide ~~novel insights into identifying~~post-hoc assessment of model sensitivity and process ~~deficiencies~~deficiency in ~~large-scale~~ hydrologic models. ~~In this study, we~~We train a random forest model to predict the Kling-Gupta Efficiency (KGE) of National Water Model (NWM) and National Hydrologic Model (NHM) streamflow predictions for 4,383 streamgages ~~across~~in the conterminous United States. Thereafter, we explain the local and global controls that 48 catchment attributes exert on KGE prediction using interpretable Shapley values. Overall, we find that soil water content is the most impactful
20   feature controlling successful model performance, suggesting that soil water storage is difficult for hydrologic models to resolve, particularly for arid locations. We identify non-linear thresholds beyond which predictive performance decreases for NWM and NHM. For example, soil water content less than 210 mm, precipitation less than 900 mm/yr, road density greater than 5 km/km^2, and lake area percent greater than 10% contributed to lower KGE values. These results suggest that improvements in how these influential processes are represented could result in the largest increases in NWM and NHM
25   predictive performance ~~of NWM and NHM~~. This study demonstrates the utility of interrogating process-based models using data-driven techniques, which has broad applicability and potential for improving the next generation of large-scale hydrologic models.

1 Introduction

Large-scale hydrologic models are important tools for understanding and forecasting the fluxes of water across the
30   earth's surface to manage floods, droughts, and other hydrologic extremes (Brunner et al., 2021; Tijerina et al., 2021). Most

often, these models convert meteorological inputs to streamflow predictions by parameterizing and calibrating internal hydrological processes. Accurate simulation of internal processes is a grand challenge of hydrology (Blöschl et al., 2019) because of the difficulty of resolving equifinality (Vrugt and Beven, 2018), scaling relationships (Savenije, 2018), epistemic uncertainties in hydrologic data (Beven, 2024), and spatial heterogeneity in watershed attributes (McDonnell et al., 2021).

35 Accurate determination of model limitations(Santos et al., 2025). The accurate determination of sensitive model parameters and drivers is crucial for improving process representation in hydrologic models and, ultimately, the management of water resources. (Pandit et al., 2025).

The National Water Model (NWM) and the National Hydrologic Model (NWM) are two process-oriented, continental-scale hydrologic models designed for use in operational decision-making (Towler et al., 2023). The NWM

40 framework applies the WRF-Hydro formulation, which includes representations for infiltration, evaporation, transpiration, overland flow, shallow subsurface flow, baseflow, channel routing, and passive reservoir routing, but not active reservoir management (Cosgrove et al., 2024). The NHM framework applies the Precipitation-Runoff Modeling System formulation, which includes representation of evaporation, transpiration, runoff, infiltration, interflow, groundwater flow, and channel routing, but not reservoir operations, water withdrawals, or stream releases (Regan et al., 2019). Perhaps the major difference

45 in the two modeling approaches is that the NWM has a focus on high-resolution (hourly) flood forecasting whereas the NHM is designed to assess general water availability at timescales from days to centuries (Towler et al., 2023). The NWM and NHM have variable success for streamflow prediction (Tijerina et al., 2021), which depends on differences between sites inwith the strength of prediction varying as a function of catchment-scale climate, land use, and physiographic regimes. physiography. As the NWM and NHM have different constructions, their sensitivity to catchment drivers is likely to differ.

50 The sensitivity of process based hydrologic models to certain catchment attributes and parameters has been interrogated using well established tools, such as sensitivity analyses (Song et al., 2015). These approaches work by exploring the range of values that model parameters may take and recording the net impact on model performance The sensitivity of process-based hydrologic models to certain catchment attributes and parameters has been interrogated using well-established statistical tools, such as sensitivity analysis (Pianosi et al., 2016; Song et al., 2015). These approaches work

55 by exploring the range of values that model parameters may take and recording the net impact on model performance (Mai, 2023). Mai et al. (2022) showed in a recent large-scale sensitivity analysis across North America that functional relationships could be derived between hydrologic processes and physiographic catchment characteristics. In those studies, the authors excluded poor performing sites from their sensitivity analyses (NSE < 0.50); however, poor performing sites may have the greatest potential for identifying sensitive processes and improving their representation in hydrologic models. Thus, there is

60 a need to understand the characteristics of catchments that lead to poor performance to improve confidence in operational decision making in diverse settings.

Machine learning has transformed the field of hydrology in recent years, providing improved predictive capabilities (Kratzert et al., 2018). These data driven approaches have highlighted that large scale hydrological datasets contain more information in them than is explained by our existing theories and perceptions (Nearing et al., 2021). To this end,

65 ~~explainable or interpretable artificial intelligence (AI) methods can be leveraged to bridge the gap between data driven~~ ~~understanding (provided by machine learning models) and process~~. Notable examples include the Sobol' (2001) and Morris (1991) methods. A drawback of traditional sensitivity analysis methods, particularly when applied to large-scale hydrologic applications (Mai et al., 2022), is that they can be computationally demanding (Sarrazin et al., 2016). Less demanding techniques, such as the Robustness Assessment Test (RAT; Nicolle et al., 2021), have been developed to evaluate model bias

70 without the need to control the calibration process but these focus only on the influence of temporal forcings, such as air temperature. Thus, there is a need to continue to develop spatial methods for assessing model sensitivity that are useful in scenarios where traditional methods are computationally intractable.

Explainable or interpretable machine learning methods have the potential to bridge the gap between data-driven insights (provided by machine learning models) and process-based understanding (contained within physically based

75 models) ~~(Park et al., 2022). Numerous explainable AI~~(Slater et al., 2025). These methods help to explain why a model gives the prediction that it does (Lundberg et al., 2020). Several explainable machine learning methods have been developed, including Partial Dependence Plots (PDP; Friedman, 2001), Local Interpretable Model-Agnostic Explainers (LIME; Ribeiro et al., 2016), and Shapley Additive Explanations (SHAP; Lundberg et al., 2020). ~~Thus, there is an opportunity to apply data-~~ ~~driven, explainable AI approaches to identify sensitive processes in physically based hydrologic models.~~

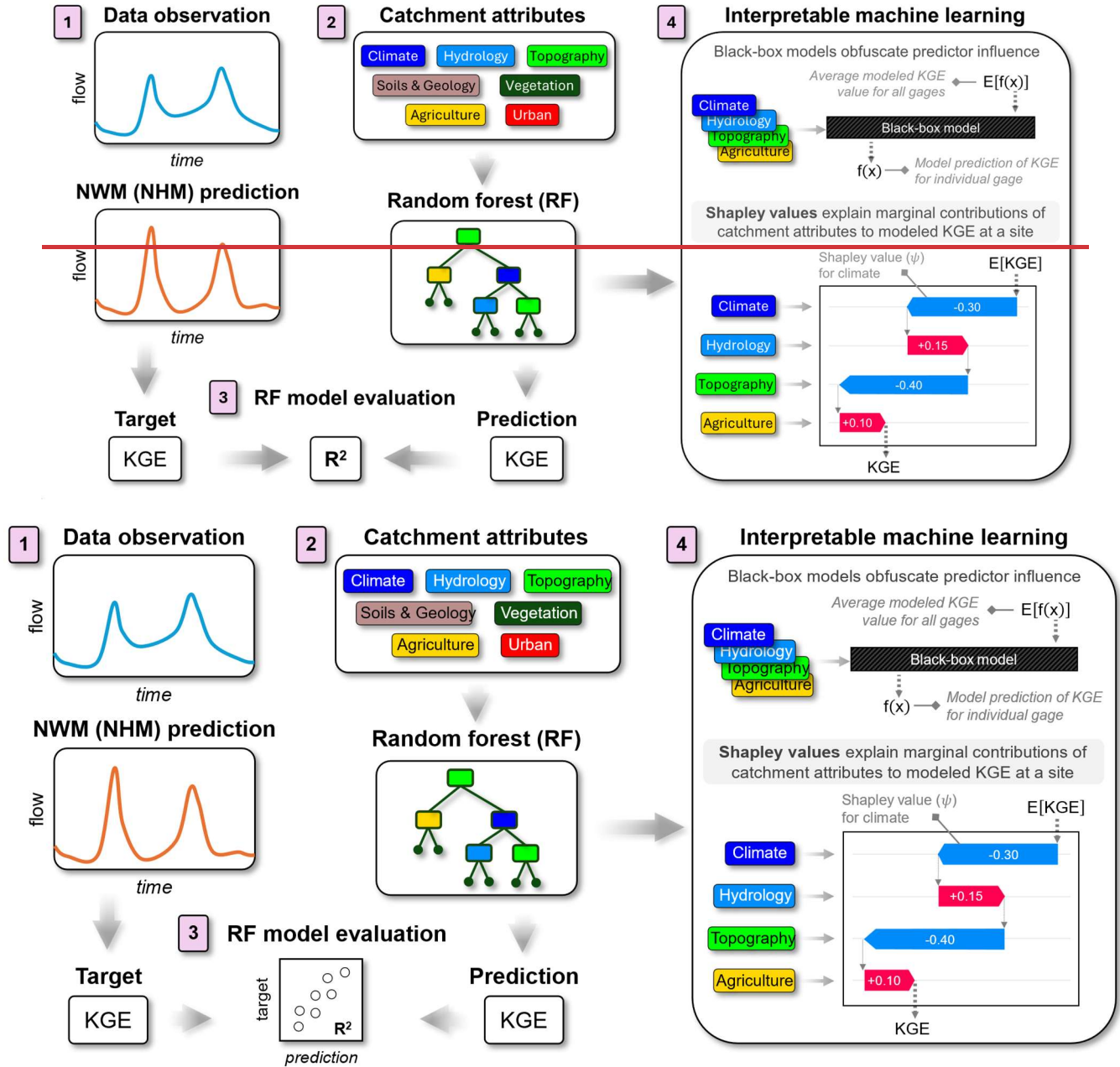
80 ~~Explainable AI can complement and enhance traditional approaches like Sobol's sensitivity analysis (Nossent et al.,~~ ~~2011) and the Generalized Likelihood Uncertainty Estimation (GLUE; Blasone et al., 2008) method in hydrologic modeling~~ ~~by providing deeper insights and interpretability. Explainable AI techniques, such as SHAP values, enhance traditional~~ ~~sensitivity analyses like Sobol by providing interpretative insights into how parameter changes influence model predictions.~~ ~~Similarly, explainable AI complements the GLUE methodology by visualizing the impact of uncertain parameters on output~~

85 ~~variability and offering local explanations for specific predictions. However, caution is necessary when inferring AI results~~ ~~because they typically only imply direct or indirect relations and may not represent causal linkages (Heskes et al., 2020).~~In hydrology, for example, these tools have been applied for the analysis of hydrologic fluxes (Brêda et al., 2024), soil moisture (Huang et al., 2023), water table depth (Ma et al., 2024), and drought intensity (De Meester and Willems, 2024). Interpretable machine learning can complement and enhance traditional sensitivity approaches (Maier et al., 2024), by

90 providing post-hoc interpretative insights into how parameter changes influence hydrologic model predictions, that is, without the need for perturbing the model parameter space. Interpretable machine learning methods are not without limitations as they only imply relations in the model which may not necessarily be causal (Heskes et al., 2020), thus caution should be exercised when interpreting results.

This paper aims to interrogate large-scale hydrologic model performance with machine learning tools to identify

95 which processes may be inadequately represented in physically based models. Thus, the questions we address are: what catchment attributes can be used to predict poor model performance, and are certain dominant hydrological processes associated with these catchment attributes? To answer these questions, we ~~built a random forest~~present a proof-of-concept approach that uses machine learning techniques to provide post-hoc assessment of model sensitivity. We did this by building

a random forest model to predict KGE values for NWM and NHM predictions at over 4,000 basins. Thereafter, model
100   predictions were interpreted using Shapley values, which highlight the physiographic and hydrologic controls of process-
based model performance (Fig. 1). This work aims to inform how the next generation of large-scale hydrologic models can
be improved for the responsible stewardship of water resources into an uncertain future.

105 **Figure 1:** Flow diagram showing the application of interpretable machine learning in this study. (1) Data observations and National Water Model (NWM) or National Hydrologic Model (NHM) predictions are used to generate a target Kling–Gupta efficiency (KGE) for each site. (2) Catchment attributes are input to a Random Forest (RF) model to predict KGE for each site. (3) The RF model is evaluated by comparing the predicted KGE to the target KGE~~.~~, using the coefficient of determination (R^2) to determine goodness of fit. (4) Shapley values (ψ) are used to explain the marginal contributions of catchment attributes that distinguish KGE prediction at a particular site, f(x),

110 from the average modeled KGE for all sites, E[(f(x)]. In the given example, the values of the climate and topography attributes at this individual gage lower the predicted KGE ($-\psi$), whereas the values of the hydrology and agriculture attributes increase the predicted KGE ($+\psi$~~) by the RF. R^2 = coefficient of determination.~~).

2 Methods

2.1 The National Water and National Hydrologic Models

115   We retrieve daily streamflow observations and predictions for gaged locations (sites) for the NWM version 2.1 and NHM version 1.0 from existing repositories ~~(Johnson et al., 2023a; Regan et al., 2019). A total of 4,614 basins~~(Johnson et al., 2023a; Regan et al., 2019; Towler et al., 2023). A total of 4,614 basins with at least 10 years of data that span the contiguous US (CONUS) are included in our analysis (U.S. Geological Survey, 2024). The date range of flow observations and predictions is from water years 1984 to 2016. ~~The accuracies of~~

120   NWM ~~(Fig. 2)~~ and NHM ~~(Fig. S1) predictions are particularly sensitive to aridity. Model~~ performance at each site was assessed ~~at each site~~ using the Kling-Gupta Efficiency (KGE), a common ~~evaluation~~ metric ~~for~~in hydrologic modeling (Gupta et al., 2009). The KGE is calculated as

$$KGE = 1 - \sqrt{(\alpha - 1)^2 + (\rho - 1)^2 + (\beta - 1)^2} \tag{1}$$

      where ρ is Pearson correlation coefficient, and α and β are the ratios of the standard deviation and the mean, respectively, of model predictions to data observations.

125   The accuracies of NWM (Fig. 2) and NHM (Fig. S1) predictions are particularly sensitive to aridity. The KGE values calculated in Equation (1) serve as the target variables for the forthcoming machine learning model (Figure 1).

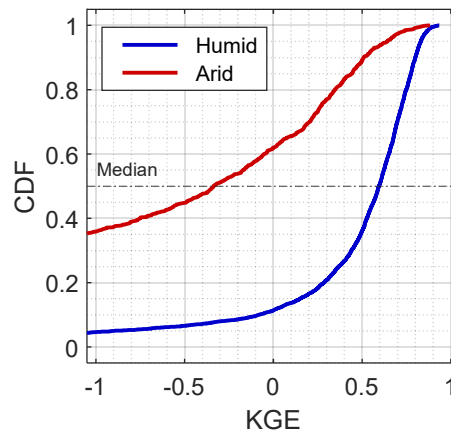



Figure 2: Cumulative distribution function (CDF) of National Water Model (NWM) performance for humid (PET/P <1, n = 3,827) and
130 arid (PET/P >1, n = 787) sites as assessed by the Kling–Gupta efficiency (KGE) evaluation metric.

2.2 Random Forest Model

      ~~A regression tree is a supervised learning approach that can predict continuous values and capture non-linear trends in a dataset (De'Ath and Fabricius, 2000). A random forest model creates an ensemble of regression trees to mitigate the potential of overfitting to a single regression tree~~Random forest modeling is an ensemble-based machine learning approach
135 for predicting continuous values and capturing non-linear trends in a dataset (Ho, 1998). ~~In this study, we~~We train a random

forest model, comprising of 1,000 regression trees, to predict the target KGE. ~~The predictor~~ at each site using catchment attributes as input variables (termed "features~~") used to train the model are 48 catchment attributes, which were aggregated based on their likelihood to impact hydrology.~~"). The features (n = 48) are derived from BasinATLAS (Linke et al., 2019) and incorporate wide ranges of climate, hydrology, topography, soils & geology, ~~natural~~undeveloped vegetation, agriculture, and urban land use. The names and descriptions of the 48 predictors can be found in Table S1, and the spatial variations of the 48 predictors across the CONUS are shown in Fig. S2. ~~One influential predictor – soil water content – is defined as soil stress, or~~The features were selected based on their likelihood to impact hydrology. Soil water content appears as an important predictor in the later analysis, and we define it here for clarity. Soil water content is defined as the annual soil water available for evapotranspiration (Trabucco and Zomer, 2010)~~. We represent soil water content as an equation~~, and the original study authors calculated it as equal to the long-term effective precipitation minus the sum of actual evapotranspiration and runoff.

~~Training of the~~The random forest model was ~~done with "in-the-bag"~~trained and ~~"out-of-bag" splits.~~validated using bootstrapping. Individual trees are grown from an "in-the-bag" bootstrap of the observation dataset. "Out-of-bag" observations not included in the bootstrap are used for model validation. The models were trained using the mean squared error objective function. The coefficient of determination (R^2) was calculated to assess predictive performance of the random forest (Pearson, 1901). Extreme values (outliers) can distort the utility of a predictive and interpretable model (Liu et al., 2018). Because the KGE metric has a small upper bound (+1) and an infinite lower bound (-∞), a small subset of very negative values can dominate model inferences. The lowest KGE value for a gaged location in the NWM dataset is -302.8, whereas the 5th percentile of KGE values -2.7. The performance at both sites would be considered "unacceptable"; thus, including extreme negative values negatively affects model predictability without providing much additional insight beyond that given by other underperforming sites. To address the disproportionate influence of a small subset of values, we consider the 5% of sites with the most negative target KGE values as outliers, reducing our dataset from 4,614 to 4,383 sites. Random forest model analyses and development were performed using the ~~treebagger~~*TreeBagger* function in MATLAB 2023 (MathWorks, ~~2023~~2024).

2.3 Shapley Values

~~The Shapley value approach is a model-agnostic, explainable AI method~~Shapley values are derived from cooperative game theory and they aim to assess how coalitions form and how these coalitions impact the payout of a game (Shapley, 1953). In the context of explainable AI, they are a model-agnostic approach that attributes each feature an importance value for a prediction, indicating the marginal benefit that the inclusion of ~~a particular~~the feature provides to the overall prediction (Lundberg et al., 2020; Lundberg and Lee, 2017). ~~The~~Thus, Shapley ~~value has~~values explain the ~~same units as those~~inner workings of a model, with influential features receiving large attribution of credit whereas non-influential features may receive little or no credit for the model prediction. ~~Further, the~~The Shapley value is also the only distribution of gain among features (e.g., predictor variables) that maximizes four properties: (1) efficiency, (2) symmetry, (3) linearity, and

(4) null player (Shapley, 1953). ~~Thus, while other model explanation techniques exist, e.g., LIME (Ribeiro et al., 2016), they violate one or more of these properties.~~Respectively, these properties ensure that (1) the total prediction is fully allocated to features, (2) features that contribute the same to the prediction should receive identical credit, (3) allocations in ensemble models should be the sum of the allocation in member models, and (4) a feature contributing nothing to the prediction should receive no allocation.

The Shapley value (ψ) of the i-th feature (catchment attribute) for the query point x (KGE) can be calculated by the characteristic value function (v) as:

$$\psi_i(v_x) = \frac{1}{M} \sum_{S \subseteq M \setminus \{i\}} \frac{|S|! \, (M - |S| - 1)!}{(M - 1)!} \, [v_x(S \cup \{i\}) - v_x(S)] \tag{2}$$

where M is the number of features, M is the set of all features, S is a set or coalition of features, |S| is the number of elements in the coalition, $v_x(S)$ is the value function of the features in the coalition for the query point x (Shapley, 1953). The value of $v_x(S)$ represents the "worth" or the expected contribution of the features in S to the cooperative prediction for the query point x. Leveraging the additive nature of Shapley values, we calculate them for each observation for all trees in the random forest and then average respective feature results across trees for a more robust statistic. All Shapley value analyses were performed using the *TreeSHAP* function in MATLAB 2023 ~~(Lundberg et al., 2020; MathWorks, 2023).~~(Lundberg et al., 2020; MathWorks, 2024). The Shapley value has the same units as those of the prediction, in this case unitless (as is KGE).

~~————~~To aid in interpretation of Shapley values, we provide a brief example. The random forest model described in Section 2.2 is trained to predict the KGE of the NWM (or NHM) model at 4,383 sites in the analysis (Fig. 1). In short, "how accurate is the NWM model at site?" The random forest model answers this question by transforming 48 catchment attribute features into a prediction of KGE. In the absence of Shapley values or interpretable AI, the process by which the catchment attributes are transformed to create the KGE prediction is uncertain, that is, it is a black box. Shapley values elucidate the marginal contribution of a feature (ψ) to the random forest, which is defined as how much the predicted KGE at a site increases ($+\psi$) or decreases ($-\psi$) when a feature is included in the model. In this way, sensitive features will have a high Shapley value magnitude, $|\psi|$, as the predicted KGE is sensitive to the value that the feature takes. Thus, Shapley values help to distinguish the catchment attributes that cause variation in predicted KGE across space. Although the full range of Shapley values for the 48 catchment attribute features are informative, we ~~focus on~~highlight and discuss the most impactful feature negatively affecting model ~~performance~~performances at each site. The most impactful feature is ~~defined as~~ the one having the lowest Shapley value (min ψ) at a site, ~~i.e., providing the most negative marginal contribution to KGE prediction.~~ meaning it reduces the predicted KGE more than any other feature.

We ~~relate our Shapley value feature importance to the spatial distributions of the~~ used the Ecological Regions of North America as a way of grouping clusters of catchments in order to facilitate the discussion of similarities (or dissimilarities) between the drivers of model performance across broad areas (Omernik, 1987)~~, where watersheds were assigned to Ecoregions based on the greatest area of an ecoregion in the watershed.~~. Ecoregions are defined by "perceived

patterns of a combination of causal and integrative factors including land use, land-surface form, potential natural vegetation, and soils. ~~We consider Level-I and Level-II ecoregions, which will be identified in figures using the following superscripts: [1]Atlantic Highlands, [2]Mixed Wood Shield, [3]Ozark, Ouachita-Appalachian Forests, [4]Mixed Wood Plains, [5]Central USA Plains, [6]Southeastern USA Plains, [7]Mississippi Alluvial and Southeast USA Coastal Plains, [8]Everglades, [9]Temperate Prairies, [10]West-Central Semi-Arid Prairies, [11]South Central Semi-Arid Prairies, [12]Texas-Louisiana Coastal Plain, [13]Tamaulipas-Texas Semi-Arid Plain, [14]Cold Deserts, [15]Warm Deserts, [16]Western Sierra Madre Piedmont, [17]Upper Gila Mountains, [18]Western Cordillera, [19]Marine West Coast Forest, and [20]Mediterranean California.~~" (Omernik, 1987). Results from individual catchments were aggregated to the ecoregion level for comparison of general trends. A catchment was assigned to an ecoregion based on the greatest area of an ecoregion contained within the drainage boundary of a catchment.

3 Results

Because general results for both the NWM and NHM were broadly similar, we focus the main text discussion on the NWM and note instances where the two models differ (detailed results from NHM analysis can be found in the Supplement). R^2 values for the ~~training and~~ testing predictions of KGE for the random forest model ~~were 0.86 (0.86) and 0.~~are shown in Fig. 3. The random forest model explains 47 ~~(0.~~% (43~~), respectively,~~%) of the variance encoded in the KGE metric for ~~the~~ NWM (NHM~~). The criteria for~~) simulations at 4,383 gages, which contain considerable variability in the processes influencing hydrologic model performance. We consider model performance satisfactory as acceptability ~~of~~criteria for R^2 varies with the complexity of a dataset ~~(Legates and McCabe, 1999), and we consider a model that explains 47% of the variance encoded in the KGE metric for 4,383 gages as acceptable to function as a surrogate for predicting NWM performance. We~~(Legates and McCabe, 1999). We proceed and apply explainable AI techniques to this random forest model to understand how catchment attributes influence KGE values of streamflow for the NWM and NHM.

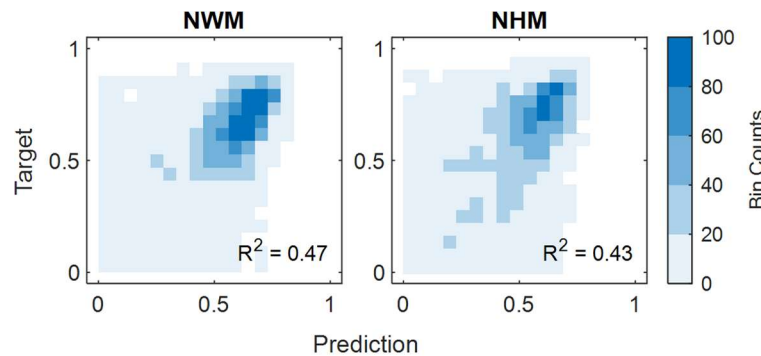



Figure 3: Evaluation of the random forest model prediction of Kling-Gupta Efficiency (KGE) at NWM and NHM sites. Results are shown for the out-of-bag (testing) samples. The density scatter plot displays the count of data points in each partitioned bin. For visual clarity, predicted and observed KGE values less than 0 are not plotted, although they are included in the calculation of R^2 for each model. NWM = National Water Model, NHM = National Hydrologic Model, R^2 = Coefficient of Determination.

We investigated the local structure of Shapley values (ψ) ~~for~~at three demonstration sites (Fig. ~~3~~4). We report how the Shapley values explain random forest model predictions of KGE, but it is important to note that these explanations ~~may~~are not ~~be a result of direct causality~~necessarily directly causal. The directionality and extent of influence by each predictor is indicated by the magnitude and sign of the predictor's Shapley value ($\pm\psi$). Each waterfall plot shows how Shapley values (ψ) of features help to distinguish one site, f(x), from the mean of all sites, E[f(x)]. These three sites were selected to demonstrate various catchment controls, such as climate at Tucannon River, WA; hydrology at Seboeis River, ME; and soils & geology at Timpas Creek, CO. At Tucannon River, the relatively high values of actual evapotranspiration and aridity index at the site cause a decrease (-ψ) in the prediction of KGE at that site. At Seboeis River, the large lake area percentage causes a decrease (-ψ) in KGE prediction, but the high soil water content causes an increase (+ψ) in KGE prediction. At the final site, Timpas Creek, the most influential feature is the low soil water content, which has a considerable negative contribution (-ψ) to KGE prediction. With an understanding of ~~the local structure of Shapely~~how Shapley values operate at a local site-scale, we proceed to a global perspective by assessing the aggregate Shapley value results of all 4,383 sites.

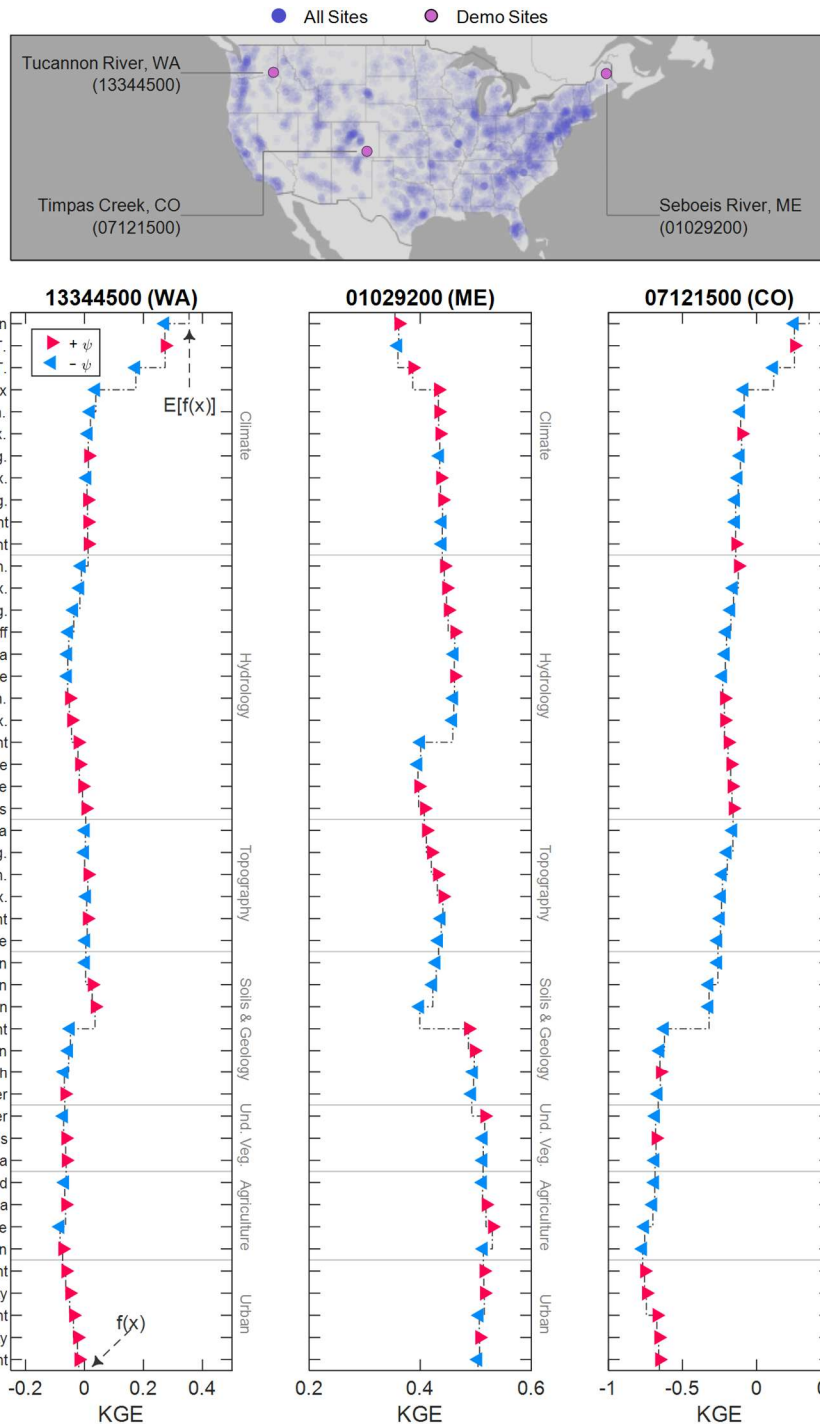



Figure 4: Local structure of Kling–Gupta efficiency (KGE) prediction for the National Water Model (NWM) as illustrated by Shapley value (ψ) waterfall plots at three demonstration sites, indicated by U.S. Geological Survey station numbers associated with streamgages and 2-letter state abbreviations. Each plot begins with the expected value of the model prediction for all sites, $E[f(x)]$, which undergoes marginal alteration ($\pm\psi$) by each of the 48 predictor features. The final model prediction, $f(x)$, is equal to $E[f(x)]$ plus the cumulative sum of all marginal contributions. Undeveloped Vegetation is abbreviated as Und. Veg.

245

The global structure of Shapley values (ψ) for six important catchment attributes is shown (Fig. 4~~5~~): soil water content, snow cover maximum, road density, precipitation, lake area, and irrigated area. The marginal contribution of the soil water content variable ($\psi_{\text{soil water content}}$) is positive (+$\psi$) in areas with high soil water content (east of the 98th meridian and in the Pacific Northwest) and negative (-ψ) in areas with lower soil water content (Great Plains, Intermountain West, and California). The Shapley dependence plot identifies 210 mm soil water content as a threshold from when $\psi_{\text{soil water content}}$ increases (+ψ) versus decreases (-ψ) the prediction of KGE. The $\psi_{\text{snow cover max.}}$ values are positive in the Rocky Mountains and the upper Midwest. Snow cover maximum has little effect on KGE predictions until a threshold of 40% is exceeded, at which point maximum snow coverage improves KGE prediction. The $\psi_{\text{road density}}$ values are negative in urban centers, when road density exceeds 5 km/km^2, suggesting high road density decreases accuracies of model predictions. Otherwise, the presence of roadways has little impact on KGE predictions at lower road densities. A threshold of 900 mm/yr in precipitation emerges; precipitation values lower than this threshold lower KGE (-$\psi_{\text{precipitation}}$) and values greater than this threshold increase KGE (+$\psi_{\text{precipitation}}$). The $\psi_{\text{lake area}}$ values are generally close to zero except for when lakes constitute a substantial portion of a watershed (> 10%), such as in Minnesota and Wisconsin and the Northeast Region. For $\psi_{\text{irrigated area}}$, watersheds with less than 3% irrigated area are unaffected by the variable, but beyond a threshold of around 10%, the presence of irrigation decreases KGE predictions.

Shapley value swarm charts show the directionality and magnitude of feature importance for all 48 predictors (Fig. ~~5~~6). Globally, the most impactful features (greatest $\overline{|\psi|}$) for KGE prediction are $\psi_{\text{soil water content}}$, $\psi_{\text{aridity index}}$, $\psi_{\text{actual ET}}$, and $\psi_{\text{precipitation}}$. Regarding directionality, higher catchment-scale values of soil water content, aridity index, actual ET, and precipitation increase KGE prediction (+ψ) whereas smaller values decrease KGE prediction (-ψ). Although these are globally the most influential variables, they are not necessarily the most influential at each individual site. We plot the spatial distribution of the most impactful feature group leading to poor KGE scores at each site, that is the predictor group having the greatest negative Shapley value (min ψ) at a site. The count of most impactful features groups at individual sites were climate (n = 761), hydrology (n = 1,290), and soils and geology (n = 1,447). Soils and geology features, most frequently low soil water contents, reduced KGE most often in the Great Plains and Intermountain West. Hydrology features, typically large values of lake and reservoir storage, reduce modeled KGE in the Midwest. Climate features did not have strong spatial coherence. Next, we assess the distribution of KGE values grouped by most impactful feature (Fig. ~~6~~7). For the NWM, sites where the most impactful features were soils & geology as well as urban land use had the lowest median KGE values. The results for NHM were similar to NWM except that areas controlled by climate have lower median KGE values for NHM than NWM.

We map the spatial linkage between ecological regions in the US and the influential features controlling KGE scores at sites contained within these regions (Fig. ~~7~~8). The ecoregions containing the most streamgages are Eastern Temperate Forest, Great Plains, Northwestern Forested Mountains, and North American Deserts. Streams in the Eastern Temperate Forest ecoregions are most frequently influenced by, in decreasing order, hydrology, climate, urban, and soils &

geology features. For the Great Plains, the most frequent controlling features are soils & geology, followed distantly by hydrology. The Northwestern Forested Mountains are influenced by soils & geology, climate, hydrology, and topography. Lastly, the North American Desert streams are controlled almost exclusively by soils & geology features.

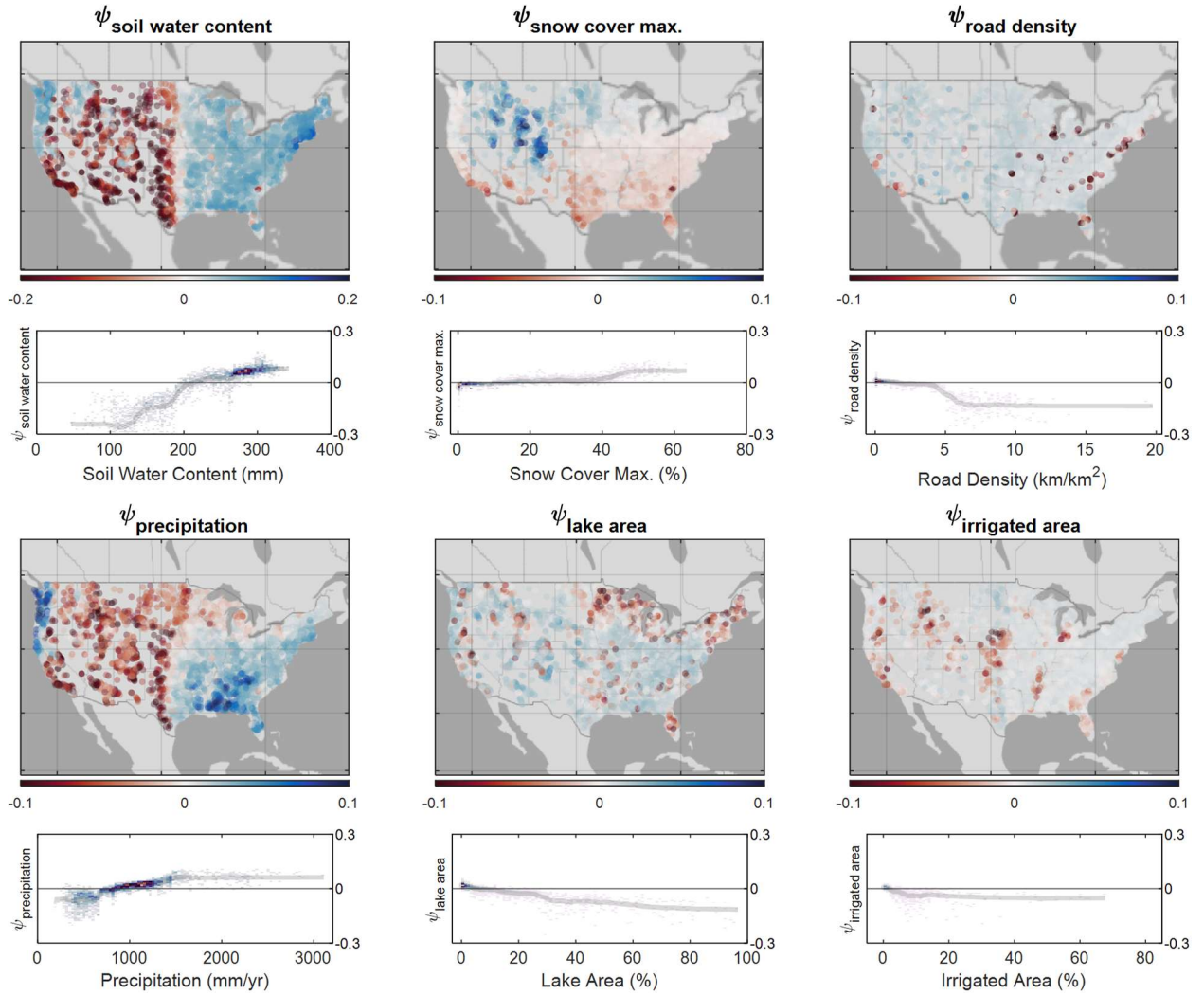



Figure 5: Spatial distribution of Shapley values (ψ) for selected influential features and their impact on Kling–Gupta efficiency (KGE) prediction for the National Water Model (NWM). ~~The colorbar represents the magnitude of ψ.~~ The partial dependence plot of each feature is shown. Features value distributions are represented with a heatmap. A moving average of feature values is indicated by a line to show general trends.

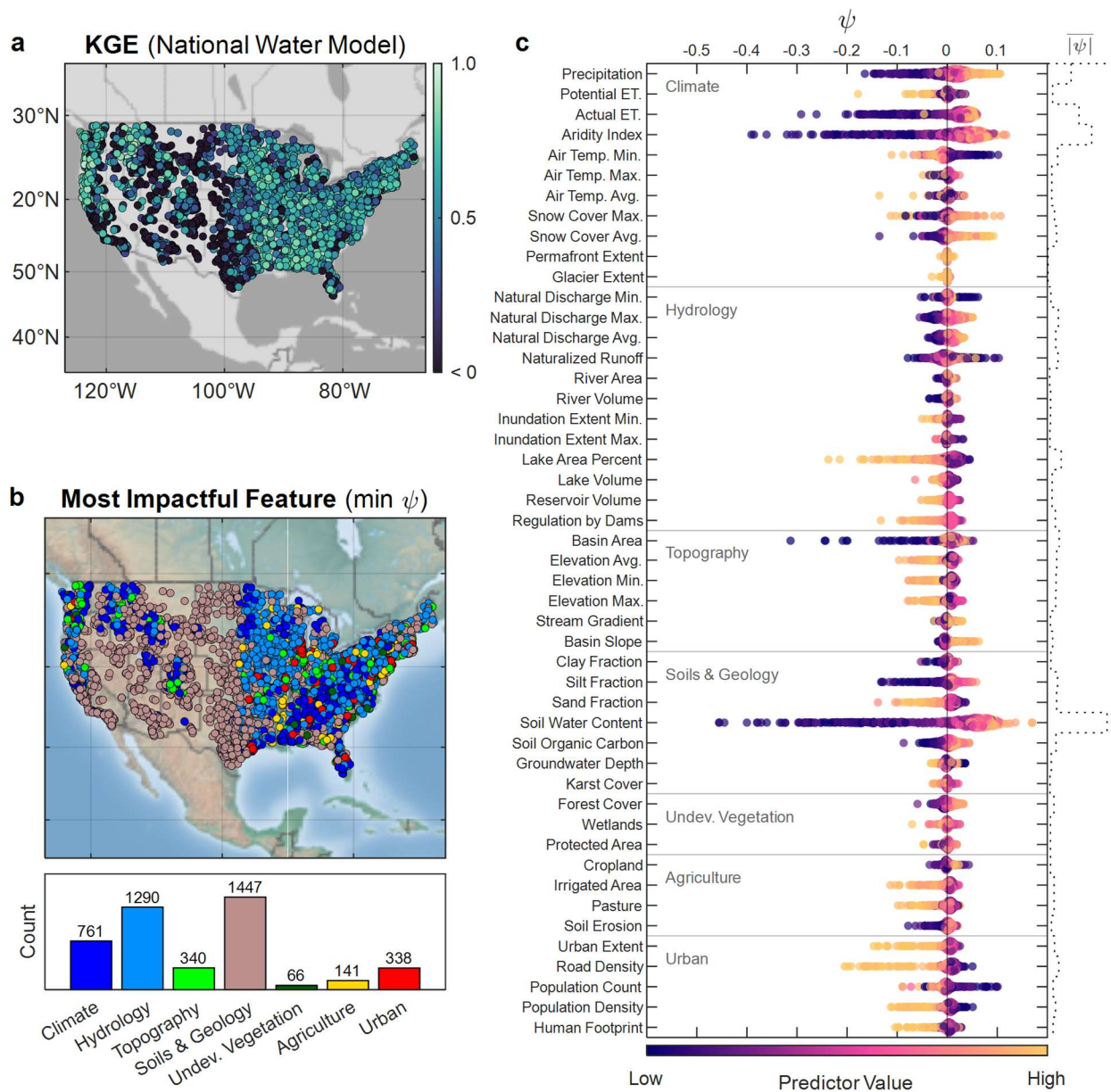



Figure 6: (a) Map of Kling–Gupta efficiency (KGE) for the National Water Model. (b) Map and histogram of the most impactful feature causing poor model performance at each site, i.e., the predictor group having the greatest negative Shapley value (ψ) at a site. (c) Swarm chart of Shapley values for KGE prediction showing feature importance for 48 predictors. The staircase plot on the right axis indicates the mean absolute Shapley value $\overline{|\psi|}$) of all observations for a predictor.

The predictor value is the magnitude of the catchment attribute.

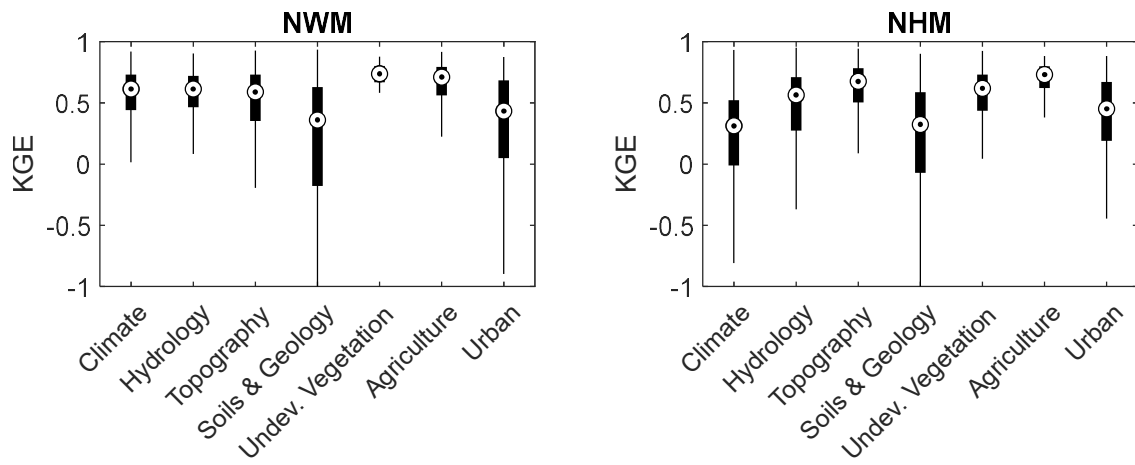



Figure 7: Kling–Gupta efficiency (KGE) performance grouped by the most important variable at each site as identified by Shapley values
for the National Water Model (NWM) and National Hydrologic Model (NHM).

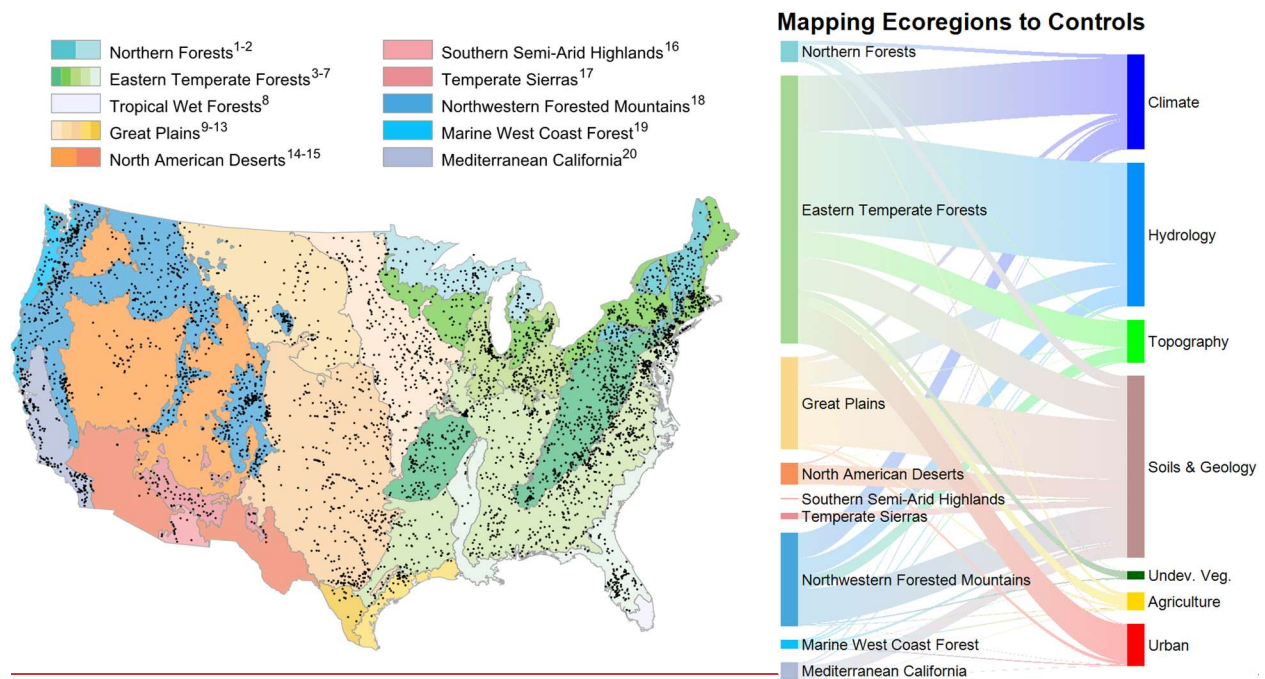
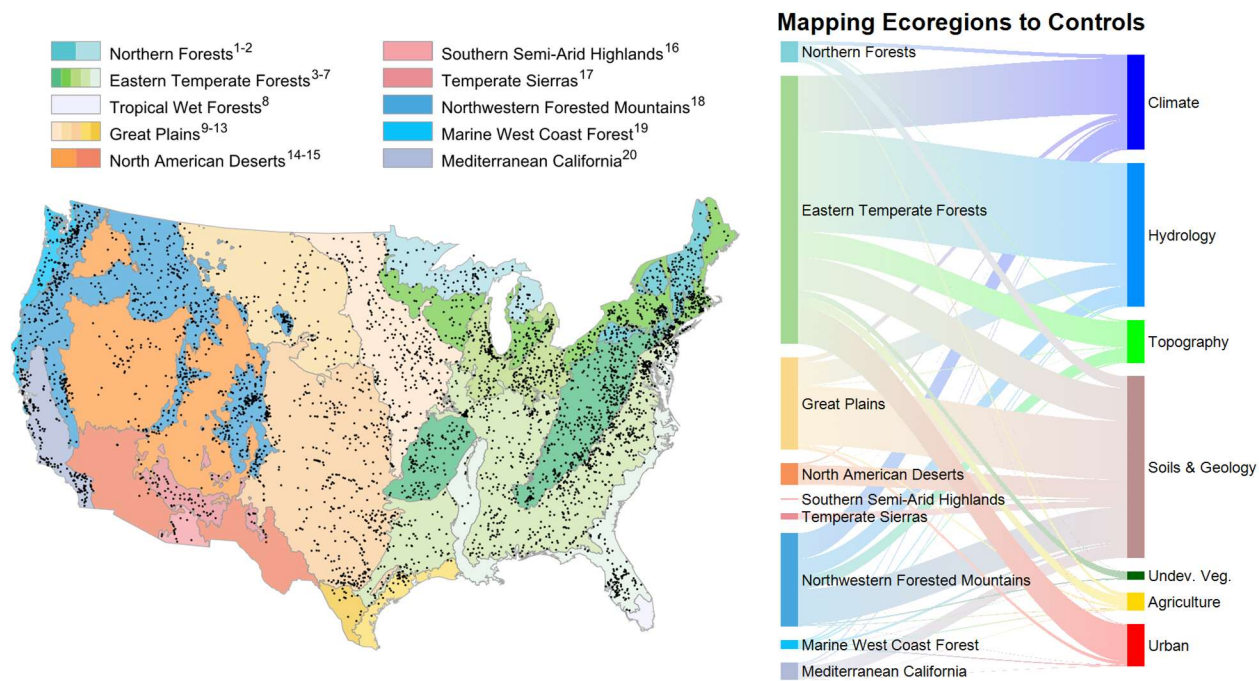



Figure 8: Map of study stream gages (black markers) and the Ecological Regions of North America (as defined in Omernik, 1987). Sankey diagram showing the pairing of ecoregions and impactful feature classes for the National Water Model (NWM) for the Kling–Gupta efficiency (KGE) evaluation metric. ~~Superscripts in ecoregion classifications are defined in Section 2.3~~Ecoregion classifications are defined using the following superscripts: [1]Atlantic Highlands, [2]Mixed Wood Shield, [3]Ozark, Ouachita-Appalachian Forests, [4]Mixed Wood Plains, [5]Central USA Plains, [6]Southeastern USA Plains, [7]Mississippi Alluvial and Southeast USA Coastal Plains, [8]Everglades, [9]Temperate Prairies, [10]West-Central Semi-Arid Prairies, [11]South Central Semi-Arid Prairies, [12]Texas-Louisiana Coastal Plain, [13]Tamaulipas-Texas Semi-Arid Plain, [14]Cold Deserts, [15]Warm Deserts, [16]Western Sierra Madre Piedmont, [17]Upper Gila Mountains, [18]Western Cordillera, [19]Marine West Coast Forest, and [20]Mediterranean California.

4 Discussion

We investigate the relative importance of catchment attributes to streamflow model performance to diagnose deficiencies in how the hydrologic models represent physical processes. Compared to other parameter-based continental-scale sensitivity analyses (e.g., Mai et al., 2022), our approach provides a post-hoc assessment of model sensitivity. That is, perturbing the parameterization of the original modeling framework is not necessary to identify model sensitivities. Rather, sensitivities are deduced (learned) through the identification of the marginal contribution of predictor features to model performance. ~~That is~~In this way, our approach identifies how catchment attributes may impact ~~the predictions of KGE at a site~~KGE—rather than how model parameters directly impact KGE. The interpretable machine learning approach we present is flexible and model agnostic, meaning it can be applied to any modeling framework.

4.1 Model diagnostics with explainable AI

The Shapely value approach used in our study is able to make both local (Fig. ~~3) and global (Fig. 4) inferences from the same model. Shapley dependence plots allow us to infer the individual (marginal) contribution of a feature to the overall model as a function of the feature's magnitude. Other approaches, such as LIME (local) and PDP (global) can only calculate one or the other. Shapley values have also been shown to match human intuition more closely compared to LIME (Lundberg and Lee, 2017), providing confidence in the approach.~~4) and global (Fig. 5) inferences from the same model. Shapley dependence plots allow us to infer the individual (marginal) contribution of a feature to the overall model as a function of the feature's magnitude. Compared to traditional sensitivity analyses, which perturb model parameters and observe the resulting impact to a performance evaluation metric (Pianosi et al., 2016), this approach identifies spatial patterns in where models perform well and where they do not, and relates that pattern to the spatial variation in catchment attributes. This indirect approach to model sensitivity allows for the identification of attributes that show a high degree of influence on model performance. This approach can serve as an interrogation tool for prioritizing which processes should be better represented within the evaluated hydrologic model structure. Below, we highlight both local and global structures that emerge from our analysis and that allow for the interrogation of NWM and NHM model performance.

Local structures emerge whereby a few sensitive attributes can dominate the overall KGE prediction at a site (Fig. ~~3~~4). This can manifest as a catchment attribute decreasing or increasing prediction accuracies (as measured by KGE) of NWM or NHM. For example, at an arid site on the Tucannon River (WA), the NWM performance is lower at this site than the nation-wide average of NWM for all sites because of high actual evapotranspiration and low precipitation conditions. Conversely, at Seboeis River (ME), the higher humidity and soil water content contributes to higher NWM prediction accuracy compared to the nation-wide average site. In some instances, multiple competing attributes offset their negative and positive contributions to KGE prediction. At the Seboeis River, the positive contribution to KGE from high soil water content is offset by the negative contribution of a large lake area percentage. Another way to interpret this would be that in the absence of lakes in the basin, the NWM would produce more accurate streamflow predictions at this site, ~~i.e.,~~that is, a higher KGE. Therefore, although the model's representation of soil water content at this site increases streamflow prediction accuracy, the simulation of lake water storage (or lack thereof) is inhibiting streamflow prediction. Importantly, the Shapley value approach can also identify features that are not influential to KGE. For example, for all three sites investigated in Fig. ~~3~~4, the natural vegetation and agricultural variables have limited influence on KGE. By elucidating the local structure of catchment controls on model performance, this approach allows for inference about which processes are not well represented by the model. Addressing these processes could be prioritized in further iterations of models to facilitate large increases in model accuracy.

Global structures emerge whereby the Shapley value approach can identify thresholds at which features become influential (Fig. ~~4~~5). Because our approach considers all sites simultaneously, we can make conclusions about the spatial coherence of influential attributes across regions (Mai et al., 2022). A few variables, most prominently soil water content, are

highly influential regardless of ~~the value of~~whether the variable ~~(i.e., whether~~ takes a small or large~~).~~ value. However, some variables largely have no influence until certain thresholds are crossed~~,~~ (Fig. 5), such as snow cover, road density, irrigation area, and lake area. The ability to resolve threshold behavior in model performance allows for better parameterization of models and identification of areas where increased data collection could improve model calibration (Zehe and Sivapalan, 2009).

This model diagnostic approach provided intuitive results that match the general understanding of streamflow controls across ecoregions (Figs. 6 and. S5). The features that commonly decreased model accuracy the most at individual sites (min ψ) were related to soils & geology, hydrology, and climate predictor groups (Fig. ~~5~~6). The influence of other predictor groups is more variable. For example, urban features (urban extent, road density, population count and density, and human footprint index) are influential in catchments near large metropolitan areas, such ~~as near~~ Chicago, New York, and Boston, but their influence is largely absent elsewhere. Urban features are the most influential predictors for just 7.7% of all gages, but these urban-controlled sites have low KGE values that are similar to ~~KGE at~~ sites controlled by the most influential variable group, soils and geology (Fig. ~~6~~7). In this way, Shapley values show utility in interrogating process-based models by allowing for the identification overarching controls across all sites in a dataset while not obscuring unique, local controls.

4.2 Natural and anthropogenic ~~process representation within the NWM~~drivers of NWM and NHM performance

4.2.1 Climate

Climate processes are of central importance to the goodness-of-fit for the NWM for many sites (Fig. ~~5~~6), as indicated by large absolute Shapley values ($\overline{|\psi|}$) for climate variables. These results align with results of multiple studies focused on climate processes as drivers for streamflow processes, such as non-perennial streamflow (Hammond et al., 2021; Price et al., 2021; Zipper et al., 2021) and peak streamflow (McMillan et al., 2018). Shapley values results show that climate processes that are related to low water availability (i.e., low values of precipitation, aridity, and ET) decrease the predictive capacity of the NWM (Fig. ~~4~~5). The inverse is also true, in that streamflow can be simulated more accurately at sites with higher precipitation and lower ET (Fig. ~~5). While prior studies~~ 6). Thus, while the NWM is recognized to have ~~observed the~~ poor performance ~~of the NWM to aridity~~in arid locations (Johnson et al., 2023b), ~~fewer have highlighted the potential impact of climate variables to improve model performance at~~our results show that it is well-suited for prediction in humid locations.

Soil water content, actual ET, and precipitation are the most influential ~~values~~features for determining KGE, all of which are highly seasonal (Elnashar et al., 2021). For example, the spatial map of KGE performance (Fig. ~~5~~6) is broadly related to precipitation amount and the Shapley value for precipitation (Fig. ~~3~~5; Lute and Luce, 2017). In areas where climate may have a lower degree of variance throughout the year, NWM accurately simulates streamflow because of the predictability of the hydrologic response in a basin. As an example, we find that the presence of a considerable snow cover (> 40%; Fig. ~~4~~5) can improve model predictability, which has been noted elsewhere (Johnson et al., 2023b) and may be

related to the predictability of seasonal snowmelt, which can dominate the water balance in cold regions. These results highlight the ability of Shapley values to elucidate the relationships between climate and streamflow and provide important insights into careful parameterization of climate forcings to increase model accuracy.

4.2.2 Hydrology

Of the variables in the hydrology category, we observed the largest effect on KGE in the NWM from lake area and upstream reservoir storage relative to annual flow volume (the degree of regulation), with KGE decreasing as lake area and the degree of regulation increase (Figs. 3 and 4). The modeling of pond and lake storage and release is a known deficiency in large-scale hydrologic modeling, and recent parameterizations have been developed to enhance representation of surface-water depression storage (Costigan and Daniels, 2012; Hay et al., 2018; Hodgkins et al., 2024).

The negative impact of lake and reservoir features on model accuracy is greater to the NHM (Fig. S3) than to the NWM (Fig. 4_5_). As noted earlier, the NHM framework does not simulate any kind of reservoir operations, water withdrawals, or stream releases (Regan et al., 2019). On the other hand, the NWM framework models passive reservoir routing (Cosgrove et al., 2024) to mitigate the confounding effects of lake and reservoir volume on model performance. ~~The successful identification of a hydrologic model sensitivity by the Shapley value approach underscores~~The Shapley value approach was able to successfully identify that the model without any provision for reservoirs (NHM) is more negatively affected by the presence of reservoirs than the model with routing capability (NWM), underscoring that the method is highly interpretable and can produce intuitive results that match our conceptual models.

4.2.3 Physiography (Topography, Soils, and Geology)

Hydrologic connectivity controls many facets of the natural flow regime and determines the ability of a watershed to store and release water ~~(Husic and Michalek, 2022).~~(Michalek et al., 2023). Parameterizations of soils, geology, and other basin characteristics are highly heterogeneous and mediate many facets of connectivity, many of which are poorly resolved in large-scale hydrologic models (Li et al., 2023). ~~Soil water content was the most impactful predictor for KGE according to the Shapley value analysis, with low values of soil water content greatly impacting the KGE (Fig. 4). Soil water content represents the annual soil water available for evapotranspiration, with complete soil saturation as an upper limit. Other factors that contribute to a high degree of hydrologic connectivity, such as high percent sand and low percent clay (Fig. 5), also highlight the inability of the NWM to resolve storage and connectivity~~For example, soil water content was the most impactful predictor for KGE according to the Shapley value analysis (Fig. 5), with low values of soil water content greatly impacting the KGE. Accurate simulation of soil moisture patterns, particularly in arid locations, is a well-recognized challenge in the NWM, which can be mitigated by the integration of soil moisture data into the model calibration process (Araki et al., 2025). Other factors that contribute to a high degree of hydrologic connectivity, including high percent sand and low percent clay (Fig. 6), also highlight the inability of the NWM to resolve storage dynamics, which likely results from

inadequate parameterization of areas that have highly seasonal soil water content (Hughes et al., 2024) and the inability of
420   the current generation of NWM to represent losing streams (Jachens et al., 2021; Lahmers et al., 2021).

We also identified predictor variables commonly associated with the physiography of headwater systems as important predictors of KGE (Fig. ~~5~~6), such as drainage area and mean elevation. Headwater systems are defined as "surface-water catchment areas and groundwater zones that contribute water, material, and energy to a headwater stream" ~~(Brinkerhoff et al., 2024; Golden et al., 2024).~~(Brinkerhoff et al., 2024; Golden et al., 2025). Headwater streams typically
425   have smaller drainage areas and higher mean elevations, which our approach found were associated with lower KGE values for NWM predictions possibly because NWM simulates atmospheric states and fluxes on a 1×1 km^2 grid cell and can misrepresent processes that are on the scale of headwater systems. These headwater systems are low-order and highly variable in their flow regimes (Rojas et al., 2020), both of which are inadequately represented in NWM.

4.2.4 Anthropogenic processes

430   Of the variables related to anthropogenic influence, we note that urban features, such as urban extent, road density, population ~~county~~count, population density, and human footprint, typically decrease KGE values for modeled streamflows (Figs. 5 and S4). The construction of urban drainage networks has been recognized to increase the connectivity of water, solutes, and sediment, and to add additional pathways of transport through the artificial routing of water ~~(Lakoba et al., 2020; Zarnaghsh and Husic, 2021).~~(Zarnaghsh and Husic, 2021). In a continental-scale analysis of the NWM, urban areas
435   exhibited some of the largest bias (Johnson et al., 2023b), in part due to the presence of constructed drainage networks. Notwithstanding this limitation, the NWM has shown some success in simulating hydrology when artificial urban channels, which differ from natural flow paths, are manually delineated within the flow grid (Pasquier et al., 2022). However, manual delineation is not feasible for applications at intended for regional or continental scales, such as NWM and NHM.

Our model identifies a threshold of around 5 km/km^2 of roadways as the initiation point whereby the presence of
440   roadways decreases accuracies of NWM and NHM predictions (Figs. 4 and S3). The sensitivity of the roadway density feature may indicate other associated infrastructure, the configuration of proximal impervious areas, and the relative amount of human alternation of surface flow generation and routing mechanisms not picked up by considering imperious area alone. Population and population density similarly likely indicate associated infrastructure that alters flow timing and magnitude of water delivery to rivers (Hopkins et al., 2019). For example, leaky infrastructure can result in elevated low flows beyond
445   natural background levels (Bhaskar et al., 2020). Regarding agriculture, irrigation return flows have been shown to be important to flow generation processes, particularly in lower elevation, arid rivers (Putman et al., 2024). These urban and agricultural features can decrease model accuracy when present, but the absence of these features does not necessarily increase model accuracy (Fig. ~~5~~6).

4.3 Limitations and Future Research

450   Our interpretable modeling approach has provided several insights into interrogating process deficiencies in the NWM and NHM. Although the inferences we derived from the Shapley values are robust, interpretable, and intuitive, the analysis approach itself is not causative (Lundberg et al., 2020). Thus, some inferences may occur due to indirect correlation (Heskes et al., 2020). We took precautions to mitigate the effect of feature correlations while constructing the random forest model, such as through random exclusion of features during tree construction and out-of-bag sampling (Fox et al., 2017).

455 Our approach provides us with confidence because, as we noted earlier, many of the inferences we derived with the Shapley values match the causative ~~"under the hood"~~and mechanistic model assessments performed by others (Hodgkins et al., 2024; Hughes et al., 2024; Jachens et al., 2021; Pasquier et al., 2022).

   The interpretable modeling approach has its own set of limitations. First, predictions made by Shapley values are a function of (1) the set of sites considered, in this case 4,383 streamgages in the United States used in NWM and NHM

460 assessment and (2) the choice and performance of the predictive model, which in this case was a reasonably accurate random forest model ($R^2 \geq 0.43$). With regard to the first point, if our analysis approach were applied to interpreting the KGE values for streamflow predictions made by applying the Soil Water and Assessment Tool (SWAT) to Europe (Abbaspour et al., 2015), the order and magnitude of influence by various features would undoubtedly change. To the second point, although our random forest model is reasonably accurate, it only explains 47% of the variance in KGE prediction for the NWM (and

465 43% for the NHM). ~~Thus, while~~While our model effectively captures dominant global trends and local structures, it still leaves more than half of the variance in ~~predicting the~~ KGE ~~is unaccounted for;~~predictions unexplained. Future studies could explore ways to further ~~reduce model variances could be explored in future studies. Further~~explain this variance. Additionally, we ~~only~~ consider only the KGE goodness-of-fit metric in this study, but if we were to interpret other goodness-of-fit metrics, such as the Nash-Sutcliffe Efficiency, there is potential that inferred controls on model performance may

470 change. This is because all goodness-of-fit metrics encode for – and are biased by – various information contained within streamflow timeseries (Clark et al., 2021). Nonetheless, of the common evaluation metrics presently applied in the hydrologic literature, use of the KGE is increasing because of its lower overall bias and provision for balanced results during low- and high-flow conditions (Althoff and Rodrigues, 2021).

   Several ~~limitations and~~ opportunities ~~for improvement~~ exist ~~regarding~~for overcoming limitations and making

475 improvements to the data inputs and model outputs. First, the spatial extent and resolution of the catchment attribute dataset may be too coarse, particularly for smaller basins. Of the 48 catchment attributes derived from the BasinATLAS dataset (Linke et al., 2019), spatial ~~resolution can~~resolutions range from 3 arc-~~second~~seconds for elevation to 5 arc-~~minute~~minutes for land use ~~(e.g., cropland and urban extents).~~. At 40º N, the median latitude of the CONUS, these arc values correspond to ~85 meters and ~7 kilometers, respectively. These datasets were aggregated to 15-arcseconds (~350 m), thus the calculated

480 attributes for smaller basins are more uncertain due to a smaller sample size of attribute estimates contained within basin bounds. A second data limitation is that the catchment attribute dataset represents snapshot-in-time value for all basins

(Linke et al., 2019). However, catchments and their characteristics, particularly land use, may change substantially over time. The hydrologic models are simulated over multiple decades (1984 to 2016), during which change may occur and be captured within the process-based representation of the models but not in the catchment attribute dataset. Improved spatial resolution and temporal evolution of catchment attributes could provide deeper insights into identifying NWM and NHM process deficiencies. Finally, the process-based models used here vary in their spatial and physical representation of hydrologic processes. Process-based model differences in routing schema, spatial groupings (hydrologic response unit vs grid-based), and subsurface properties could result in slight differences but are unlikely to impact or explain broad, CONUS scale patterns observed in our analysis.

Looking forward, the National Oceanic and Atmospheric Administration (NOAA), the developers of NWM, are expanding modeling capacity with their Next Generation Water Resources Modeling Framework (NextGen; Ogden et al., 2021). In addition to a uniformly applied national hydrologic model, there will be tools for identifying the best model/parameterization for each individual location and then modeling regions as patchworks of individual/local models (Cosgrove et al., 2024). In addition to assessing overall flow performance, this approach could be used for specific components of the flow regime, such as high and low flows. For example, studies that have focused on individual components of non-perennial drying regimes have used a random forest approach coupled with partial-dependency analysis (e.g., Price et al., 2021). The Shapley value approach used in this study could be used in a similar way to evaluate magnitude and directionality of impact between predictor values and flow regimes across systems. Further, modules are planned for purely data-driven approaches, like Long-Short Term Memory models ~~(Frame et al., 2021).~~(Frame et al., 2025, 2021). Our interpretable modeling approach provides a starting point to inform the parametrization of local-scale and regional-scale applications in the next generation of hydrologic models.

5 Conclusions

The interpretable machine learning technique we present is flexible and model agnostic. We use the technique to identify potential process-based deficiencies in two continental scale hydrologic models: ~~NWM and NHM.~~the National Water Model and the National Hydrologic Model. Compared to other parameter-based continental-scale sensitivity analyses, our approach provides a post-hoc assessment of model sensitivity. This method allows for the identification of thresholds after which a feature begins to negatively impact streamflow model performance. Globally, soil water content was the most common feature influencing the accuracies of streamflow simulations, followed by aridity, evapotranspiration, and precipitation. We interpret the results to indicate that the present formulations of NWM and NHM have limited ability to resolve soil water storage and release, particularly in arid locations. Locally, the presence of lakes and ~~flow regulation~~reservoirs were related to decreased model accuracy as ~~were~~was the presence of roadways and irrigation canals. Our results suggest that further refining how these influential processes are represented in large scale hydrological models would result in the largest increases in model accuracies. This study demonstrates the utility of interrogating process-based

models using data-driven techniques and explainable AI, which has broad applicability and potential for improving simulation of large-scale hydrology and water quality.

Code availability

The data and code used for the random forest model, the Shapley value analysis, and generation of figures can be found at the following Open Science Framework link: https://doi.org/10.17605/OSF.IO/MNQCZ.

Author contributions

AH: Conceptualization, Methodology, Formal analysis, Visualization, Writing – original draft, Writing – review & editing. **JH**: Methodology, Formal analysis, Writing – original draft, Writing – review & editing. **AP**: Methodology, Writing – original draft, Writing – review & editing. **JK**: Writing – original draft, Writing – review & editing.

Competing interests

The contact author has declared that none of the authors have any competing interests.

Acknowledgements

This work was supported by the National Science Foundation (Award Nos. 2229616 and 2438017). This work was performed at the HPC facilities operated by the Center for Research Computing at the University of Kansas supported in part through the National Science Foundation MRI Award OAC-2117449. Any use of trade, firm, or product names is for descriptive purposes only and does not imply endorsement by the U.S. Government. This work has been reviewed by the U.S. Forest Service. This product has been peer reviewed and approved for publication consistent with USGS Fundamental Science Practices (https://pubs.usgs.gov/circ/1367/).

References

535   Abbaspour, K.C., Rouholahnejad, E., Vaghefi, S., Srinivasan, R., Yang, H., Kløve, B., 2015. A continental-scale hydrology and water quality model for Europe: Calibration and uncertainty of a high-resolution large-scale SWAT model. J. Hydrol. 524, 733–752. https://doi.org/10.1016/j.jhydrol.2015.03.027

Althoff, D., Rodrigues, L.N., 2021. Goodness-of-fit criteria for hydrological models: Model calibration and performance assessment. J. Hydrol. 600, 126674. https://doi.org/10.1016/j.jhydrol.2021.126674

540   Araki, R., Ogden, F.L., McMillan, H.K., 2025. Testing Soil Moisture Performance Measures in the Conceptual-Functional Equivalent to the WRF-Hydro National Water Model. JAWRA J. Am. Water Resour. Assoc. 61.

Beven, K., 2024. A brief history of information and disinformation in hydrological data and the impact on the evaluation of hydrological models. Hydrol. Sci. J. 69, 519–527. https://doi.org/10.1080/02626667.2024.2332616

Bhaskar, A.S., Hopkins, K.G., Smith, B.K., Stephens, T.A., Miller, A.J., 2020. Hydrologic Signals and Surprises in U.S.
545   Streamflow Records During Urbanization. Water Resour. Res. 56, 1–22. https://doi.org/10.1029/2019WR027039

Blasone, R.S., Vrugt, J.A., Madsen, H., Rosbjerg, D., Robinson, B.A., Zyvoloski, G.A., 2008. Generalized likelihood uncertainty estimation (GLUE) using adaptive Markov Chain Monte Carlo sampling. Adv. Water Resour. 31, 630–648. https://doi.org/10.1016/j.advwatres.2007.12.003

Blöschl, G., Bierkens, M.F.P., Chambel, A., Cudennec, C., Destouni, G., Fiori, A., Kirchner, J.W., McDonnell, J.J.,
550   Savenije, H.H.G., Sivapalan, M., Stumpp, C., Toth, E., Volpi, E., Carr, G., Lupton, C., Salinas, J., Széles, B., Viglione, A., Aksoy, H., Allen, S.T., Amin, A., Andréassian, V., Arheimer, B., Aryal, S.K., Baker, V., Bardsley, E., Barendrecht, M.H., Bartosova, A., Batelaan, O., Berghuijs, W.R., Beven, K., Blume, T., Bogaard, T., Borges de Amorim, P., Böttcher, M.E., Boulet, G., Breinl, K., Brilly, M., Brocca, L., Buytaert, W., Castellarin, A., Castelletti, A., Chen, X., Chen, Yangbo,Y., Chen, Yuanfang,Y., Chifflard, P., Claps, P., Clark, M.P., Collins, A.L., Croke, B., Dathe,
555   A., David, P.C., de Barros, F.P.J., de Rooij, G., Di Baldassarre, G., Driscoll, J.M., Duethmann, D., Dwivedi, R., Eris, E., Farmer, W.H., Feiccabrino, J., Ferguson, G., Ferrari, E., Ferraris, S., Fersch, B., Finger, D., Foglia, L., Fowler, K., Gartsman, B., Gascoin, S., Gaume, E., Gelfan, A., Geris, J., Gharari, S., Gleeson, T., Glendell, M., Gonzalez Bevacqua, A., González-Dugo, M.P., Grimaldi, S., Gupta, A.B., Guse, B., Han, D., Hannah, D., Harpold, A., Haun, S., Heal, K., Helfricht, K., Herrnegger, M., Hipsey, M., Hlaváčiková, H., Hohmann, C., Holko, L., Hopkinson, C.,
560   Hrachowitz, M., Illangasekare, T.H., Inam, A., Innocente, C., Istanbulluoglu, E., Jarihani, B., Kalantari, Z., Kalvans, A., Khanal, S., Khatami, S., Kiesel, J., Kirkby, M., Knoben, W., Kochanek, K., Kohnová, S., Kolechkina, A., Krause, S., Kreamer, D., Kreibich, H., Kunstmann, H., Lange, H., Liberato, M.L.R., Lindquist, E., Link, T., Liu, J., Loucks, D.P., Luce, C., Mahé, G., Makarieva, O., Malard, J., Mashtayeva, S., Maskey, S., Mas-Pla, J., Mavrova-Guirguinova, M., Mazzoleni, M., Mernild, S., Misstear, B.D., Montanari, A., Müller-Thomy, H., Nabizadeh, A., Nardi, F., Neale, C.,

Nesterova, N., Nurtaev, B., Odongo, V.O., Panda, S., Pande, S., Pang, Z., Papacharalampous, G., Perrin, C., Pfister, L., Pimentel, R., Polo, M.J., Post, D., Prieto Sierra, C., Ramos, M.-H., Renner, M., Reynolds, J.E., Ridolfi, E., Rigon, R., Riva, M., Robertson, D.E., Rosso, R., Roy, T., Sá, J.H.M., Salvadori, G., Sandells, M., Schaefli, B., Schumann, A., Scolobig, A., Seibert, J., Servat, E., Shafiei, M., Sharma, A., Sidibe, M., Sidle, R.C., Skaugen, T., Smith, H., Spiessl, S.M., Stein, L., Steinsland, I., Strasser, U., Su, B., Szolgay, J., Tarboton, D., Tauro, F., Thirel, G., Tian, F., Tong, R., Tussupova, K., Tyralis, H., Uijlenhoet, R., van Beek, R., van der Ent, R.J., van der Ploeg, M., Van Loon, A.F., van Meerveld, I., van Nooijen, R., van Oel, P.R., Vidal, J.-P., von Freyberg, J., Vorogushyn, S., Wachniew, P., Wade, A.J., Ward, P., Westerberg, I.K., White, C., Wood, E.F., Woods, R., Xu, Z., Yilmaz, K.K., Zhang, Y., 2019. Twenty-three unsolved problems in hydrology (UPH) – a community perspective. Hydrol. Sci. J. 64, 1141–1158. https://doi.org/10.1080/02626667.2019.1620507

Brêda, J.P.L.F., Melsen, L.A., Athanasiadis, I., Van Dijk, A., Siqueira, V.A., Verhoef, A., Zeng, Y., van der Ploeg, M., 2024. Predictor Importance for Hydrological Fluxes of Global Hydrological and Land Surface Models. Water Resour. Res. 60.

Brinkerhoff, C.B., Gleason, C.J., Kotchen, M.J., Kysar, D.A., Raymond, P.A., 2024. Ephemeral stream water contributions to United States drainage networks. Science (80-.). 384, 1476–1482. https://doi.org/10.1126/science.adg9430

Brunner, M.I., Slater, L., Tallaksen, L.M., Clark, M., 2021. Challenges in modeling and predicting floods and droughts: A review. Wiley Interdiscip. Rev. Water 8, 1–32. https://doi.org/10.1002/wat2.1520

Clark, M.P., Vogel, R.M., Lamontagne, J.R., Mizukami, N., Knoben, W.J.M., Tang, G., Gharari, S., Freer, J.E., Whitfield, P.H., Shook, K.R., Papalexiou, S.M., 2021. The Abuse of Popular Performance Metrics in Hydrologic Modeling. Water Resour. Res. 57, 1–16. https://doi.org/10.1029/2020WR029001

Cosgrove, B., Gochis, D., Flowers, T., Dugger, A., Ogden, F., Graziano, T., Clark, E., Cabell, R., Casiday, N., Cui, Z., Eicher, K., Fall, G., Feng, X., Fitzgerald, K., Frazier, N., George, C., Gibbs, R., Hernandez, L., Johnson, D., Jones, R., Karsten, L., Kefelegn, H., Kitzmiller, D., Lee, H., Liu, Y., Mashriqui, H., Mattern, D., McCluskey, A., McCreight, J.L., McDaniel, R., Midekisa, A., Newman, A., Pan, L., Pham, C., RafieeiNasab, A., Rasmussen, R., Read, L., Rezaeianzadeh, M., Salas, F., Sang, D., Sampson, K., Schneider, T., Shi, Q., Sood, G., Wood, A., Wu, W., Yates, D., Yu, W., Zhang, Y., 2024. NOAA's National Water Model: Advancing operational hydrology through continental-scale modeling. JAWRA J. Am. Water Resour. Assoc. 1–26. https://doi.org/10.1111/1752-1688.13184

Costigan, K.H., Daniels, M.D., 2012. Damming the prairie: Human alteration of Great Plains river regimes. J. Hydrol. 444–445, 90–99. https://doi.org/10.1016/j.jhydrol.2012.04.008

De'Ath, G., Fabricius, K.E., 2000. Classification and regression trees: A powerful yet simple technique for ecological data analysis. Ecology 81, 3178–3192. https://doi.org/10.1890/0012-9658(2000)081[3178:CARTAP]2.0.CO;2

De Meester, J., Willems, P., 2024. Analysing spatial variability in drought sensitivity of rivers using explainable artificial intelligence. Sci. Total Environ. 931, 172685.

Elnashar, A., Wang, L., Wu, B., Zhu, W., Zeng, H., 2021. Synthesis of global actual evapotranspiration from 1982 to 2019.

Earth Syst. Sci. Data 13, 447–480. https://doi.org/10.5194/essd-13-447-2021

600   Fox, E.W., Hill, R.A., Leibowitz, S.G., Olsen, A.R., Thornbrugh, D.J., Weber, M.H., 2017. Assessing the accuracy and stability of variable selection methods for random forest modeling in ecology. Environ. Monit. Assess. 189. https://doi.org/10.1007/s10661-017-6025-0

Frame, J.M., Araki, R., Bhuiyan, S.A., Bindas, T., Rapp, J., Bolotin, L., Deardorff, E., Liu, Q., Haces-Garcia, F., Liao, M., Frazier, N., Ogden, F.L., 2025. Machine Learning for a Heterogeneous Water Modeling Framework. J. Am. Water
605   Resour. Assoc. 61, 1–10.

Frame, J.M., Kratzert, F., Raney, A., Rahman, M., Salas, F.R., Nearing, G.S., 2021. Post-Processing the National Water Model with Long Short-Term Memory Networks for Streamflow Predictions and Model Diagnostics. J. Am. Water Resour. Assoc. 57, 885–905. https://doi.org/10.1111/1752-1688.12964

Friedman, J., 2001. Greedy Function Approximation: A Gradient Boosting Machine. Ann. Stat. 29, 1189–1232.

610   Golden, H.E., Christensen, J.R., McMillan, H.K., Kelleher, C.A., Lane, C.R., Husic, A., Li, L., Ward, A.S., Hammond, J., Seybold, E.C., Jaeger, K.L., Zimmer, M., Sando, R., Jones, C.N., Segura, C., Mahoney, D.T., Price, A.N., Cheng, F.Y., 2024., 2025. Advancing the Sciencescience of Headwater Streamflowheadwater streamflow for Global global water protection. Nat. Water Protection. Nat. Water In Review3, 16–26.

Gupta, H. V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009. Decomposition of the mean squared error and NSE
615   performance criteria: Implications for improving hydrological modelling. J. Hydrol. 377, 80–91. https://doi.org/10.1016/j.jhydrol.2009.08.003

Hammond, J.C., Zimmer, M., Shanafield, M., Kaiser, K., Godsey, S.E., Mims, M.C., Zipper, S.C., Burrows, R.M., Kampf, S.K., Dodds, W., Jones, C.N., Krabbenhoft, C.A., Boersma, K.S., Datry, T., Olden, J.D., Allen, G.H., Price, A.N., Costigan, K., Hale, R., Ward, A.S., Allen, D.C., 2021. Spatial Patterns and Drivers of Nonperennial Flow Regimes in
620   the Contiguous United States. Geophys. Res. Lett. 48, 1–11. https://doi.org/10.1029/2020GL090794

Hay, L., Norton, P., Viger, R., Markstrom, S., Steven Regan, R., Vanderhoof, M., 2018. Modelling surface-water depression storage in a Prairie Pothole Region. Hydrol. Process. 32, 462–479. https://doi.org/10.1002/hyp.11416

Heskes, T., Sijben, E., Bucur, I.G., Claassen, T., 2020. Causal shapley values: Exploiting causal knowledge to explain individual predictions of complex models. Adv. Neural Inf. Process. Syst.

625   Ho, T.K., 1998. The random subspace method for constructing decision forests. IEEE Trans. Pattern Anal. Mach. Intell. 20, 832–844. https://doi.org/10.1109/34.709601

Hodgkins, G.A., Over, T.M., Dudley, R.W., Russell, A.M., LaFontaine, J.H., 2024. The consequences of neglecting reservoir storage in national-scale hydrologic models: An appraisal of key streamflow statistics. J. Am. Water Resour. Assoc. 60, 110–131. https://doi.org/10.1111/1752-1688.13161

630   Hopkins, K.G., Fanelli, R.M., Bhaskar, A.S., Woznicki, S.A., 2019. Changes in event-based streamflow magnitude and timing after suburban development with infiltration-based stormwater management. Hydrol. Process. https://doi.org/10.1002/hyp.13593

Huang, F., Shangguan, W., Li, Q., Li, L., Zhang, Y., 2023. Beyond prediction: An integrated post-hoc approach to interpret complex model in hydrometeorology. Environ. Model. Softw. 167.

635 Hughes, M., Jackson, D.L., Unruh, D., Wang, H., Hobbins, M., Ogden, F.L., Cifelli, R., Cosgrove, B., DeWitt, D., Dugger, A., Ford, T.W., Fuchs, B., Glaudemans, M., Gochis, D., Quiring, S.M., RafieeiNasab, A., Webb, R.S., Xia, Y., Xu, L., 2024. Evaluation of Retrospective National Water Model Soil Moisture and Streamflow for Drought-Monitoring Applications. J. Geophys. Res. Atmos. 129, 1–25. https://doi.org/10.1029/2023JD038522

Husic, A., Michalek, A., 2022. Structural Hillslope Connectivity Is Driven by Tectonics More Than Climate and Modulates
640 Hydrologic Extremes and Benefits. Geophys. Res. Lett. 49, e2022GL099898. https://doi.org/10.1029/2022GL099898

Jachens, E.R., Hutcheson, H., Thomas, M.B., Steward, D.R., 2021. Effects of Groundwater-Surface Water Exchange Mechanism in the National Water Model over the Northern High Plains Aquifer, USA. J. Am. Water Resour. Assoc. 57, 241–255. https://doi.org/10.1111/1752-1688.12869

645 Johnson, J.M., Blodgett, D.L., Clarke, K.C., Pollak, J., 2023a. Restructuring and serving web-accessible streamflow data from the NOAA National Water Model historic simulations. Sci. Data 10, 1–10. https://doi.org/10.1038/s41597-023-02316-7

Johnson, J.M., Fang, S., Sankarasubramanian, A., Rad, A.M., Kindl da Cunha, L., Jennings, K.S., Clarke, K.C., Mazrooei, A., Yeghiazarian, L., 2023b. Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous
650 Formulations and Diagnostic Model Selection. J. Geophys. Res. Atmos. 128, 1–21. https://doi.org/10.1029/2023JD038534

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. https://doi.org/10.5194/hess-22-6005-2018

Lahmers, T.M., Hazenberg, P., Gupta, H., Castro, C., Gochis, D., Dugger, A., Yates, D., Read, L., Karsten, L., Wang, Y.H.,
655 2021. Evaluation of NOAA National Water Model Parameter Calibration in Semiarid Environments Prone to Channel Infiltration. J. Hydrometeorol. 22, 2939–2969. https://doi.org/10.1175/JHM-D-20-0198.1

Lakoba, V., Wind, L., DeVilbiss, S., Lofton, M., Bretz, K., Weinheimer, A., Moore, C., Baciocco, C., Hotchkiss, E., Hession, W.C., 2020. Salt Dilution and Flushing Dynamics of an Impaired Agricultural-Urban Stream. ACS ES&T Water. https://doi.org/10.1021/acsestwater.0c00160

660 Legates, D.R., McCabe, G.J., 1999. Evaluating the use of "goodness-of-fit" Measures in hydrologic and hydroclimatic model validation. Water Resour. Res. 35, 233–241. https://doi.org/10.1029/1998WR900018

Li, C., Yu, G., Wang, J., Horton, D.E., 2023. Toward Improved Regional Hydrological Model Performance Using State-Of-The-Science Data-Informed Soil Parameters. Water Resour. Res. 59, 1–23. https://doi.org/10.1029/2023WR034431

Linke, S., Lehner, B., Ouellet Dallaire, C., Ariwi, J., Grill, G., Anand, M., Beames, P., Burchard-Levine, V., Maxwell, S.,
665 Moidu, H., Tan, F., Thieme, M., 2019. Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. Sci. Data 6, 1–16. https://doi.org/10.1038/s41597-019-0300-6

Liu, D., Guo, S., Wang, Z., Liu, P., Yu, X., Zhai, Q., Zou, H., 2018. Statistics for sample splitting for the calibration and validation of hydrological models. Stoch. Environ. Res. Risk Assess. 1–18. https://doi.org/10.1007/s00477-018-1539-8

Lundberg, S.M., Erion, G., Chen, H., DeGrave, A., Prutkin, J.M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S.I., 2020. From local explanations to global understanding with explainable AI for trees. Nat. Mach. Intell. 2, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 4766–4775.

Lute, A., Luce, C.H., 2017. National Forest Climate Change Maps: Your Guide to the Future [WWW Document]. URL https://www.fs.usda.gov/rm/boise/AWAE/projects/national-forest-climate-change-maps.html (accessed 7.1.17).

Ma, Y., Leonarduzzi, E., Defnet, A., Melchior, P., Condon, L.E., Maxwell, R.M., 2024. Water Table Depth Estimates over the Contiguous United States Using a Random Forest Model. Groundwater 62, 34–43.

Mai, J., 2023. Ten strategies towards successful calibration of environmental models. J. Hydrol. 620, 129414. https://doi.org/10.1016/j.jhydrol.2023.129414

Mai, J., Craig, J.R., Tolson, B.A., Arsenault, R., 2022. The sensitivity of simulated streamflow to individual hydrologic processes across North America. Nat. Commun. 13, 455. https://doi.org/10.1038/s41467-022-28010-7

Maier, H., Taghikhah, F., Nabavi, E., Razavi, S., Gupta, H., Wu, W., Radford, D.A.G., Huang, J., 2024. How much X is in XAI: Responsible use of "Explainable" artificial intelligence in hydrology and water resources. J. Hydrol. X 25, 100185.

MathWorks, 20232024. Shapley Values for Machine Learning Model [WWW Document]. URL https://www.mathworks.com/help/stats/shapley-values-for-machine-learning-model.html (accessed 1.10.23).

McDonnell, J.J., Spence, C., Karran, D.J., van Meerveld, H.J., Harman, C.J., 2021. Fill and Spill: A Process Description of Runoff Generation at the Scale of the Beholder. Water Resour. Res. 57, 1–13. https://doi.org/10.1029/2020WR027514

McMillan, S.K., Wilson, H.F., Tague, C.L., Hanes, D.M., Inamdar, S., Karwan, D.L., Loecke, T., Morrison, J., Murphy, S.F., Vidon, P., 2018. Before the storm: antecedent conditions as regulators of hydrologic and biogeochemical response to extreme climate events. Biogeochemistry 141, 487–501. https://doi.org/10.1007/s10533-018-0482-6

Nearing, G.S., Kratzert, F., Sampson, A.K., Pelissier, C.S., Klotz, D., Frame, J.M., Prieto, C., Gupta, H. V., 2021. What Role Does Hydrological Science Play in the Age of Machine Learning? Water Resour. Res. 57. https://doi.org/10.1029/2020WR028091

Nossent, J., Elsen, P., Bauwens, W., 2011. Sobol' sensitivity analysis of a complex environmental model. Environ. Model. Softw. 26, 1515–1525. https://doi.org/10.1016/j.envsoft.2011.08.010

Michalek, A.T., Villarini, G., Husic, A., 2023. Climate change projected to impact structural hillslope connectivity at the global scale. Nat. Commun. 14, 6788.

Morris, M.D., 1991. Factorial sampling plans for preliminary computational experiments. Technometrics 33, 161–174.

Nicolle, P., Andreássian, V., Royer-Gaspard, P., Perrin, C., Thirel, G., Coron, L., Santos, L., 2021. Technical note: RAT- A robustness assessment test for calibrated and uncalibrated hydrological models. Hydrol. Earth Syst. Sci. 25, 5013–5027.

Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., Cosgrove, B., Cui, S., Kindl da Cunha, L., Farthing, M., Flowers, T., Frame, J., Frazier, N., Graziano, T., Gutenson, J., Johnson, D., McDaniel, R., Moulton, J., Loney, D., Peckham, S., Mattern, D., Jennings, K., Williamson, M., Savant, G., Tubbs, C., Garrett, J., Wood, A., Johnson, J., 2021. The Next Generation Water Resources Modeling Framework: Open Source, Standards Based, Community Accessible, Model Interoperability for Large Scale Water Prediction, in: AGU Fall Meeting Abstracts.

Omernik, J.M., 1987. Ecoregions of the Conterminous United. Ann. Assoc. Am. Geogr. 77, 118–125.

Park, J., Lee, W.H., Kim, K.T., Park, C.Y., Lee, S., Heo, T.Y., 2022. Interpretation of ensemble learning to predict water quality using explainable artificial intelligence. Sci. Total Environ. 832, 155070. https://doi.org/10.1016/j.scitotenv.2022.155070

Pandit, A., Hogan, S., Mahoney, D.T., Ford, W.I., Fox, J.F., Wellen, C., Husic, A., 2025. Establishing performance criteria for evaluating watershed-scale sediment and nutrient models at fine temporal scales. Water Res. 274, 123156.

Pasquier, U., Vahmani, P., Jones, A.D., 2022. Quantifying the City-Scale Impacts of Impervious Surfaces on Groundwater Recharge Potential: An Urban Application of WRF–Hydro. Water (Switzerland) 14. https://doi.org/10.3390/w14193143

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. London, Edinburgh, Dublin Philos. Mag. J. Sci. 2, 559–572. https://doi.org/10.1080/14786440109462720

Pianosi, F., Beven, K., Freer, J., Hall, J.W., Rougier, J., Stephenson, D.B., Wagener, T., 2016. Sensitivity analysis of environmental models: A systematic review with practical workflow. Environ. Model. Softw. 79, 214–232.

Price, A.N., Jones, C.N., Hammond, J.C., Zimmer, M.A., Zipper, S.C., 2021. The Drying Regimes of Non-Perennial Rivers and Streams. Geophys. Res. Lett. 48, 1–12. https://doi.org/10.1029/2021GL093298

Putman, A.L., Longley, P.C., Mcdonnell, M.C., Reddy, J., Katoski, M., Miller, O.L., Renée Brooks, J., 2024. Isotopic evaluation of the National Water Model reveals missing agricultural irrigation contributions to streamflow across the western United States. Hydrol. Earth Syst. Sci. 28, 2895–2918. https://doi.org/10.5194/hess-28-2895-2024

Regan, R.S., Juracek, K.E., Hay, L.E., Markstrom, S.L., Viger, R.J., Driscoll, J.M., LaFontaine, J.H., Norton, P.A., 2019. The U. S. Geological Survey National Hydrologic Model infrastructure: Rationale, description, and application of a watershed-scale model for the conterminous United States. Environ. Model. Softw. 111, 192–203. https://doi.org/10.1016/j.envsoft.2018.09.023

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. "Why should i trust you?" Explaining the predictions of any classifier. Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. 1135–1144. https://doi.org/10.1145/2939672.2939778

Rojas, M., Quintero, F., Krajewski, W.F., 2020. Performance of the National Water Model in Iowa Using Independent Observations. J. Am. Water Resour. Assoc. 56, 568–585. https://doi.org/10.1111/1752-1688.12820

Santos, L., Andréassian, V., Sonnenborg, T.O., Lindström, G., de Lavenne, A., Perrin, C., Collet, L., Thirel, G., 2025. Lack of robustness of hydrological models: a large-sample diagnosis and an attempt to identify hydrological and climatic drivers. Hydrol. Earth Syst. Sci. 29, 683–700.

Sarrazin, F., Pianosi, F., Wagener, T., 2016. Global Sensitivity Analysis of environmental models: Convergence and validation. Environ. Model. Softw. 79, 135–152.

Savenije, H.H.G., 2018. HESS Opinions: Linking Darcy's equation to the linear reservoir. Hydrol. Earth Syst. Sci. 22, 1911–1916. https://doi.org/10.5194/hess-22-1911-2018

Shapley, L.S., 1953. A Value for n-Person Games, in: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games (AM-28), Volume II. Princeton University Press, pp. 307–318. https://doi.org/10.1515/9781400881970-018

Slater, L., Blougouras, G., Deng, L., Deng, Q., Ford, E., Hoek van Dijke, A., Huang, F., Jiang, S., Liu, Y., Moulds, S., Schepen, A., Yin, J., Zhang, B., 2025. Challenges and opportunities of ML and explainable AI in hydrology. Under Rev.

Sobol', I., 2001. Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. 55, 271–280.

Song, X., Zhang, J., Zhan, C., Xuan, Y., Ye, M., Xu, C., 2015. Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. J. Hydrol. 523, 739–757. https://doi.org/10.1016/j.jhydrol.2015.02.013

Tijerina, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M.M., Sampson, K., Gochis, D., Maxwell, R., 2021. Continental Hydrologic Intercomparison Project, Phase 1: A Large-Scale Hydrologic Model Comparison Over the Continental United States. Water Resour. Res. 57, 1–27. https://doi.org/10.1029/2020WR028931

Towler, E., Foks, S.S., Dugger, A.L., Dickinson, J.E., Essaid, H.I., Gochis, D., Viger, R.J., Zhang, Y., 2023. Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States. Hydrol. Earth Syst. Sci. 27, 1809–1825. https://doi.org/10.5194/hess-27-1809-2023

Trabucco, A., Zomer, R., 2010. Global High-Resolution Soil-Water Balance [WWW Document]. CGIAR Consort. Spat. Inf. URL https://csidotinfo.wordpress.com. (accessed 1.11.23).

U.S. Geological Survey, 2024. USGS water data for the Nation: U.S. Geological Survey National Water Information System database [WWW Document]. URL https://doi.org/10.5066/F7P55KJN (accessed 12.20.23).

Vrugt, J.A., Beven, K.J., 2018. Embracing equifinality with efficiency: Limits of Acceptability sampling using the DREAM(LOA)algorithm. J. Hydrol. 559, 954–971. https://doi.org/10.1016/j.jhydrol.2018.02.026

Zarnaghsh, A., Husic, A., 2021. Degree of anthropogenic land disturbance controls fluvial sediment hysteresis. Environ. Sci. Technol. 55, 13737–13748. https://doi.org/https://doi.org/10.1021/acs.est.1c00740

Zehe, E., Sivapalan, M., 2009. Threshold behaviour in hydrological systems as (human) geo-ecosystems: Manifestations, controls, implications. Hydrol. Earth Syst. Sci. 13, 1273–1297. https://doi.org/10.5194/hess-13-1273-2009

Zipper, S.C., Hammond, J.C., Shanafield, M., Zimmer, M., Datry, T., Jones, C.N., Kaiser, K.E., Godsey, S.E., Burrows,

770   R.M., Blaszczak, J.R., Busch, M.H., Price, A.N., Boersma, K.S., Ward, A.S., Costigan, K., Allen, G.H., Krabbenhoft, C.A., Dodds, W.K., Mims, M.C., Olden, J.D., Kampf, S.K., Burgin, A.J., Allen, D.C., 2021. Pervasive changes in stream intermittency across the United States. Environ. Res. Lett. 16. https://doi.org/10.1088/1748-9326/ac14ec