JGR Atmospheres

RESEARCH ARTICLE
10.1029/2023JD038534

Comprehensive Analysis of the NOAA National Water Model: A Call for Heterogeneous Formulations and Diagnostic Model Selection

J. Michael Johnson[1,2,3], Shiqi Fang[4], Arumugam Sankarasubramanian[4], Arash Modaresi Rad[1,3], Luciana Kindl da Cunha[3,5], Keith S. Jennings[1,3], Keith C. Clarke[2], Amir Mazrooei[6], and Lilit Yeghiazarian[7]

[1]Lynker, Fort Collins, CO, USA, [2]Department of Geography, University of California, Santa Barbara, Santa Barbara, CA, USA, [3]NOAA/NWS Office of Water Prediction, [4]North Carolina State University, Raleigh, NC, USA, [5]West Consultants, Sacramento, CA, USA, [6]Research Applications Laboratory, National Center for Atmospheric Research, Boulder, NC, USA, [7]University of Cincinnati, Cincinnati, OH, USA

Abstract With an increasing number of continental-scale hydrologic models, the ability to evaluate performance is key to understanding uncertainty and making improvements to the model(s). We hypothesize that any model, running a single set of physics, cannot be "properly" calibrated for the range of hydroclimatic diversity as seen in the contenintal United States. Here, we evaluate the NOAA National Water Model (NWM) version 2.0 historical streamflow record in over 4,200 natural and controlled basins using the Nash-Sutcliffe Efficiency metric decomposed into relative performance, and conditional, and unconditional bias. Each of these is evaluated in the contexts of meteorologic, landscape, and anthropogenic characteristics to better understand where the model does poorly, what potentially causes the poor performance, and what similarities systemically poor performing areas share. The primary objective is to pinpoint traits in places with good/bad performance and low/high bias. NWM relative performance is higher when there is high precipitation, snow coverage (depth and fraction), and barren area. Low relative skill is associated with high potential evapotranspiration, aridity, moisture-and-energy phase correlation, and forest, shrubland, grassland, and imperviousness area. We see less bias in locations with high precipitation, moisture-and-energy phase correlation, barren, and grassland areas and more bias in areas with high aridity, snow coverage/fraction, and urbanization. The insights gained can help identify key hydrological factors underpinning NWM predictive skill; enforce the need for regionalized parameterization and modeling; and help inform heterogenous modeling systems, like the NOAA Next Generation Water Resource Modeling Framework, to enhance ongoing development and evaluation.

Plain Language Summary Water-related issues challenge societies ability to respond to extreme events and plan for the future. Hydrologic models can help better understanding changing water supply and extreme events. To this end, NOAA has implemented a National Water Model (NWM) to forecast the real-time conditions of U.S. waterways and the hydrologic fluxes on the landscape. Here, we evaluate the performance of the NWM version 2.0 streamflow outputs by comparing a 26-year historic simulation to observed data. We diagnose where the model is performing well (and poorly) in the contexts of landscape, climate conditions, and human influence using a large sample basin set. The insights gained identify key factors driving NWM skill and suggest different model formulations are needed in different places. Lastly, we show that understanding why the NWM performs the way it does can help diagnostically select different physics options within the NOAA Next Generation Water Resource Modeling Framework to reduce error in the model output through a more deliberate process representation.

Key Points:

- The relative error and biases in the National Water Model 2.0 streamflow are evaluated in the contexts of categorized basin characteristics
- Aridity, moisture-energy phase correlation, forest and grass cover limit model skill suggesting challenges in modeling evapotranspiration
- Similar understandings can inform regionally heterogeneous models while large biases present opportunity for post-processing model output

Supporting Information:

Supporting Information may be found in the online version of this article.

Correspondence to:

J. M. Johnson,
jjohnson@lynker.com

Citation:

Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., Kindl da Cunha, L., Jennings, K. S., et al. (2023). Comprehensive analysis of the NOAA National Water Model: A call for heterogeneous formulations and diagnostic model selection. *Journal of Geophysical Research: Atmospheres*, *128*, e2023JD038534. https://doi.org/10.1029/2023JD038534

Received 17 JAN 2023
Accepted 3 NOV 2023
Corrected 12 FEB 2024

This article was corrected on 12 FEB 2024. See the end of the full text for details.

Disclaimer: The views expressed in this article do not necessarily represent the views of NOAA or the United States.

Author Contributions:

Conceptualization: J. Michael Johnson, Arumugam Sankarasubramanian, Keith C. Clarke, Amir Mazrooei

1. Introduction

In 2012, the National Academies challenged climate modelers to address an expanding range of scientific problems through more accurate projections of environmental conditions (Bretherton et al., 2012). The hydrologic community has faced a similar challenge with calls for higher resolution forecasts and projections across increasingly large domains (Archfield et al., 2015; Bierkens, 2015; Wood et al., 2011). These forecasts are not only critical for enhanced flood prediction and emergency response (Johnson et al., 2018, 2019, 2022; Maidment, 2016; Salas et al., 2017) but for seasonal supply forecasts that support agriculture, reservoir operations, and commerce

Data curation: J. Michael Johnson, Luciana Kindl da Cunha, Amir Mazrooei
Formal analysis: J. Michael Johnson, Arumugam Sankarasubramanian, Arash Modaresi Rad, Keith S. Jennings
Funding acquisition: Arumugam Sankarasubramanian, Lilit Yeghiazarian
Investigation: J. Michael Johnson, Arash Modaresi Rad
Methodology: J. Michael Johnson, Shiqi Fang, Arumugam Sankarasubramanian, Luciana Kindl da Cunha, Amir Mazrooei
Resources: J. Michael Johnson, Arumugam Sankarasubramanian, Keith C. Clarke
Software: J. Michael Johnson, Luciana Kindl da Cunha
Supervision: Arumugam Sankarasubramanian, Keith C. Clarke
Validation: J. Michael Johnson, Shiqi Fang, Arash Modaresi Rad
Visualization: J. Michael Johnson, Arash Modaresi Rad, Luciana Kindl da Cunha
Writing – original draft: J. Michael Johnson, Arumugam Sankarasubramanian, Keith C. Clarke, Amir Mazrooei
Writing – review & editing: J. Michael Johnson, Shiqi Fang, Arumugam Sankarasubramanian, Arash Modaresi Rad, Luciana Kindl da Cunha, Keith S. Jennings

in the face of global change (Hirabayashi et al., 2013; Mazrooei et al., 2015; Van Loon et al., 2016; Wens et al., 2019).

Traditionally, the hydrologic modeling community has used catchment and land surface models (LSM) to represent the energy and water components of the earth system (Archfield et al., 2015). For example, official streamflow forecasts in the U.S. are issued by the 13 river forecasting centers (RFC) across ~3,600 catchments (Adams, 2016; Burnash, 1995; Salas et al., 2017). To increase spatial coverage, many modeling systems use grid based LSMs to simulate hydrologic and energy fluxes. The ability for LSMs to provide discretized water balance states has long been recognized (Maurer et al., 2001; Nijssen et al., 2001) and many studies have produced reanalysis products and/or evaluated the long-term state of water fluxes in these outputs (Livneh et al., 2013; Maurer et al., 2002; Pekel et al., 2016).

While many land surface models (LSMs) can be used for continental-scale hydrologic modeling, they were historically built to provide land surface boundary conditions in coupled climate models. In that role, LSMs have a stronger focus on closing the energy balance than most catchment models (Archfield et al., 2015). However, large-scale LSMs have two primary limitations for producing accurate hydrologic predictions. The first is that computing fluxes at a grid scale limits the ability to produce river flow in channels without a separate routing model (Li et al., 2016). The second is that when the same models and parameters are applied across the entire domain, location-specific performance tends to degrade. For example, Cai, Yang, Xia, et al. (2014) compared four LSMs across the continental United States (CONUS) using the North American Land Data Assimilation System (NLDAS) test bed (Cai, Yang, Xia, et al., 2014) and in each model, the relative bias in the continental evaluations were larger than those in regional studies (Abdulla et al., 1996; Cai, Yang, David, et al., 2014; Christensen et al., 2004).

Further, as model domains expand, the methods used to evaluate, and synthesis findings become more complex. To date, most hydrologic studies focus on a small number of watersheds to provide comprehensive assessments. These localized insights cannot easily inform general hydrologic concepts across diverse regions (Gupta et al., 2014; Newman et al., 2015). Because of this, there is a fundamental need to facilitate large-sample hydrologic studies with large-sample basin data sets.

In 2016, the NOAA National Weather Service Office of Water Prediction undertook the role of providing reach-level forecasts for the entire U.S. steam network to enhance the authoritative forecasts provided by the RFCs through the National Water Model (NWM). The WRF-Hydro based NWM provides a continental-scale modeling framework that integrates an operational forcing model, a high-resolution land surface model, and high-resolution overland flow, shallow subsurface flow, conceptual baseflow, channel routing, and passive reservoir routing modules. The resolution of each of these components, paired with the geographic extent, make this the only operational model of its class.

Today, the NWM is in its fifth version (v2.2) and some releases include a multi-decade historical simulation (NOAA National Water Model CONUS Retrospective Dataset, n.d.). Versions 1.2 and 2.0 of the historical simulations used the NLDAS/NARR meteorological forcings (Cosgrove et al., 2003, 2020; Mitchell et al., 2004; Mo et al., 2012) while v2.1 used the Analysis of Record for Calibration data set (H. Kim & Villarini, 2022; Kitzmiller et al., 2018). In this study, we evaluate version 2.0 of the NWM, despite the release of a historic simulation associated with version of 2.1, given there is a larger community understanding of the NLDAS forcings.

While the NWM historic simulations lack aspects of the operational model, including data assimilation and reservoir management, the historical products provide an opportunity to better understand where and why the WRF-Hydro implementation of the NWM performs well/poorly to provide guidance on the areas and processes that might be prioritized in ongoing model development.

To date, the NWM has seen several regional and CONUS wide evaluations and model intercomparisons. For example, Salas et al. evaluated an uncalibrated version of WRF-Hydro for the summer of 2015 at 5,700 gauges, providing a benchmark for the evolving hydrology program within the National Weather Service (Salas et al., 2017). Lin et al. evaluated streamflow prediction in Texas, finding that dry regions are strongly affected by a positive bias (Lin et al., 2018). Rojas et al. evaluated NWM v1.0 in Iowa finding performance was linked to the size of the contributing basins with the best performance occurring in basins larger than 10,000 km^2 (Rojas et al., 2020). Other efforts have focused on addressing a range of model intercomparison questions to identify optimal model parametrization, the best performing climate and forcings, and suitable physics formulations (Clark et al., 2015;

Eyring et al., 2016; Kollet et al., 2017). In 2021, Tijerina et al. focused on model biases arising from the simulated streamflow using Flow-CLM (PF-CONUS) version 1.0 and the NWM version 1.2 configuration of WRF-Hydro. Their work highlighted the need for a regionalized modeling framework. Towler et al. (2023) used the National Hydrology Model (NHM) v1.0 and NWM v2.1 to evaluate model performance against a climatological benchmark that incorporated seasonality, spatial patterns, and human influences. Their proposed climatological benchmark offers a framework to screen sites for targeted model application, diagnostics, and development.

Some applications have also focused on using the historical simulations to study issues such as seasonal low flow in the Colorado River basin (Hansen et al., 2019), the one-way surface-groundwater flux in the Northern High Plains Aquifer during extreme flow events (Jachens et al., 2020), operational flood map generation (Johnson et al., 2019), cross section representation (Brackins et al., 2021), and reservoir inflow performance (Viterbo et al., 2020). In the latter, the authors specifically found that NWM inflows in snow-driven basins outperformed those in rain-driven and that basin area, upstream management, and calibrated basin area influenced the ability to reproduce daily reservoir inflows. Together, these studies highlight the utility of the NWM for operations and scientific research, as well as some regional drivers that impact performance. To date, there has been no published, CONUS wide, evaluation of the NWM streamflow outputs over the full 26-year record of simulation provided by v2.0. Given the attention, funding, and mission of the NWM, our first goal is introducing such an evaluation to the literature using a large-sample basin data set.

Looking forward, the NOAA Office of Water Prediction has recognized the limitations of a large scale LSM and acknowledged that improvements from calibration alone are beginning to plateau (Ogden et al., 2021). This phenomenon is not unique to CONUS and the NWM as there is no single best hydrologic model, or model configuration, that can optimize performance across large spatiotemporal domains.

This acknowledgment sparked the NOAA supported Next Generation Water Resource Modeling Framework (NextGen) as a means for running heterogenous model formulations in a single application based on an open source, standards-based, framework (Blodgett & Dornblut, 2018; Blodgett & Johnson, 2022; Ogden et al., 2021; Peckham et al., 2013). The NextGen framework provides an opportunity to regionally configure streamflow generation processes but introduces the questions of (a) what regional traits are currently limiting model skill, (b) what areas of the country most critically need improvement, and (c) what processes (determined by geophysical characteristics) are driving performance and bias. With the increasing advancements of the NextGen Framework, there is a need for a comprehensive understanding, along with methods for identifying, the specific regions and types of processes where performance is suboptimal.

Our hypothesis is that any model, running a single set of physics, cannot be "properly" calibrated for the range of hydroclimatic diversity as seen in the CONUS. However, an evaluation of a model's performance and bias in relation to geospatial catchment characteristics can reveal patterns that speak to a given model formulations strengths and weaknesses across space.

The role of this paper is three-fold. First it introduces a general, interpretable framework for evaluating hydrologic model performance and bias across large basin data sets in relation to catchment characteristics. Second it evaluates the full 26-year NWM v2.0 simulation to help the research community better understand the state of the current NWM across in relation to these basin characteristics. Lastly, it highlights the role NextGen can play in improving model skill, the need for studies like ours to inform the parameterization and selection of heterogeneous models and needed areas of research related to the NextGen framework.

2. Data

This section outlines our basin selection, the streamflow records compared, and the creation of catchment characteristics.

2.1. Gaging Locations and Streamflow Records

Gage locations were selected from the Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II) data set (Falcone, 2011). One of the GAGES-II goals was to identify watersheds with minimally disturbed hydrologic conditions ("reference gages") within 12 major ecoregions. The classification of reference, or natural, basins in the GAGES-II data set goes beyond those in the USGS Hydro-Climatic Data Network, which focused on gages

Table 1
Catchment Characteristics Evaluated in This Study

Name	Source	Description	Range	Units
Area	USGS Gages-II	Drainage Area of Basin	20–19,916	km^2
Mean PPT	NLDAS-2 monthly primary forcing "File A" data	Monthly total mm, summarized to mean Annual Rainfall	14–399	cm
Mean PET	NLDAS-2 monthly primary forcing "File A" data	Monthly total potential evaporation (PEVAP), summarized to mean Annual Potential evaporation	58–313	cm
Mean Aridity	Computed	Mean Annual (PET/PPT)	0–19	Unitless
Mean Correlation	Computed	Mean {cor(PET, PPT)}	−1–1	Unitless
Impervious Percent	NLCD Impervious 2019	Mean Imperviousness	0–57	%
Water	NLCD Landcover 2019	NLCD Class 1	0–100	%
Urban	NLCD Landcover 2019	NLCD Class 2	0–99	%
Barren	NLCD Landcover 2019	NLCD Class 3	0–99	%
Forest	NLCD Landcover 2019	NLCD Class 4	0–99	%
Shrubland	NLCD Landcover 2019	NLCD Class 5	0–95	%
Herbaceous	NLCD Landcover 2019	NLCD Class 7	0–95	%
Agriculture	NLCD Landcover 2019	NLCD Class 8	0–80	%
Wetland	NLCD Landcover 2019	NLCD Class 9	0–15	%
Total Dams	USACE NID	Number of dams in drainage basin	1–2,040	Count
Total Storage	USACE NID	Sum of maximum storage of all dams in drainage basin	0–149	km^3
Snow Depth Mean	NLDAS VIC Land Surface Model L4 Hourly 0.125° × 0.125° V002	Snowfall (frozen precipitation) (kg/m^2)	0–89	cm
Snow Fraction	NLDAS VIC Land Surface Model L4 Hourly 0.125° × 0.125° V002	Mean Annual Snow Cover Fraction	0–69	%

that experienced natural flow regimes at some point in the past (Slack et al., 1993). The USGS site IDs associated with these gages were used to collect daily streamflow data from the National Water Information System (NWIS) using the dataRetrieval R package (De Cicco et al., 2018) and only those with at least 10 years of daily observed flow between 1 January 1993 and 31 December 2018, a total drainage area between 20 and 20,000 km^2, and that were completely within CONUS were retained. Figure S1 in Supporting Information S1 shows the locations of the controlled and natural basins overlayed on a map of 26-year mean aridity index (AI) values in CONUS.

The historical record for NWM v2.0 is approximately 40 TB in size, 10 TB of which is the channel point files containing streamflow. Johnson et al. (2023) restructured this data set to support broad scale applications and the data are accessible through the nwmTools R package (Johnson & Blodgett, 2020; Johnson et al., 2023). Hourly records were summarized to daily averages to remain consistent with the NWIS observations, and, in total, 4,236 basins are available for analysis with natural basins making up ∼21% of the data set.

2.2. Basin Characteristics

All physical and machine learning models rely on accurate geospatial data to discretize and parameterize the models. High-quality data sets are essential for hydrological modeling and evaluation. The utility of the catchment characteristics includes but is not limited to categorizing performance, building statistical and data-driven models (Kratzert et al., 2019), regionalizing parameters from gauged to ungauged basins (Guo et al., 2021), informing modeling efforts focusing on the dominant hydrological processes for each landscape and hydroclimate (Jehn et al., 2020), better understanding hydrological organization, scaling, and similarity (Peters-Lidard et al., 2017), and providing an additional tool to guarantee that the "right answers" are being obtained for the "right reasons" (Kirchner, 2006). Here, we define and construct a set of landscape characteristics to help characterize NWM performance. Table 1 identifies the catchment characteristics tested and their source, description, range, and units.

2.2.1. Landscape Characteristics

Noah-MP is a spatially distributed LSM with multiple options for land-atmosphere interaction processes (Niu et al., 2011). To determine parameter values for specific computational elements, the model relies heavily on land cover and soil inputs. In total, 49 parameters are assigned based off the land cover assigned to a cell using the MPTABLE (Barlage, 2017). Noted limitations of this lookup approach are that all pixels with the same vegetation have the same parameters, across space and time (except for two cases of climate Seasonality and Asynchrony Index (SAI) and Leaf Area Index (LAI) (Barlage, 2017; D. H. D. Kim et al., 2023). To explore the impacts of land cover on model performance, the percentage of each Anderson level 1 land cover class (9 in total) from the 2019 National Land Cover Data set (NLCD) was determined (Anderson, 1976; Homer et al., n.d.; L. Yang et al., 2018). The total impervious surface was also determined from the 2019 NLCD Impervious data product.

2.2.2. Meteorological Characteristics

Following Lin's et al. (2018) analysis of the NWM in Texas, Cai, Yang, Xia's, et al. (2014) broad evaluation of LSMs, and Peterson's et al. (2012) evaluation of LSM's, we identified several energy and moisture flux variables that could influence model performance. These include monthly potential evaporation (PET; $\frac{kg}{m^2}$), precipitation (PPT; $\frac{kg}{m^2}$), Aridtiy Index (AI), moisture-and-energy phase correlation, mean snow depth, and mean snow coverage fraction. PET and PPT were obtained from the primary forcing data of NLDAS-2 for January 1993 through December 2018. For each basin the mean monthly PET and PPT were summarized over the basin area using a method that weighted partially covered grid cells by the percentage of containment. AI was calculated as the ratio of annual mean PPT to annual mean PET $\left(\frac{PET}{PPT}\right)$ to help categorize basins as energy- or moisture-limited, where an AI < 0.3 is humid, an AI between 0.3 and 1 is semi-humid, between 1 and 2 temperate, between 2 and 3 semi-arid, and greater than 3 arid.

The covariability between the monthly cycles of moisture and energy is estimated by the correlation between monthly PPT and PET (ρ(PPT,PET)) (Abdulla & Lettenmaier, 1997; Sankarasubramanian & Vogel, 2002). These values range from -1 to $+1$ and when covariability is greater than -0.4 or less than $+0.4$ there is evidence the precipitation and temperature cycles are out-of-phase (Petersen et al., 2012). The Spearman correlation coefficient was determined for each NLDAS cell using the mean monthly PET and PPT over the 26 years. From this, a mean value was determined for each basin. Overall, this term expresses the correlation between the precipitation and PET, or the moisture-and-energy phase correlation. From here on out, we refer to this term as the phase correlation (see Figure S2 in Supporting Information S1 for more information).

Lastly, snow cover fraction and Water Equivalent of Accumulated Snow Depth (WEASD; kg/m^2) were taken from the NLDAS-2 Noah Land Surface Model L4 Hourly 0.125° × 0.125° V002 outputs and summarized to a mean annual basin value.

2.2.3. Anthropogenic Characteristics

The anthropogenic influence in each basin is approximated by counting the number of 2019 United States Army Corp of Engineers National Inventory of Dams (USACE NID; National Inventory of Dams, 2019) in each basin as well as the cumulative storage (NID_STORAGE). In total, 3,970 of the 91,457 dams (4.34%) in the USA have either 0 or "NA" storage reported. In these cases, these dams did not contribute to the total storage, but were included in the total dam count.

3. Methods

3.1. Goodness of Fit Metrics

To assess model performance, we focus on how well the NWMMv2.0 simulations capture the observed USGS streamflow at a daily timescale. To do this, the Nash-Sutcliffe Efficiency (NSE) was calculated for each location (Equation 1; Nash & Sutcliffe, 1970).

$$\mathrm{NSE} = 1 - \frac{\sum_{t=1}^{T} \left(Q_m^t - Q_o^t\right)^2}{\sum_{t=1}^{T} \left(Q_o^t - \mathrm{mean}\, Q_o^t\right)^2} \quad (1)$$

where Q_o is the observed and Q_m is the modeled streamflow, both at time (t).

An NSE of 1.0 represents perfect agreement between the modeled and observed values and an NSE of 0.0 occurs when the modeled error variance is equal to the observed variance from the mean. NSE can become negative when the error variance in the modeled record is greater than in the observed record, suggesting the observed mean would be a better predictor than the model. Here we chose NSE decomposition over widely used metrics (i.e., Kling–Gupta efficiency) as the error decomposition empowers useful and insightful diagnostics.

Subjective NSE thresholds have been suggested by several authors (Criss & Winston, 2008; McCuen et al., 2006; Moriasi et al., 2007; Ritter & Muñoz-Carpena, 2013) and we adopt those used for categorizing performance on monthly time steps (as there are none for daily steps), stating a NSE greater than 0.75 is "very good," a NSE between 0.65 and 0.75 is "good," an NSE between 0.5 and 0.65 is "satisfactory," and those less than 0.5 are "unsatisfactory" (Moriasi et al., 2007). These could be considered too strict for the daily evaluation performed here, but, they provide a general qualitative categorization.

With more than 4,000 sites being evaluated, the lower NSE limit of $-\infty$ can be problematic and in these cases, a Normalized NSE (NNSE) rescaled to the range of $\{0,1\}$ is computed (Equation 2; Nossent & Bauwens, 2012).

$$\text{NNSE} = \frac{1}{2 - \text{NSE}} \tag{2}$$

With this transformation, values of 1 are still interpreted as a perfect fit and values <0.5 represent cases where the NSE is less than 0 and the mean of the observed data is better than the model.

To further support evaluation, NSE can be decomposed into components representing the overall agreement of the model (A term), as well as conditional (B term) and unconditional (C term) bias making it easier to determine how different types of error are interrelated and what might cause a particular model—or location—to perform well or poorly (Murphy, 1988; Weglarczyk, 1998) (Equations 3–6). This disaggregation is shown in Equations 3–6.

$$\text{NSE} = \text{A} - \text{B} - \text{C} \tag{3}$$

$$\text{A} = r^2 \tag{4}$$

$$\text{B} = \left(r - \frac{\sigma_s}{\sigma_o}\right)^2 \tag{5}$$

$$\text{C} = \left(\frac{(\mu_s - \mu_o)}{\sigma_o}\right)^2 \tag{6}$$

where r is the Pearson correlation coefficient; σ_o is the standard deviation of the observed flows; σ_s is the standard deviation of the simulated flows; μ_o is the mean of the observed flows; and μ_s is the mean of the simulated flows. The relationship among A, B, and C is illustrated in Figure 1.

3.2. Analysis of Variance ANOVA (Type II)

We used a series of analysis of variance (ANOVA) tests to find statistically significant catchment characteristics for accurately predicting streamflow. The principal test for ANOVA is the F statistic which is the ratio of variance caused by a treatment compared to the variance due to random chance. The ANOVA test assumes independence of observations, absence of significant outliers, data normality, and homogeneity of variances. The p-value associated with the F statistic can be used to tell if there is a statistically significant difference between the categorical groups and the probability of getting a result at least as extreme assuming there is no difference in means.

In practice, a small p-value does not always translate to a practical significance and should be considered alongside the effect size which represents the magnitude of the difference between groups (Sullivan & Feinn, 2012). While a p-value can determine if an effect exists, it will not reveal the size of the effect. Thus, gaging both practical (effect size) and statistical significance (p-value) is essential. The effect size reported here is the η^2 squared.

$$\eta^2 = \frac{\text{SS}_{\text{effect}}}{\text{SS}_{\text{total}}} \tag{7}$$

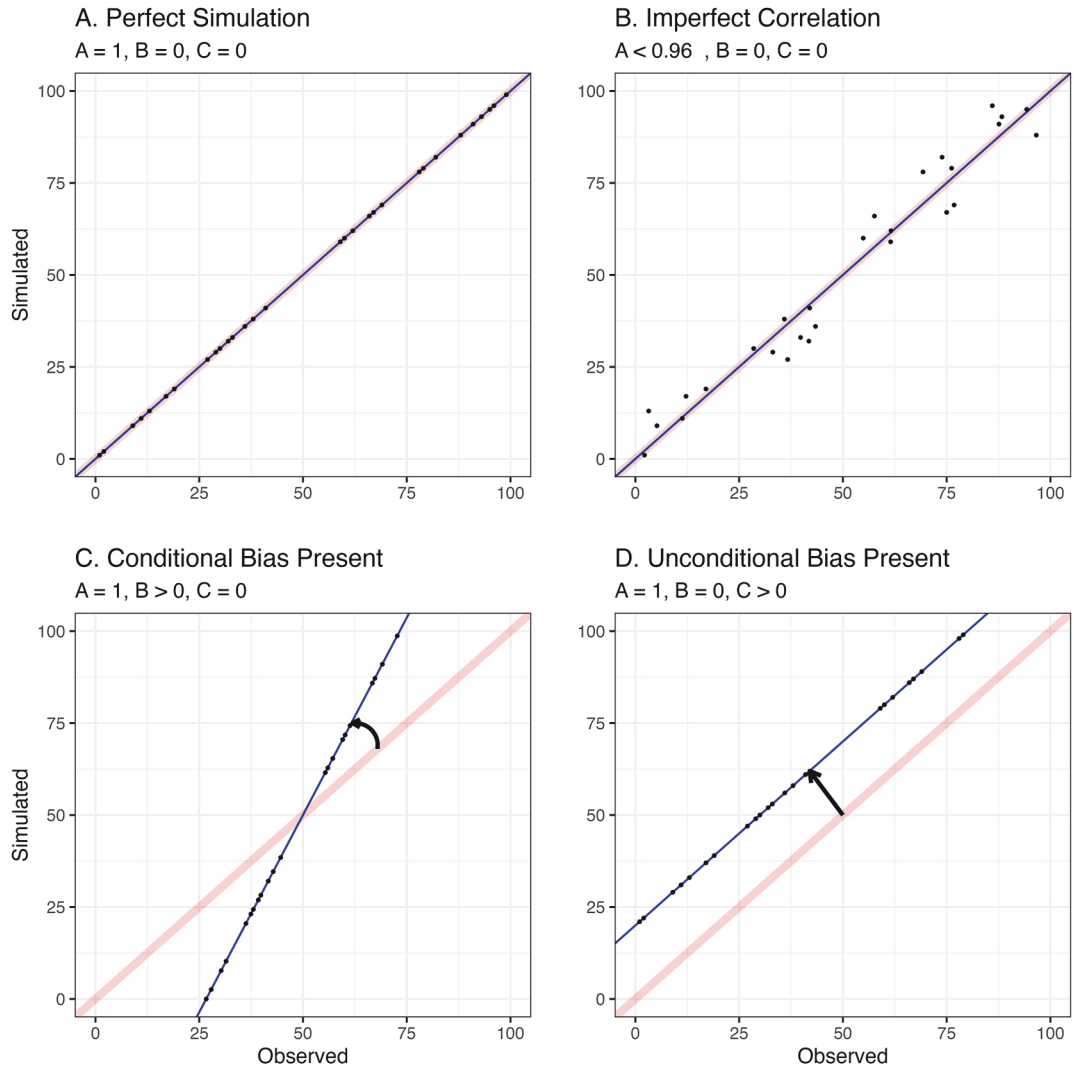

Figure 1. Conceptual diagram illustrating how Nash-Sutcliffe Efficiency (NSE)-A, B, and C appear in a scatter plot of observed versus simulated flows. The pink trend line represents a perfect simulation while blue trend line represents a simulation with error and bias. Panel a shows a perfect simulation where A = 1 and there is no bias (B = C = 0). Panel b shows an example where there is no bias (B = C = 0) and high, but imperfect correlation (A < 1). Panel c shows the presence of conditional bias illustrated by the rotation of the regression line around the **1**:**1** plot center, thus B > 0. Panel d shows the presence of unconditional bias represented by the offset of the hypothetical regression line from a 1:1 line (C > 0).

where SS_{effect} is the sum of squares of an effect for one variable and SS_{total} is the total sum of squares in the ANOVA model.

The value for η^2 can range from 0 to 1 and describes the proportion of variance that can be explained by a given variable in the model after accounting for variance explained by other variables in the model. A general baseline for interpreting η^2 states that (Cohen, 2013):

- $\eta^2 > 0.01$ indicates a small effect
- $\eta^2 > 0.06$ indicates a medium effect
- $\eta^2 > 0.14$ indicates a large effect

For our tests, we run independent ANOVA tests for each catchment characteristic in Table 1, on each NSE component, for natural basins and controlled basins (18 characteristics, 3 NSE metrics, 2 groups = 78 tests). Also, it should be noted that since multicollinearity is common in earth system models, variance inflation can occur, and one can expect the total explained variance from all variables to exceed 1.

Since all predictor variables are continuous, and ANOVA is based on categorical groupings, we use a Jenks natural break classification to identify natural groupings within the complete set of data. Jenks natural breaks is a clustering method to determine a predefined number of groups that minimize each group's average deviation from the group mean, while maximizing each groups mean deviation from the mean of other classes. For each characteristic, we started with four natural classes; however, in cases where natural groups were formed that resulted in any group having less than 10% of the overall population, we decreased the number of classes. In some cases, there are literature-driven values that we use in lieu of these clusters. For example, the classification for aridity and the Peterson 2012 classification for phase correlation are used.

We chose ANOVA over other statistical methods (e.g., regression) to identify the traits of locations with good/bad performance and high/low bias. If we understand where the model does poorly, what causes the poor performance, and what similarities systemically poor performing areas have, we can better understand the model and appropriately apply its output. Future work can use this understanding to revise the formulations and parametrizations—particularly with the advent of the NextGen system, and other efforts could seek to build on this to provide regression-based post-processing or error diagnostics. In each of these cases, understanding the most influential characteristics will be an advantageous start.

4. Results

4.1. NNSE

To understand the variability in the NWM performance, the NNSE results are visualized in Figure 2.

Figure 2a maps the NWIS gauges, split by controlled and natural categories, and colored by NNSE. On the left, the control basins show strong performance in the northeast, east, and south but exhibit weak performance west of the 100th meridian. The exception to this is along the western side of the Sierra Nevada Range where the AI is lower than the west region at large. In the controlled basins there is a qualitative impact of cities on NWM performance, with low skill surrounds the Orlando, Charlotte, New York, Detroit, Chicago, and Nashville metropolitan areas in an otherwise well-performing east. In the humid west, the San Francisco Bay Area and Portland also underperform compared to their surroundings.

The natural basins demonstrate a more consistent performance east of the 100th meridian, however, performance begins to degrade west of this line. The extent of relative performance loss is less than in controlled basins.

In all basins, there are clear systematic drops in performance between the 105th and 95th meridians (see Figure 2b). When focusing on the controlled basins, the 50th percentile of locations achieves "satisfactory" performance, while west of the 95th meridian, the 75th percentile drops well below this mark. Not only does performance drop, but the variability increases as evidenced by the spread between the 25th and 75th percentiles. There is a slight recovery in performance starting around the 115th meridian, however variability remains high.

In examining the natural basins, the 75th percentile shows "satisfactory" performance, until the 100th meridian; however, the spread in variation is not as large as in controlled basins. West of the 105th meridian, the spread in variability increases, but to a lower level than in the controlled basins.

Figure 2c illustrates the empirical cumulative distribution function of NNSE grouped by basin and aridity classification. In this plot, the ideal curve would stay as low as possible on the *y*-axis for as long as possible along the *x*-axis. Humid basins outperform arid basins; natural basins outperform controlled basins; and the difference between controlled and natural classification is more noticeable in the humid basins. More than 55% of the controlled humid basins achieve "satisfactory" or better performance with over 75% of the natural humid basins meeting this goal. In the arid regions, approximately 85% of the basins (regardless of classification) exhibit "unsatisfactory" performance. Among those with "satisfactory" or better performance, the distinction between natural and controlled is non-existent.

4.2. NSE-A: Relative Performance

NSE-A represents the coefficient of determination between observed and simulated streamflow values. NSE-A values are mapped in Figure 3a while Figure 3b plots the 25th, 50th, and 75th percentile NSE-A, grouped by whole-degree longitude bands smoothed with a 5° rolling mean. Concerning NSE-A, the NWM performs better in the eastern part of the CONUS and along the west coast. The variability in NSE-A is greater in the west than in

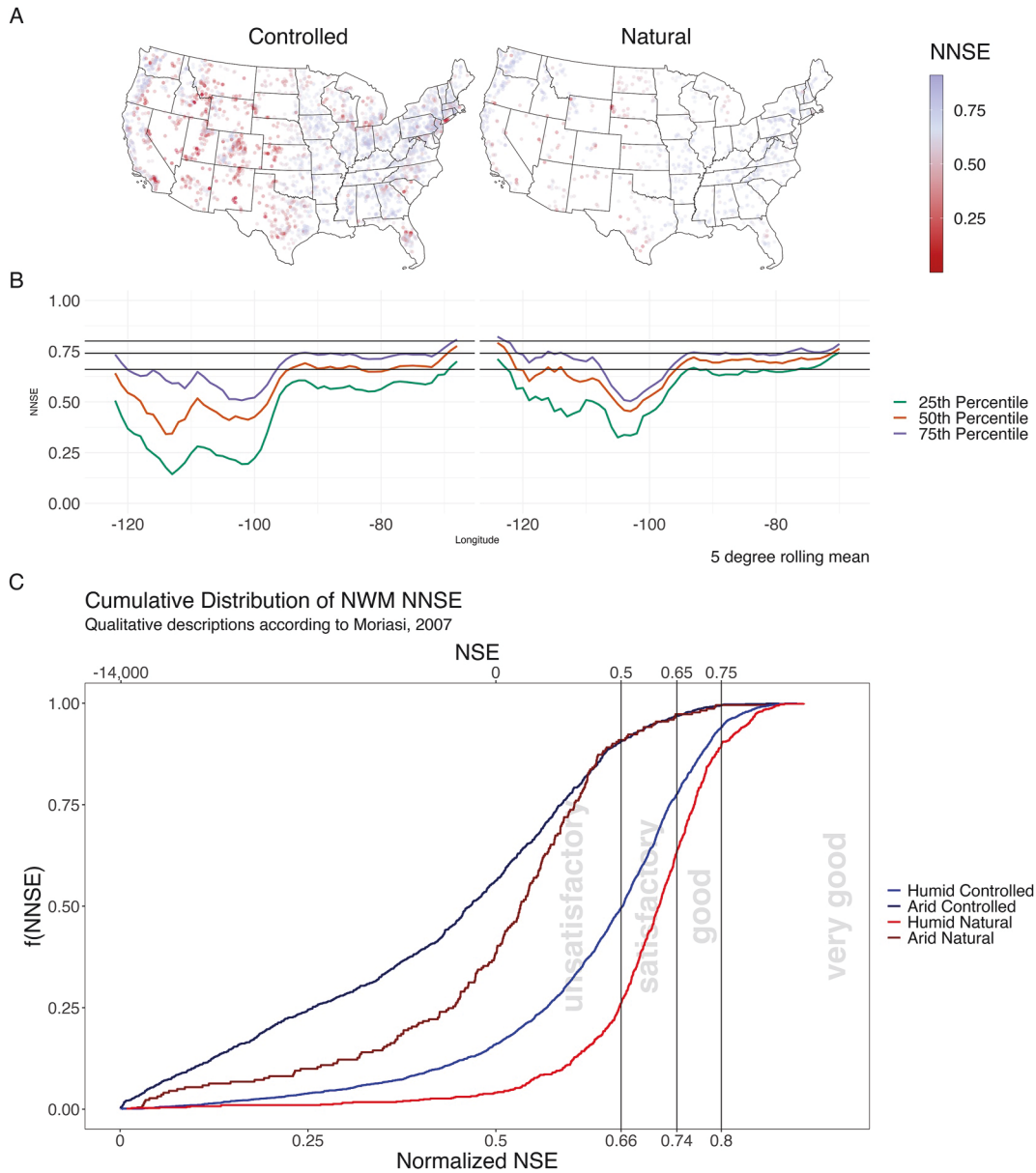

Figure 2. (a) Normalized Nash-Sutcliffe Efficiency (NNSE) mapped by gage location, the midway color aligns with "Satisfactory" performance. (b) Shows the 25th, 50th, and 75th percentile NNSE for each band of longitude smoothed with a 5° rolling mean. The vertical lines at NNSE = 0.66, 0.74, and 0.80 represent the categories of "Unsatisfactory," "Satisfactory," "Good," and "Very Good." (c) NNSE cumulative distributions grouped by aridity and Geospatial Attributes of Gages for Evaluating Streamflow classification. The vertical lines represent the same qualitative groupings as panel b. Here red curves represent arid basins (aridity index [AI] > 2), and blue curves represent humid basins (AI < 2).

the east, with the exception of natural basins in the humid west coast. Across the CONUS, the variation in NSE-A in controlled basins is greater than in natural basins however, the pattern in the longitudinal profiles are largely the same. Using the catchment characteristics identified in Table 1, a series of ANOVA tests were conducted to examine the effects of each characteristic on NSE-A in natural and controlled basins. Only those tests that yielded a statistically ($p < 0.05$) and practically ($\eta^2 > 0.01$) significant result are shown in Figure 3c. In each of these panels, a horizontal line is used to mark the mean NSE-A across all basins.

4.2.1. Meteorological Characteristics

We found a significant relationship between AI and NSE-A indicating that basins with lower values of AI had higher NSE-A (Figure 3ca). The effect size suggests that 45% of the variance in NSE-A can be explained

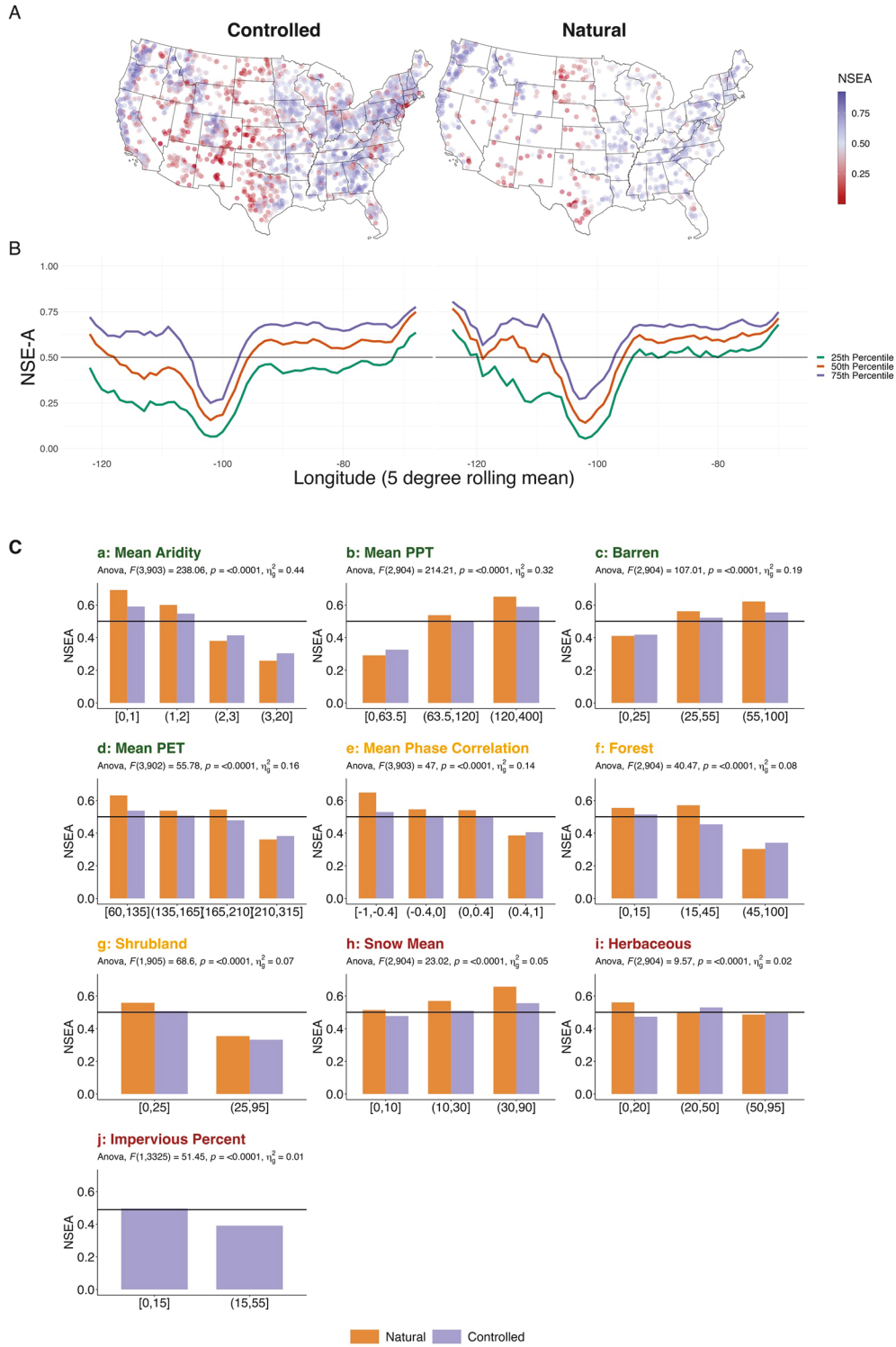

Figure 3. (a) Nash-Sutcliffe Efficiency (NSE)-A split by natural and controlled basins. (b) 25th, 50th, and 75th percentile NSE-A for each band of longitude smoothed with a 5° rolling mean. (c) Mean NSE-A is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically ($p > 0.05$) and practically ($\eta^2 > 0.01$) significant are shown. Plots are ordered according to effect size and titles are colored according to Cohen's effect size classification where green is a large effect size, orange a medium and red a small. A high value on the *y*-axis indicates better model performance. The black horizontal line across all plots is the mean NSE-A across all basins.

by the aridity of a basin. In both controlled and natural basins NSE-A decreases by a factor of 2 when comparing very humid to very arid basins. The second and fourth largest effect sizes belong to PPT (Figure 3cb) and PET (Figure 3cd), respectively. Naturally, these are highly correlated with aridity, however, evaluating them independently shows that as there is more precipitation, the NWM can better predict streamflow. The contrast between dry and wet basins is slightly lower in controlled basins.

Unlike PPT and aridity, which show a steady pattern across groupings, the middle two sections of PET hover around the mean NSE-A. This suggests that only "extreme" low PET or "extreme" high PET have a high impact on performance. In all but basins with very high PET, natural basins perform better than controlled basins. In natural basins, out of phase moisture and energy correspond to better performance. In both controlled and natural basins, inphase moisture and energy produce worse performance. With respect to overall variance in NSE-A, PPT explains 31%, PET 18%, and mean phase correlation 13%. Lastly, as mean snow coverage (Figure 3ch) increases, so does general NSE-A performance.

When mean annual snow depth is between 0 and 10 cm, the NSE-A across all basins is near the overall mean. As snow depth increases, relative performance improves, particularly in natural basins. In a broad sense, across groupings, more PPT and snow increase model performance, while more PET, aridity, and phase correlation decrease model performance. Of course, some of these factors are correlated; for example, snowy basins are generally not arid.

4.2.2. Landscape Characteristics

We found a significant relationship between barren land and NSE-A, as higher percentages of barren land had higher NSE-A (see Figure 3cc). This relationship is particularly evident in natural basins and the effect size of 20% highlights the significance of barren land. Imperviousness percentage (Figure 3cj) has the opposite effect and is only significant in controlled basins (as expected). When imperviousness is <15%, basins perform at the expected NSE-A mean; however, when more than 15% of the basin is impervious, the observed NSE-A performance worsens.

The opposite is observed for forest (Figure 3cf) and shrubland (Figure 3cg). As each of these increases, NSE-A decreases. Vegetative classes (forests and shrublands) possess significant biomass that respond differently based on season and location. These changes impact both PET and actual ET which impacts model performance. In this case, we found that increased vegetation coverage in a basin corresponds to lower NSE-A. This effect is exacerbated in models where the same parameters (e.g., LAI) are applied to different hydroclimate regions (e.g., Arizona and Maine) (Johnson & Clarke, 2021). This pattern is also evident in the herbaceous land cover (grasslands, Figure 3ci); however, the effect is smaller, and the pattern differs considerably when comparing controlled and natural basins.

Overall, 10 characteristics were statistically and practically significant in describing the variation in relative performance. Among these, Aridity, PPT, PET, phase correlation (meteorological factors), barren, forest, and shrubland (landscape features) demonstrated a medium or strong relationship with NWM NSE-A.

4.3. NSE-B: Conditional Bias

When comparing the NNSE and NSE-A longitudinal plots (Figures 2b and 3b), NSE-A exhibits a U-shaped pattern, indicating model performance recovery west of the 100th meridian, while the NNSE plots so not show the same recovery. This suggests there are structured biases in the model—particularly in the west—that yield poor overall performance, despite relatively high NSE-A (e.g., Equation 3).

Figure 4 maps NSE-B for the natural and controlled basins. Here, NSE-B values are truncated to 1.0 for visualization purposes, meaning anything listed as 1.0 is ≥1.0. The number of dropped sites is listed in the subtitle of each plot. Beneath each map is a longitudinal average smoothed with a 5° rolling mean, developed in the same manner as Section 4.1.

Larger NSE-B values are observed in the arid west, and the longitudinal percentile plots indicate the amount and variability of NSE-B is nearly zero in natural basins east of the 100th meridian and less than 0.15 in controlled basins. In all basins, NSE-B spikes between the 95th and 105th meridians. Natural basins show model recovery (less conditional bias) west of the 110th meridian (the Rocky Mountains). In contrast, the controlled basins do

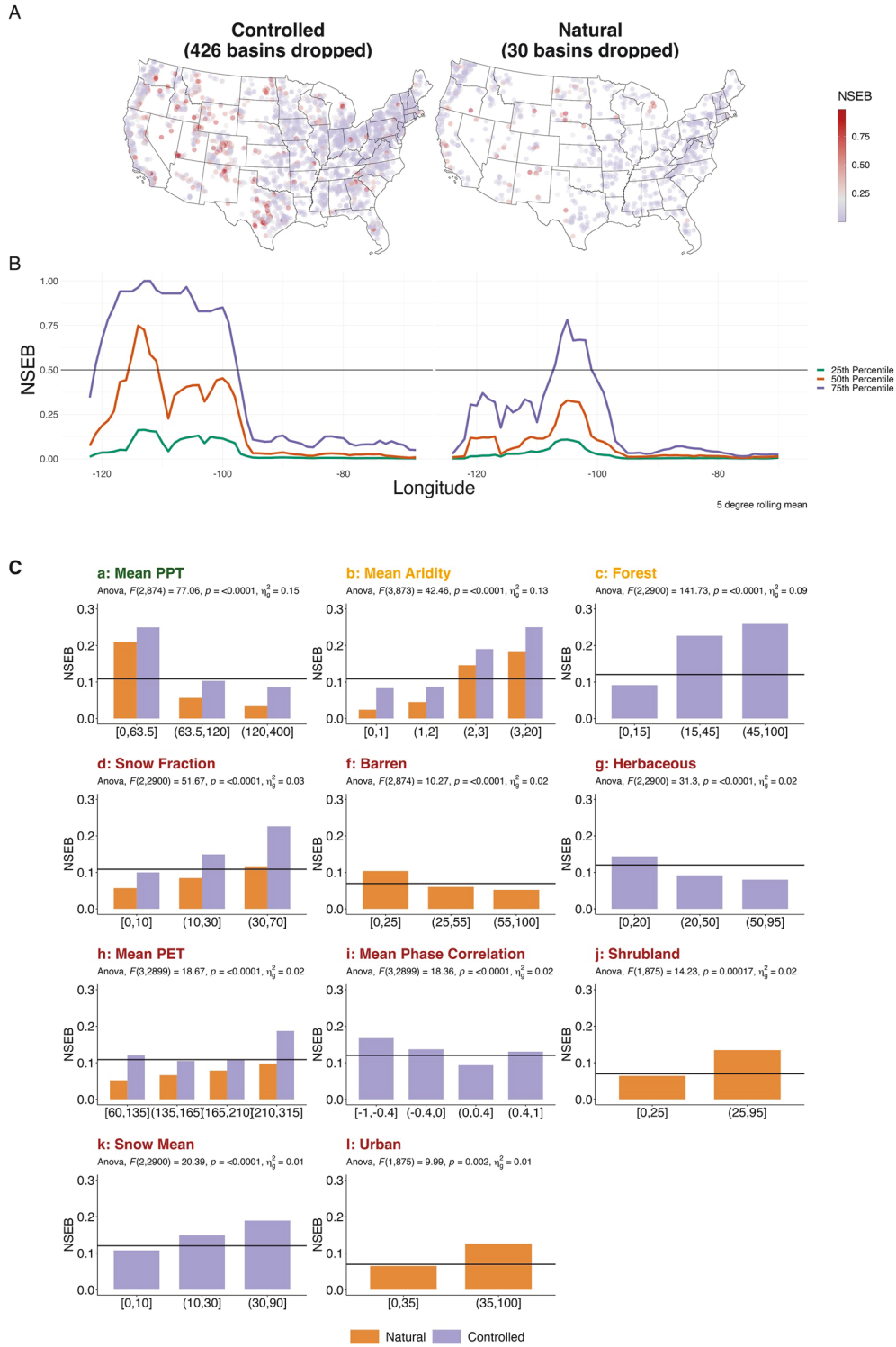

Figure 4. (a) Nash-Sutcliffe Efficiency (NSE)-B split by natural and controlled basins. (b) 25th, 50th, and 75th percentile NSE-B for each band of longitude smoothed with a 5° rolling mean. (c) Mean NSE-B is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically ($p > 0.05$) and practically ($\eta^2 > 0.01$) significant are shown. Plots are ordered according to effect size and plot titles are colored according to Cohen's effect size classification where green is a large effect size, orange a medium and red a small. The black horizontal line across all plots is the mean NSE-B across all basins.

not recover—and even increase—until the humid west coast is reached. In all basins, variability and conditional bias are larger in controlled basins. In the controlled basins, the influence of large cities is evident with deep red clusters occurring around Tampa, Atlanta, Columbus, Milwaukee, Denver, San Antonio, Salt Lake, Reno, and Missoula among others. While not strictly a quantitative analysis, this high conditional bias near urban centers should caution the application of NWM in high-flow forecasting and underscores the need to better represent urban, non-riverine hydrology. Figure 4c is arranged in the same way as Figure 3c with the exception that a low value on the *y*-axis is desirable, indicating minimal conditional bias. Again, only statistically significant characteristics are shown in the plots. Across the board, conditional bias is lower in the natural basins, but all basins demonstrate the same patterns.

4.3.1. Meteorological Characteristics

Starting with Figures 4ca and 4cb, dry (PPT < 63.5 cm), arid (AI > 3) basins have larger than average conditional bias while wet (PPT > 12 cm), humid (AI < 2) basins exhibit less than average conditional bias. The effect of PET (Figure 4cg) is only significant in controlled basins when PET exceeds 210 cm/year almost doubling the average conditional bias. Conversely, mean phase correlation is significant in basins that are notably out of phase (<−0.4) and conditional bias increases by a factor of 1.5. Overall PPT, AI, PET, and phase correlation explain 15%, 12%, 2%, and 2% of the variance in conditional bias, respectively.

While higher average snow depth is related to higher NSE-A for all basins, it results in greater NSE-B in controlled basins highlighting the challenges of modeling diverse snow processes (Figure 4cj). This could also be a product of the primary functions of local reservoirs as those in snowy basins may be designed to store runoff and snowmelt for the dry season. Snow Fraction (Figure 4cd) also influences NSE-B suggesting that as more of a basin is covered with snow more conditional bias can be expected. There is a difference between the natural and controlled basins in that even at high levels of snow coverage, natural basins exhibit average conditional bias. In contrast, conditional bias increases dramatically as snow coverage increases in controlled basins.

4.3.2. Landscape Characteristics

Forest and herbaceous land covers are the only significant types with respect to NSE-B in controlled basins (Figure 4cc). When forest coverage is <15%, conditional bias remains near the overall average. However, when coverage exceeds 15%, conditional bias grows by a factor of 2.5. While significant, the influence of barren land is less pronounced than the other factors present in Figure 4c. Barren land is only significant in natural basins and there is less conditional bias with more barren coverage. A nearly identical pattern exists for herbaceous coverage, except its influence is significant in controlled basins. Shrub and urban landscapes are significant in natural basins and when they exceed 25% and 35% of the basin respectively, they lead to an almost 1.5 times increase in conditional bias. While the idea that urban land cover influences natural basins is at first counter to our expectations, the takeaway is that when urbanization appears in what's deemed a natural basin, its impact is high. In basins already deemed controlled, the presence of urban land cover is not a significant factor. In cases where large urbanization occurs in natural basins, we can assume the basin to be erroneously classified, or, that the basin has been urbanized post GAGES-II creation.

Overall, 11 characteristics were statistically and practically significant in describing the variation in conditional bias. Among these, PPT and Aridity were meteorological factors with a medium or larger effect size while forest was the only landcover with a medium or larger effect size.

4.4. NSE-C: Unconditional Bias

Figure 5a maps NSE-C (unconditional bias) for the natural and controlled basins where NSE-C values are truncated to 1.0, meaning anything listed as 1.0 is ≥1.0. The number of truncated sites is listed in the subtitle of each plot. Beneath each map is a longitudinal average smoothed with a 5° rolling mean developed in the same way as Section 4.1. Figure 5c is arranged in the same way as Figure 4c. Across the board, bias in the natural basins is lower than in controlled basis and land cover impacts are more prominent in controlled basins while meteorological properties influence all basins. When compared to the population mean (the horizontal bar in each panel in Figure 5c), natural basins exhibit significantly less unconditional bias than the controlled basins.

4.4.1. Meteorological Characteristics

Higher values of aridity (Figure 5cc) and snow fraction (Figure 5cd) tend to have higher values of NSE-C. Equally as PPT (Figure 5cb) and phase correlation (Figure 5cj; only in controlled basins) increase, NSE-C values

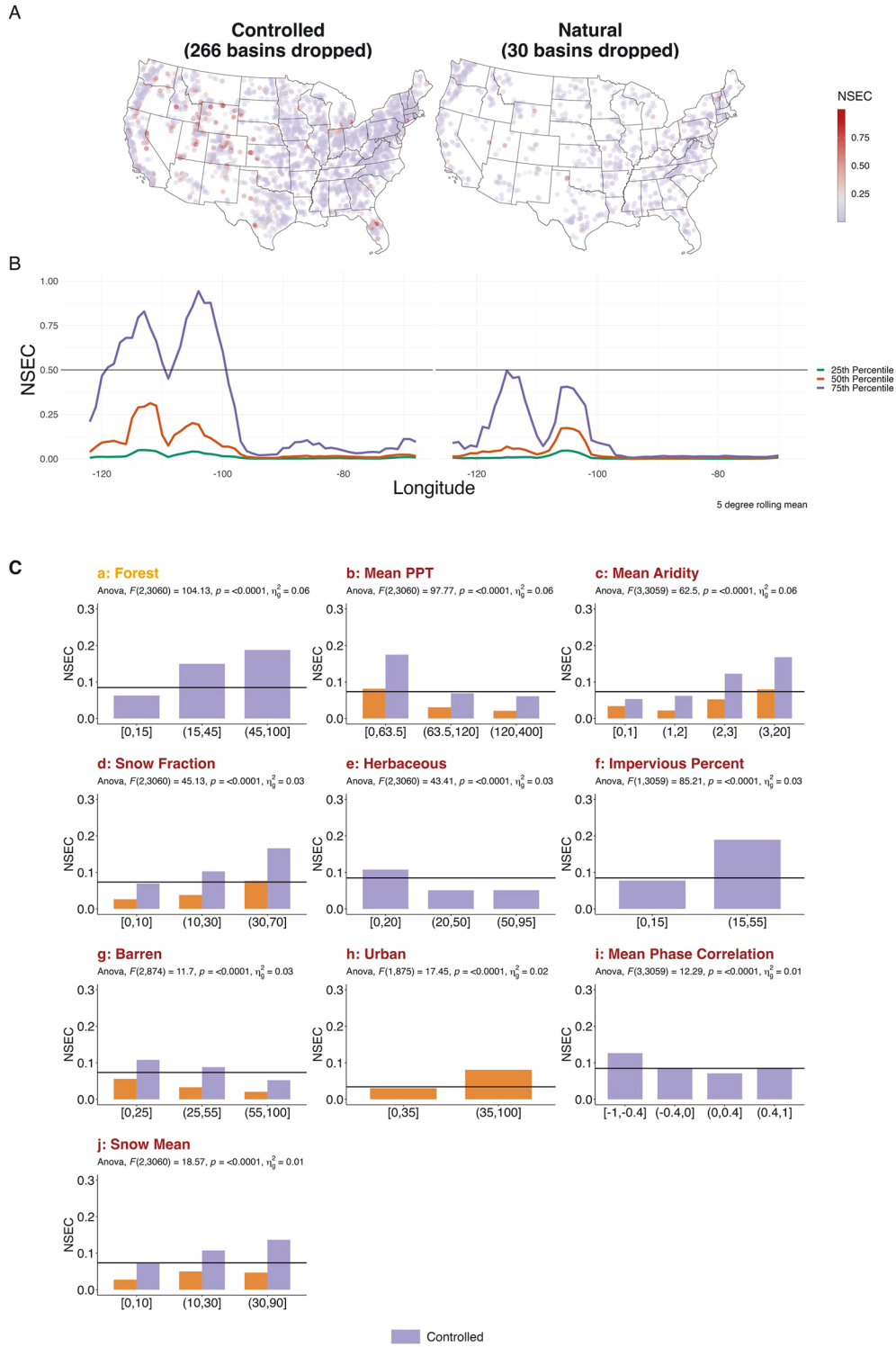

Figure 5. (a) Nash-Sutcliffe Efficiency (NSE)-C split by natural and controlled basins. (b) 25th, 50th, and 75th percentile NSE-C for each band of longitude smoothed with a 5° rolling mean. (c) Mean NSE-C is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically ($p > 0.05$) and practically ($\eta^2 > 0.01$) significant are shown. Plots are ordered according to effect size and titles are colored according to Cohen's effect size classification where green is a large effect size, orange a medium and red a small. The black horizontal line across all plots is the mean NSE-C across all basins.

Table 2
Significant Catchment Characteristics and Their Impact on Model Performance and Bias

	Variable	NSE-A	NSE-B	NSE-C	As "variable" increases, NWM…
Meteorological	PPT	*Controlled, natural*	*Controlled, natural*	*Controlled, natural*	Performance increases & bias decreases in all basins
	PET	*Controlled, natural*	*Controlled*		Performance decreases in all basins & bias increases in controlled basins
	Aridity	*Controlled, natural*	*Controlled, natural*	*Controlled, natural*	Performance decreases & bias increases in all basins
	Phase Correlation	*Controlled, natural*	*Controlled*	*Controlled*	Performance decreases in all basins & bias decreases in controlled basins
	Snow Coverage	*Controlled, natural*	*Controlled*	*Controlled, natural*	Performance increases & bias increases in all basins
	Snow Fraction		*Controlled, natural*	*Controlled, natural*	Bias increases in all basins
Landscape	Impervious Percent	*Controlled*		*Controlled*	Performance decreases & bias increases in controlled basins
	Urban		*Natural*	*Natural*	Bias increases in natural basins
	Barren	*Controlled, natural*	*Natural*	*Controlled, natural*	Performance increases & bias decreases in all basins
	Forest	*Controlled, natural*	*Controlled*	*Controlled*	Performance decreases in all basins & bias increases in controlled basins
	Shrubland	*Controlled, natural*	*Natural*		Performance decreases in all basins & bias increases in natural basins
	Herbaceous	*Controlled, natural*	*Controlled*	*Controlled*	Performance decreases in all basins & bias decreases in controlled basins

Note. In each cell, the impacted basin class is listed assuming the variable is increasing. Green colors indicate improvement, while red cells show degradation. The last column summarizes the overall effect in plain language.

are lower. In all cases, the worst-performing category (e.g., low PPT) of natural basins results in unconditional bias near the population average which then improves in the respective direction of the characteristic. In contrast, when looking at controlled basins, the best-performing category (e.g., high PPT) is generally near the population average while unconditional bias exponentially increases when moving away from the best-performing category. The exception to this pattern is mean snow coverage (Figure 5ci) where unconditional bias in controlled basins increases across groupings and remains nearly level for natural basins. The large takeaway is that when looking at the unstructured bias in the NWM, the bulk of it is in controlled basins where moisture and energy cycles are out of sync, and exhibit low PPT, high aridity, and/or high snow quantities (both mean and fraction).

4.4.2. Landscape Characteristics

In natural basins, urban (Figure 5ch) and barren (Figure 5ce) land cover emerged as the only significant types. NSE-C associated with urban coverage increases by a factor of 2 when more than 35% of the basin is urbanized. These basins are likely urbanized post GAGES-II classification. In contrast, increasing barren land cover (Figure 5ce) results in lower NSE-C in all basin types. In controlled basins, impervious surface (Figure 5cg), forest (Figure 5ca) and herbaceous (Figure 5cf) land cover are significant. NSE-C is larger (by a factor of 2) in basins with more than 15% impervious/forested while NSE-C is lower when grass coverage exceeds 20%.

Overall, 10 characteristics were statistically and practically significant in describing the variation in unconditional bias; among these, forest coverage was the only factor with a medium or larger effect size. In summary, as basins become more impervious (controlled) and urban (natural), unconditional bias increases. Meanwhile as controlled basins become more herbaceous, and all basins become more barren, unconditional bias decreases.

5. Discussion

5.1. Summary

Table 2 shows a broad summary of NWM2.0 performance and bias in the context of significant catchment characteristics. The spatial performance of the model is not unusual compared to other large scale streamflow simulations. However, the evaluation process used allows us to better understand the drivers behind components of NSE with respect to a suite of catchment characteristics.

Our results show a clear distinction between natural and controlled basins. Towler et al. (2023) also found that most underperforming basins in the NWM and USGS NHM have anthropogenic influences. Some of the challenge with modeling these systems are a lack of information about human impacts on the water cycle. At the drainage basin scale, none of the storage values tested here proved to be a significant indicator of performance.

For the most part, the characteristics impacting NSE-A were the same across controlled and natural basins. The exception being impervious percentage which was associated with a decrease in NSE-A in controlled basins.

Concerning bias, catchment characteristics were more closely associated with either natural or controlled basins. This suggests one of the principal differentiators in class performance is the bias generated in relation to certain characteristics. The exception to this is in Aridity, PPT and Snow Fraction which influence bias across all basin types (more on this these below).

A large part of the natural/controlled distinction is exacerbated by the calibration process that typically calibrates natural basins and transfers model parameters via receiver-donor relationships. The NWM for example, is calibrated to streamflow in a selection of natural GAGES-II basins. One possible solution to this, when focusing on continental scale models, is to avoid calibration solely on natural basins.

Initially, one of our hypotheses was that performance would be sensitive to characteristics used to parameterize the Land Surface Model. Broadly, highly vegetated surfaces like forests (increase bias) and herbaceous covers (decrease bias) showed a relationship with NSE-B/C in controlled basins. Conversely, more sparsely vegetated surfaces like urban (increase bias), barren (decrease bias) and shrubland (increase bias) had a relationship with NSE-B/C in natural basins.

Although this study did not consider the uncertainty arising from the NLDAS-2 forcings, recent studies have emphasized the importance, and sensitivity these have on predictions (Newman et al., 2015; Van Beusekom et al., 2022). That said, core meteorological characteristics including AI, PPT, and Snow Fraction significantly impacted skill and bias across all basin type. We highlighted a drop in NSE-A, and a rise in NSE-B/C in the middle of the country starting around the 95th meridian. Towler et al. also reported that sites in the central and mid-western regions of the U.S. underperform in both the NHM and NWM and other studies have seen the same behavior in models including VIC, ParFlow-CONUS, and SAC-SMA (Ghimire et al., 2023; Newman et al., 2015; Tijerina et al., 2021). Further, these limitations are consistent with the evaluation of LSM-driven streamflow by Cai, Yang, Xia, et al. (2014) who showed difficulty representing streamflow in the north central region of the country while "most models perform well east of the 95th meridian". The 100th meridian is known as a non-permanent divide splitting the continent into an "arid west" and a "humid east," based on differentiations in terms of vegetation, hydrology, crops, and farm economy (Seager et al., 2018) and is indicative of the role of energy, aridity, rainfall, and vegetation dynamics have on a model's relative skill and bias. In 2015, Newman et al. used SAC-SMA to evaluate 671 basins and found model performance varied regionally with the largest contributing factors being aridity and precipitation intermittency, contribution of snowmelt, and runoff seasonality. The similarities between their findings, and the influential factors impacting NSE-A, were striking given two different modeling frameworks (one physics based and one conceptual). Combined, it provides a signal that modeling communities at large need to better understand how to represent these regions more accurately without decreasing the skill achieved in other parts of the country. Some of this can be improved through the development of new model formations aimed at challenging areas like the Layered Green and Ampt infiltration with Redistribution (LGAR) soils routine for arid regions (La Follette et al., 2023). That said, accomplishing this goal will likely require the use of different model schemes in various areas of the country, a prospect becoming more promising with the rise of multi-model systems like SUMMA, the Unified Forecast System, and the Next Generation Water Prediction Framework.

5.2. A Multi Model Experiment

This study has demonstrated that model skill can be broken down into relative performance (NSE-A) and biases (NSE-B, NSE-C). It also showed how each of these components can be better understood in relation to a suite of catchment characteristics. Considering how this information can be used to enhance modeling efforts in the U.S., we discuss the role of electing spatially appropriate model combinations (Niu et al., 2011; Ogden et al., 2021) and explore the potential of hybrid modeling approaches to improve NSE-A while reducing NSE-B and C.

The findings of this paper highlight general reasons for the underperformance of the WRF-Hydro NWM in certain regions. Some of these points explicitly at regions characterized by dominant vegetation dynamics, high urbanization, and water limiting climates. The principal driver in all basins was AI, PET, PPT, and forest coverage. Here

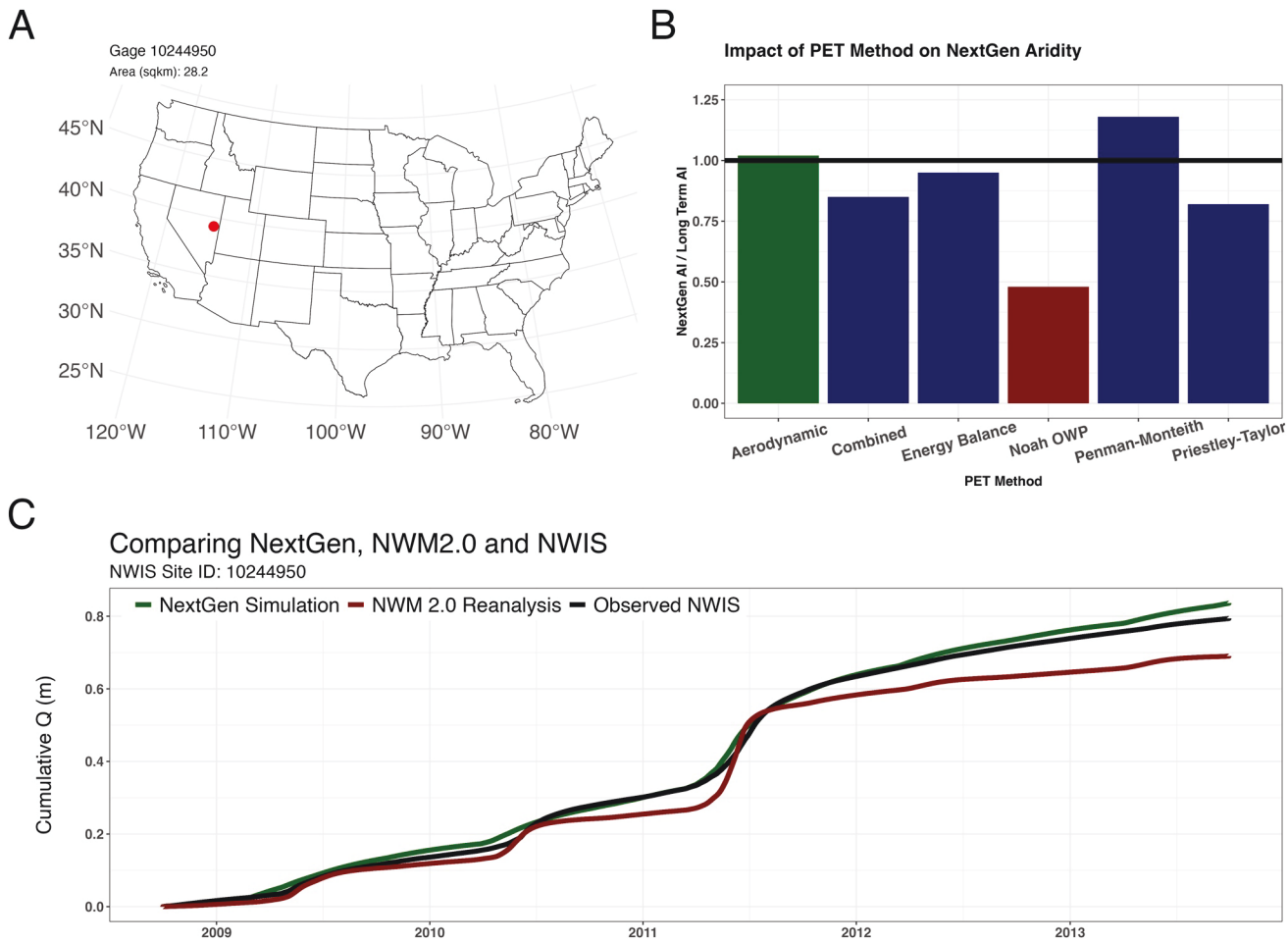

Figure 6. (a) A poor performing natural, arid basin in Nevada was selected. (b) 6 simulations were run using NextGen and the ratio of the simulated aridity index (AI) to the catchment AI was computed. The red bar approximates what was used in National Water Model (NWM) 2.0 while the ideal aerodynamic method (closest to 1) is in green. (c) Cumulative discharge plots of the USGS observations, NWM 2.0, and the aerodynamic NextGen simulation are shown highlighting the power of location driven processes.

we showcase how knowledge about the relationship between AI and model skill can help identify more optimal models in systems that allow for multiple combination schemes to be used (like NextGen or Noah-MP; Ogden et al., 2021; Niu et al., 2011). To highlight this, we selected one of the most biased natural, and arid basins in our study (NWIS ID 10244950). This basin presents an AI of 3.86, mean annual PPT of 464.51 mm and total runoff that was underestimated by the NWM 2.0 historical simulation (see Figure 6a). In the NWM 2.0 historic record, this basin presented an NSE-A of 0.45, an NSE-B of 0.25 and an extreme NSE-C of 7.3.

The NextGen framework was used to simulate streamflow in the basin over a 5-year period using the Conceptual Functional Equivalent model with the Xinanjiang rainfall-runoff partitioning module (https://github.com/NOAA-OWP/cfe) and six different PET methods including: (a) Noah-OWP-modular (https://github.com/NOAA-OWP/noah-owp-modular), (b) energy balance, (c) aerodynamic, (d) combined, (e) Priestley-Taylor, and (f) Penman-Monteith (https://github.com/NOAA-OWP/evapotranspiration). To identify the "best" of these, the AI of the catchment was compared with the AI produced by each simulation over the five year period.

In Figure 6b, the ratio of simulated AI to the long term catchment AI is shown. A ratio close to one indicates good agreement, while ratios smaller (larger) than one indicate the NextGen CFE formulation underestimates (overestimates) PET. For this basin, the aerodynamic method (green) estimates the AI ratio best while, the Noah-OWP-modular method (effectively a modular Noah-MP variant) significantly underpredicts PET.

Figure 6c displays the cumulative discharge from the NWM 2.0, the observed USGS flows, and the NextGen simulation showing the tailored model more accurately captured streamflow with a relative performance of 0.73

(compared to 0.45), a conditional bias of 0.0012 (compared to 0.25) and an unconditional bias of 0.0021 (compared to 7.3). Thus, one of the basins with the most bias and marginal relative performance was turned into a "good" simulation. While one basin does not allow us to draw broad scale conclusions, the potential to enhance a simulation by targeting an area with consistently poor performance in many models, is a promising sign for community efforts.

6. Conclusions

The NWM offers an unprecedented step forward in the hydrologic forecasting capabilities of the United States. Its innovation lies not only in advancing forecasting operations, but also in the developing an operational, near-real time, high-resolution LSM with minimal lag and comparatively sophisticated routing. However, this advancement necessitates the evaluation and diagnosis of the model in ways that explain not only *how* the model is performing but *why* it is performing that way. To achieve this, there must exist a comprehensive set of catchment characteristics that can be used to classify basin types in low and high dimensional space. These types of evaluations provide the opportunity to study the limitations of physical model process, identify improved physical representations that can be applied heterogeneously, and to explore opportunities for assimilating new data sources and postprocess output to enhance forecasts, for the appropriate reasons.

All models have ingrained assumptions (stated or unstated) that influence their performance. Most of these models are based on hydrological processes developed for pristine headwater basins in a particular location and for a specific event types. These assumptions imply that no single model is best everywhere, or, for all types of events. A framework like the one presented here offers a unique way to compare model results (either model-to-model or model-to-observation) that directly target questions related to model parametrization; process representation; and the presence of conditional and unconditional biases. Future research could use this decomposition framework to further diagnose error contributions from the entire modeling cycle including forcings, parameter estimation, process selection and calibration/regionalization. Moreover, this approach can be applied to other model development and intercomparison efforts. Its application to the NWM v2.0 historical data provides increased transparency for the public, catering to those seeking to use and improve NWM model outputs.

Data Availability Statement

The GAGES-II data set can be accessed at (Falcone, 2011). All streamflow data can be accessed from the USGS NWIS portal (U.S. Geologic Survey, 2023) or the NWM reanalysis archives (Johnson et al., 2023). Land cover data is accessed from the Multi Resolution Land Characteristics Consortium (Dewitz, 2021) and NLDAS data by NASA EarthData GES DISC service (NASA GES DISC, 2023). The complete data workflow including data download, processing, analysis, and image creation can be found on Github and Zenodo (Johnson, 2023).

Acknowledgments
Funding to publish this article provided by NSF through Grants OIA #1937099 and #2033607.

References

Abdulla, F. A., & Lettenmaier, D. P. (1997). Development of regional parameter estimation equations for a macroscale hydrologic model. *Journal of Hydrology*, *197*(1–4), 230–257. https://doi.org/10.1016/s0022-1694(96)03262-3

Abdulla, F. A., Lettenmaier, D. P., Wood, E. F., & Smith, J. A. (1996). Application of a macroscale hydrologic model to estimate the water balance of the Arkansas-Red River Basin. *Journal of Geophysical Research*, *101*(D3), 7449–7459. https://doi.org/10.1029/95jd02416

Adams, T., III. (2016). Flood forecasting in the United States NOAA/National Weather Service. In *Flood forecasting* (pp. 249–310). Elsevier.

Anderson, J. R. (1976). *A land use and land cover classification system for use with remote sensor data* (Vol. 964). U.S. Government Printing Office.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., et al. (2015). Accelerating advances in continental domain hydrologic modeling. *Water Resources Research*, *51*(12), 10078–10091. https://doi.org/10.1002/2015wr017498

Barlage, M. (2017). The Noah-MP land surface model.

Bierkens, M. F. P. (2015). Global hydrology 2015: State, trends, and directions. *Water Resources Research*, *51*(7), 4923–4947. https://doi.org/10.1002/2015wr017173

Blodgett, D. L., & Dornblut, I. (2018). OGC WaterML 2: Part 3-surface hydrology features (HY_Features)-conceptual model Version 1.0.

Blodgett, D. L., & Johnson, J. M. (2022). Hydrologic modeling and river corridor applications of HY_Features concepts. Retrieved from http://www.opengis.net/doc/PER/22-040

Brackins, J., Moragoda, N., Rahman, A., Cohen, S., & Lowry, C. (2021). The role of realistic channel geometry representation in hydrological model predictions. *JAWRA Journal of the American Water Resources Association*, *57*(2), 222–240. https://doi.org/10.1111/1752-1688.12865

Bretherton, C., Balaji, V., Delworth, T., Dickinson, R., Edmonds, J., Famiglietti, J., & Smarr, L. (2012). A national strategy for advancing climate modeling.

Burnash, R. (1995). The NWS river forecast system-catchment modeling. In *Computer models of watershed hydrology* (pp. 311–366).

Cai, X., Yang, Z.-L., David, C. H., Niu, G.-Y., & Rodell, M. (2014). Hydrological evaluation of the Noah-MP land surface model for the Mississippi River Basin. *Journal of Geophysical Research: Atmospheres*, *119*(1), 23–38. https://doi.org/10.1002/2013jd020792

Cai, X., Yang, Z.-L., Xia, Y., Huang, M., Wei, H., Leung, L. R., & Ek, M. B. (2014). Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed. *Journal of Geophysical Research: Atmospheres*, *119*(24), 13751–13770. https://doi.org/10.1002/2014JD022113

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The effects of climate change on the hydrology and water resources of the Colorado River basin. *Climatic Change*, *62*(1–3), 337–363. https://doi.org/10.1023/b:clim.0000013684.13621.1f

Clark, M. P., Fan, Y., Lawrence, D. M., Adam, J. C., Bolster, D., Gochis, D. J., et al. (2015). Improving the representation of hydrologic processes in Earth System Models. *Water Resources Research*, *51*(8), 5929–5956. https://doi.org/10.1002/2015wr017096

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.

Cosgrove, B. A., Gochis, D. J., Clark, E. P., & Flowers, T. (2020). NOAA's National Water Model: A dynamically evolving operational hydrologic forecasting framework.

Cosgrove, B. A., Lohmann, D., Mitchell, K. E., Houser, P. R., Wood, E. F., Schaake, J. C., et al. (2003). Real-time and retrospective forcing in the North American Land Data Assimilation System (NLDAS) project. *Journal of Geophysical Research*, *108*(D22), 2002JD003118. https://doi.org/10.1029/2002jd003118

Criss, R. E., & Winston, W. E. (2008). Do Nash values have value? Discussion and alternate proposals. *Hydrological Processes*, *22*(14), 2723–2725. https://doi.org/10.1002/hyp.7072

De Cicco, L. A., Lorenz, D., Hirsch, R. M., Watkins, W., & Johnson, M. (2018). dataRetrieval: R packages for discovering and retrieving water data available from U.S. federal hydrologic web services (manual) [Software]. U.S. Geological Survey. https://doi.org/10.5066/P9X4L3GE

Dewitz, J. (2021). *National Land Cover Database (NLCD) 2019 products [Land Cover L48]*. U.S. Geological Survey. https://doi.org/10.5066/P9KZCM54

Eyring, V., Bony, S., Meehl, G. A., Senior, C. A., Stevens, B., Stouffer, R. J., & Taylor, K. E. (2016). Overview of the Coupled Model Intercomparison Project Phase 6 (CMIP6) experimental design and organization. *Geoscientific Model Development*, *9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016

Falcone, J. A. (2011). *GAGES-II: Geospatial attributes of gages for evaluating streamflow*. U.S. Geological Survey. Retrieved from https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml

Ghimire, G. R., Hansen, C., Gangrade, S., Kao, S. C., Thornton, P. E., & Singh, D. (2023). Insights from dayflow: A historical streamflow reanalysis dataset for the conterminous United States. *Water Resources Research*, *59*(2), e2022WR032312. https://doi.org/10.1029/2022wr032312

Guo, Y., Zhang, Y., Zhang, L., & Wang, Z. (2021). Regionalization of hydrological modeling for predicting streamflow in ungauged catchments: A comprehensive review. *Wiley Interdisciplinary Reviews: Water*, *8*(1), e1487. https://doi.org/10.1002/wat2.1487

Gupta, H. V., Perrin, C., Blöschl, G., Montanari, A., Kumar, R., Clark, M., & Andréassian, V. (2014). Large-sample hydrology: A need to balance depth with breadth. *Hydrology and Earth System Sciences*, *18*(2), 463–477. https://doi.org/10.5194/hess-18-463-2014

Hansen, C., Shafiei Shiva, J., McDonald, S., & Nabors, A. (2019). Assessing retrospective national water model streamflow with respect to droughts and low flows in the Colorado River Basin. *JAWRA Journal of the American Water Resources Association*, *55*(4), 964–975. https://doi.org/10.1111/1752-1688.12784

Hirabayashi, Y., Mahendran, R., Koirala, S., Konoshima, L., Yamazaki, D., Watanabe, S., et al. (2013). Global flood risk under climate change. *Nature Publishing Group*, *3*(9), 816–821. https://doi.org/10.1038/nclimate1911

Homer, C., Dewitz, J., Yang, L., Jin, S., Danielson, P., Xian, G., et al. (n.d.). *Completion of the 2011 National Land Cover Database for the conterminous United States–Representing a decade of land cover change information*. Photogrammetric Engineering.

Jachens, E. R., Hutcheson, H., Thomas, M. B., & Steward, D. R. (2020). *Effects of groundwater-surface water exchange mechanism in the National Water Model over the northern high plains aquifer, USA*. JAWRA Journal of the American Water Resources Association.

Jehn, F. U., Bestian, K., Breuer, L., Kraft, P., & Houska, T. (2020). Using hydrological and climatic catchment clusters to explore drivers of catchment behavior. *Hydrology and Earth System Sciences*, *24*(3), 1081–1100. https://doi.org/10.5194/hess-24-1081-2020

Johnson, J. M. (2023). Analysis code and data for comprehensive analysis of the NOAA National Water Model: A call for heterogeneous formulations and diagnostic model selection (Version v1.0) [Computer software]. https://doi.org/10.5281/zenodo.10080619

Johnson, J. M., & Blodgett, D. L. (2020). NOAA National Water Model reanalysis data at RENCI [Dataset]. HydroShare. Retrieved from http://www.hydroshare.org/resource/89b0952512dd4b378dc5be8d2093310f

Johnson, J. M., Blodgett, D. L., Clarke, K. C., & Pollak, J. (2023). Restructuring and serving web-accessible streamflow data from the NOAA National Water Model historic simulations. *Scientific Data*, *10*(1), 725. https://doi.org/10.1038/s41597-023-02316-7

Johnson, J. M., & Clarke, K. C. (2021). An area preserving method for improved categorical raster resampling. *Cartography and Geographic Information Science*, *48*(4), 1–13. https://doi.org/10.1080/15230406.2021.1892531

Johnson, J. M., Coll, J. M., Ruess, P. J., & Hastings, J. T. (2018). Challenges and opportunities for creating intelligent hazard alerts: The "FloodHippo" prototype. *Journal of the American Water Resources Association*, *54*(4), 872–881. https://doi.org/10.1111/1752-1688.12645

Johnson, J. M., Munasinghe, D., Eyelade, D., & Cohen, S. (2019). An integrated evaluation of the National Water Model (NWM)–Height Above Nearest Drainage (HAND) flood mapping methodology. *Natural Hazards and Earth System Sciences*, *19*(11), 2405–2420. https://doi.org/10.5194/nhess-19-2405-2019

Johnson, J. M., Narock, T., Singh-Mohudpur, J., Fils, D., Clarke, K. C., Saksena, S., et al. (2022). Knowledge graphs to support real-time flood impact evaluation. *AI Magazine*, *43*(1), 40–45. https://doi.org/10.1002/aaai.12035

Kim, D. H. D., Johnson, J. M., Clarke, K. C., & McMillan, H. K. (2023). Untangling the impacts of land cover representation and resampling in distributed hydrological model predictions. *Environmental Modelling & Software*, 105893.

Kim, H., & Villarini, G. (2022). Evaluation of the Analysis of Record for Calibration (AORC) rainfall across Louisiana. *Remote Sensing*, *14*, 3284. https://doi.org/10.3390/rs14143284

Kirchner, J. W. (2006). Getting the right answers for the right reasons: Linking measurements, analyses, and models to advance the science of hydrology. *Water Resources Research*, *42*(3), W03S04. https://doi.org/10.1029/2005wr004362

Kitzmiller, D. H., Wu, W., Zhang, Z., Patrick, N., & Tan, X. (2018). The analysis of record for calibration: A high-resolution precipitation and surface weather dataset for the United States. In *AGU fall meeting abstracts* (Vol. 2018, p. H41H-06).

Kollet, S., Sulis, M., Maxwell, R. M., Paniconi, C., Putti, M., Bertoldi, G., et al. (2017). The integrated hydrologic model intercomparison project, IH-MIP2: A second set of benchmark results to diagnose integrated hydrology and feedbacks. *Water Resources Research*, *53*(1), 867–890. https://doi.org/10.1002/2016wr019191

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 Resources Research*, *55*(12), 11344–11354. https://doi.org/10.1029/2019wr026065

La Follette, P., Ogden, F. L., & Jan, A. (2023). Layered Green and Ampt infiltration with redistribution. *Water Resources Research*, *59*(7), e2022WR033742. https://doi.org/10.1029/2022wr033742

Li, W., Sankarasubramanian, A., Ranjithan, R. S., & Sinha, T. (2016). Role of multimodel combination and data assimilation in improving streamflow prediction over multiple time scales. *Stochastic Environmental Research and Risk Assessment*, *30*(8), 2255–2269. https://doi.org/10.1007/s00477-015-1158-6

Lin, P., Rajib, M. A., Yang, Z., Somos-Valenzuela, M., Merwade, V., Maidment, D. R., et al. (2018). Spatiotemporal evaluation of simulated evapotranspiration and streamflow over Texas using the WRF-Hydro-RAPID modeling framework. *JAWRA Journal of the American Water Resources Association*, *54*(1), 40–54. https://doi.org/10.1111/1752-1688.12585

Livneh, B., Rosenberg, E. A., Lin, C., Nijssen, B., Mishra, V., Andreadis, K. M., et al. (2013). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States: Update and extensions. *Journal of Climate*, *26*(23), 9384–9392. https://doi.org/10.1175/jcli-d-12-00508.1

Maidment, D. R. (2016). Conceptual framework for the national flood interoperability experiment. *JAWRA Journal of the American Water Resources Association*, *53*(2), 245–257. https://doi.org/10.1111/1752-1688.12474

Maurer, E. P., O'Donnell, G. M., Lettenmaier, D. P., & Roads, J. O. (2001). Evaluation of the land surface water budget in NCEP/NCAR and NCEP/DOE reanalyses using an off-line hydrologic model. *Journal of Geophysical Research*, *106*(D16), 17841–17862. https://doi.org/10.1029/2000jd900828

Maurer, E. P., Wood, A. W., Adam, J. C., Lettenmaier, D. P., & Nijssen, B. (2002). A long-term hydrologically based dataset of land surface fluxes and states for the conterminous United States. *Journal of Climate*, *15*(22), 3237–3251. https://doi.org/10.1175/1520-0442(2002)015<3237:althbd>2.0.co;2

Mazrooei, A., Sinha, T., Sankarasubramanian, A., Kumar, S., & Peters-Lidard, C. D. (2015). Decomposition of sources of errors in seasonal streamflow forecasting over the U.S. Sunbelt. *Journal of Geophysical Research: Atmospheres*, *120*(23), 11809–11825. https://doi.org/10.1002/2015JD023687

McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the Nash–Sutcliffe efficiency index. *Journal of Hydrologic Engineering*, *11*(6), 597–602. https://doi.org/10.1061/(asce)1084-0699(2006)11:6(597)

Mitchell, K. E., Lohmann, D., Houser, P. R., Wood, E. F., Schaake, J. C., Robock, A., et al. (2004). The multi-institution North American Land Data Assimilation System (NLDAS): Utilizing multiple GCIP products and partners in a continental distributed hydrological modeling system. *Journal of Geophysical Research*, *109*(D7), D07S90. https://doi.org/10.1029/2003jd003823

Mo, K. C., Chen, L. C., Shukla, S., Bohn, T. J., & Lettenmaier, D. P. (2012). Uncertainties in North American land data assimilation systems over the contiguous United States. *Journal of Hydrometeorology*, *13*(3), 996–1009. https://doi.org/10.1175/jhm-d-11-0132.1

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, *50*(3), 885–900. https://doi.org/10.13031/2013.23153

Murphy, A. H. (1988). Skill scores based on the mean square error and their relationships to the correlation coefficient. *Monthly Weather Review*, *116*(12), 2417–2424. https://doi.org/10.1175/1520-0493(1988)116<2417:ssbotm>2.0.co;2

NASA GES DISC. (2023). North American Land Data Assimilation System Data Archive. Retrieved from https://disc.gsfc.nasa.gov/datasets?keywords=NLDAS

Nash, J. E., & Sutcliffe, J. V. (1970). River flow forecasting through conceptual models part I—A discussion of principles. *Journal of Hydrology*, *10*(3), 282–290. https://doi.org/10.1016/0022-1694(70)90255-6

National Inventory of Dams. (2019). US Army Corps of Engineers: Federal Emergency Management Agency.

Newman, A. J., Clark, M. P., Sampson, K., Wood, A., Hay, L. E., Bock, A., et al. (2015). Development of a large-sample watershed-scale hydrometeorological data set for the contiguous USA: Data set characteristics and assessment of regional variability in hydrologic model performance. *Hydrology and Earth System Sciences*, *19*(1), 209–223. https://doi.org/10.5194/hess-19-209-2015

Nijssen, B., O'Donnell, G. M., Lettenmaier, D. P., Lohmann, D., & Wood, E. F. (2001). Predicting the discharge of global rivers. *Journal of Climate*, *14*(15), 3307–3323. https://doi.org/10.1175/1520-0442(2001)014<3307:ptdogr>2.0.co;2

Niu, G.-Y., Yang, Z.-L., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 1. Model description and evaluation with local-scale measurements. *Journal of Geophysical Research*, *116*(D12), 1381–1419. https://doi.org/10.1029/2010jd015139

NOAA National Water Model CONUS Retrospective Dataset. (n.d.). Retrieved from https://registry.opendata.aws/nwm-archive

Nossent, J., & Bauwens, W. (2012). Application of a normalized Nash-Sutcliffe efficiency to improve the accuracy of the Sobol' sensitivity analysis of a hydrological model. In *EGUGA* (p. 237).

Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., et al. (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* (Vol. 2021, pp. H43D-01).

Peckham, S. D., Hutton, E. W., & Norris, B. (2013). A component-based approach to integrated modeling in the geosciences: The design of CSDMS. *Computers & Geosciences*, *53*, 3–12. https://doi.org/10.1016/j.cageo.2012.04.002

Pekel, J.-F., Cottam, A., Gorelick, N., & Belward, A. S. (2016). High-resolution mapping of global surface water and its long-term changes. *Nature*, *540*(7633), 418–422. https://doi.org/10.1038/nature20584

Petersen, T., Devineni, N., & Sankarasubramanian, A. (2012). Seasonality of monthly runoff over the continental United States: Causality and relations to mean annual and mean monthly distributions of moisture and energy. *Journal of Hydrology*, *468*, 139–150. https://doi.org/10.1016/j.jhydrol.2012.08.028

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E., Van Emmerik, T., Uijlenhoet, R., et al. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017

Ritter, A., & Muñoz-Carpena, R. (2013). Performance evaluation of hydrological models: Statistical significance for reducing subjectivity in goodness-of-fit assessments. *Journal of Hydrology*, *480*, 33–45. https://doi.org/10.1016/j.jhydrol.2012.12.004

Rojas, M., Quintero, F., & Krajewski, W. F. (2020). Performance of the national water model in Iowa using independent observations. *JAWRA Journal of the American Water Resources Association*, *56*(4), 568–585. https://doi.org/10.1111/1752-1688.12820

Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., et al. (2017). Towards real-time continental scale streamflow simulation in continuous and discrete space. *JAWRA Journal of the American Water Resources Association*, *51*(12), 10078–10121. https://doi.org/10.1111/1752-1688.12586

Sankarasubramanian, A., & Vogel, R. M. (2002). Annual hydroclimatology of the United States. *Water Resources Research*, *38*(6), 19–21. https://doi.org/10.1029/2001wr000619

Seager, R., Lis, N., Feldman, J., Ting, M., Williams, A. P., Nakamura, J., et al. (2018). Whither the 100th meridian? The once and future physical and human geography of America's arid–humid divide. Part I: The story so far. *Earth Interactions*, *22*(5), 1–22. https://doi.org/10.1175/ei-d-17-0011.1

Slack, J. R., Lumb, A. M., & Landwehr, J. M. (1993). *Hydro-climatic data network (HCDN) streamflow data set, 1874-1988*. U.S. Geological Survey.

Sullivan, G. M., & Feinn, R. (2012). Using effect size—Or why the *P* value is not enough. *Journal of Graduate Medical Education*, *4*(3), 279–282. https://doi.org/10.4300/JGME-D-12-00156.1

Tijerina, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M. M., Sampson, K., et al. (2021). Continental hydrologic intercomparison project, phase 1: A large-scale hydrologic model comparison over the continental United States. *Water Resources Research*, *57*(7), e2020WR028931. https://doi.org/10.1029/2020wr028931

Towler, E., Foks, S. S., Dugger, A. L., Dickinson, J. E., Essaid, H. I., Gochis, D., et al. (2023). Benchmarking high-resolution hydrologic model performance of long-term retrospective streamflow simulations in the contiguous United States. *Hydrology and Earth System Sciences*, *27*(9), 1809–1825. https://doi.org/10.5194/hess-27-1809-2023

U.S. Geological Survey. (2023). USGS water data for the nation: U.S. Geological Survey National Water Information System Database. https://doi.org/10.5066/F7P55KJN

Van Beusekom, A. E., Hay, L. E., Bennett, A. R., Choi, Y. D., Clark, M. P., Goodall, J. L., et al. (2022). Hydrologic model sensitivity to temporal aggregation of meteorological forcing data: A case study for the contiguous United States. *Journal of Hydrometeorology*, *23*(2), 167–183. https://doi.org/10.1175/jhm-d-21-0111.1

Van Loon, A. F., Gleeson, T., Clark, J., Van Dijk, A. I. J. M., Stahl, K., Hannaford, J., et al. (2016). Drought in the Anthropocene. *Nature Geoscience*, *9*(2), 89–91. https://doi.org/10.1038/ngeo2646

Viterbo, F., Read, L., Nowak, K., Wood, A. W., Gochis, D., Cifelli, R., & Hughes, M. (2020). General assessment of the operational utility of National Water Model reservoir inflows for the Bureau of Reclamation Facilities. *Water*, *12*(10), 2897. https://doi.org/10.3390/w12102897

Weglarczyk, S. (1998). The interdependence and applicability of some statistical quality measures for hydrological models. *Journal of Hydrology*, *206*(1–2), 98–103. https://doi.org/10.1016/s0022-1694(98)00094-8

Wens, M., Johnson, J. M., Zagaria, C., & Veldkamp, T. I. E. (2019). Integrating human behavior dynamics into drought risk assessment—A sociohydrologic, agent-based approach. *Wiley Interdisciplinary Reviews: Water*, *105*(32), e1345. https://doi.org/10.1002/wat2.1345

Wood, E. F., Roundy, J. K., Troy, T. J., van Beek, L. P. H., Bierkens, M. F. P., Blyth, E., et al. (2011). Hyperresolution global land surface modeling: Meeting a grand challenge for monitoring Earth's terrestrial water. *Water Resources Research*, *47*(5), W05301. https://doi.org/10.1029/2010wr010090

Yang, L., Jin, S., Danielson, P., Homer, C., Gass, L., Bender, S. M., et al. (2018). A new generation of the United States National Land Cover Database: Requirements, research priorities, design, and implementation strategies. *ISPRS Journal of Photogrammetry and Remote Sensing*, *146*, 108–123. https://doi.org/10.1016/j.isprsjprs.2018.09.006

Erratum

The originally published version of this article contained errors in the affiliations and the reference list. Authors J. Michael Johnson, Arash Modaresi Rad, Luciana Kindl da Cunha, and Keith S. Jennings should be affiliated with "NOAA/NWS Office of Water Prediction" in addition to their existing affiliations. The following disclaimer has been added: "The views expressed in this article do not necessarily represent the views of NOAA or the United States." The reference for Office of Water Prediction (2022) was updated to the following: Ogden, F., Avant, B., Bartel, R., Blodgett, D., Clark, E., Coon, E., et al. (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* (Vol. 2021, pp. H43D-01). In addition, the citations for this reference have been updated throughout the article, and it was also added before Peckham et al., 2013 in the first sentence of the eleventh paragraph of the Introduction. The errors have been corrected, and this may be considered the authoritative version of record.