Posted on 19 Jan 2023 — The copyright holder is the author/funder. All rights reserved. No reuse without permission. — https://doi.org/10.22541/essoar.167415214.45806648/v1 — This a preprint and has not been peer reviewed. Data may be preliminary.

Comprehensive analysis of the NOAA National Water Model: A call for heterogeneous formulations and diagnostic model selection

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

[1]Lynker
[2]North Carolina State University
[3]West Consultants
[4]University of California Santa Barbara
[5]National Center for Atmospheric Research, NCAR
[6]University of Cincinnati

January 19, 2023

Abstract

With an increasing number of continental-scale hydrologic models, the ability to evaluate performance is key to understanding uncertainty in prediction and making improvements to the model(s). In 2016, the NOAA National Water Model (NWM) was put into operations to improve the spatial and temporal resolution of hydrologic prediction in the U.S. Here, we evaluate the NWM 2.0 historical streamflow record in natural and controlled basins using the Nash Sutcliffe Efficiency metric decomposed into relative error, conditional, and unconditional bias. Each of these is evaluated in the contexts of categorized meteorologic, landscape, and anthropogenic characteristics to assess model performance and diagnose error types. Broadly speaking greater rainfall and snow coverage leads to improved performance while larger potential evapotranspiration (PET), aridity, and phase correlation reduce performance. More rainfall and phase correlation reduce overall bias, while increasing PET, aridity, snow coverage/fraction increase model bias. With respect to landscape traits, more barren and agricultural land yeild improved performance while more forest, shrubland, grassland and imperviousness tend to decrease performance. Lastly, more barren and herbaceous land tend to decrease bias, while greater imperviousness, urban, forest, and shrubland cover increase bias. The insights gained can help identify key hydrological factors in NWM predictions; enforce the need for regionalized physics and modeling; and help develop hybid post-processing methods to improve prediction. Finally, we demonstrate how the NOAA Next Generation Water Resource Modeling Framework can help reduce the structural bias through the application of heterogenous model processes and highlight opportunities for ongoing development and evaluation.

Hosted file

954071_0_art_file_10613013_rhhcdt.docx available at https://authorea.com/users/577649/articles/619945-comprehensive-analysis-of-the-noaa-national-water-model-a-call-for-heterogeneous-formulations-and-diagnostic-model-selection

Hosted file

954071_0_supp_10613015_rqqcdt.docx available at https://authorea.com/users/577649/articles/619945-comprehensive-analysis-of-the-noaa-national-water-model-a-call-for-heterogeneous-formulations-and-diagnostic-model-selection

Comprehensive analysis of the NOAA National Water Model: A call for heterogeneous formulations and diagnostic model selection

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

[1] Lynker, Fort Collins CO, USA

[2] University of California, Santa Barbara, Department of Geography

[3] North Carolina State University, Raleigh NC, USA

[4] West Consultants, Sacramento, CA

[5] Research Applications Laboratory, National Center for Atmospheric Research

[6] University of Cincinnati, Ohio, USA

Corresponding author: J. Michael Johnson (jjohnson@lynker.com)

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, phase correlation, forest and grass cover are key characteristics suggesting limited ability to model regional evapotranspiration
- Similar results can inform regional physics in heterogeneous models while large biases indicates opportunity for successful post-processing

Abstract

With an increasing number of continental-scale hydrologic models, the ability to evaluate performance is key to understanding uncertainty in prediction and making improvements to the model(s). In 2016, the NOAA National Water Model (NWM) was put into operations to improve the spatial and temporal resolution of hydrologic prediction in the U.S. Here, we evaluate the NWM 2.0 historical streamflow record in natural and controlled basins using the Nash Sutcliffe Efficiency metric decomposed into relative error, conditional, and unconditional bias. Each of these is evaluated in the contexts of categorized meteorologic, landscape, and anthropogenic characteristics to assess model performance and diagnose error types. Broadly speaking greater rainfall and snow coverage leads to improved performance while larger potential evapotranspiration (PET), aridity, and phase correlation reduce performance. More rainfall and phase correlation reduce overall bias, while increasing PET, aridity, snow coverage/fraction increase model bias. With respect to landscape traits, more barren and agricultural land yeild improved performance while more forest, shrubland, grassland and imperviousness tend to decrease performance. Lastly, more barren and herbaceous land tend to decrease bias, while greater imperviousness, urban, forest, and shrubland cover increase bias. The insights gained can help identify key hydrological factors in NWM predictions; enforce the need for regionalized physics and modeling; and help develop hybid post-processing methods to improve prediction. Finally, we demonstrate how the NOAA Next Generation Water Resource Modeling Framework can help reduce the structural bias through the application of heterogenous model processes and highlight opportunities for ongoing development and evaluation.

Plain Language Summary

Water related issues are posing greater challenges to society both in terms of responding to extreme events and planning for the future. One approach to better understanding water supply and extreme events is through hydrologic models. NOAA has implemented a National Water Model (NWM), intended 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. We diagnose where the model is performing well (and poorly) in the contexts of landscape, weather conditions, and human influence. The insights gained can help identify key factors driving NWM skill and enforce that different physics are needed in different places. Lastly, we show how understanding why the NWM is performing the way it does can help us diagnostically select different physics options or our modeling approach within the NOAA Next Generation Water Resource Modeling Framework and to reduce error in the existing model output. Overall, this research provides a method for diagnosing performance in continental-scale, high-resolution, processed-based hydrologic models and demonstrates how that information can be used to guide use of the outputs and improve the model itself.

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 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 to represent the energy and water components of the earth system (Archfield et al., 2015). For example, offical streamflow forecasts in the US are issued by the 13 river forecasting centers (RFC) across ~3,600 sub-catchments (Adams III, 2016; Burnash, 1995; Salas et al., 2017). To increase spatial coverage, many modeling systems use grid-based Land Surface Models (LSM) 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 (Liu et al., 2012; Livneh et al., 2013; Maurer et al., 2002; Pekel et al., 2016).

While many land surface models can be used for continental-scale hydrologic modeling, they were historically built to provide land surface boundary conditions in coupled models. In that role, LSMs have a stronger focus on closing the energy balance than most catchment models. Additionally, 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 seperate routing models (Li et al., 2016). The second is that when the same equations and parameters (Johnson & Clarke, 2021) are applied across the entire domain, location specific performance tends to degrade. For example Cai et al. compared four LSMs across the continental United States (CONUS) using the North American Land Data Assimilation System (NLDAS) test bed (Cai et al., 2015) and in each model, the relative bias in the continental evaluations was larger than those in regional studies (Abdulla et al., 1996; Cai et al., 2014; Christensen et al., 2004).

In 2016, NOAA undertook the role of providing reach-level forecasts for the entire US to enhance the authoritative forecasts provided by the RFCs through the National Water Model (NWM). To meet these requirements, the NWM had to be a LSM with a high-fidelity routing component. The WRF-hydro community modeling framework was chosen to implement a 1km^2 version of the Noah-MP LSM (Niu et al., 2011; Z.-L. Yang et al., 2011) with the WRF-Hydro routing model (D. J. Gochis et al., 2013; J. Gochis & Chen, 2003) to provide hourly streamflow forecasts at ~2.7 million locations across the continental US. One of the biggest hurdles with the current NWM is the ambiguity in model reliability and a lack of published knowledge about model performance. This is a primary gap we hope this research can fill.

Today the NWM is in its fourth version (v2.2), and through its evolution, and with each release, a multi-decade historic simulation has also been produced (*NOAA National Water Model Reanalysis Model Data on AWS*, n.d.). The performance of the operational, or experimental model has seen regional evaluations. 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) and 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).

Some applications have also focused on using the the historic data 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.

Looking forward, the NOAA Office of Water Prediction has recognized the limitations of a large scale LSM and that improvements from calibration alone are beginning to plateau (Office of Water Prediction, 2022). This limitation sparked the NOAA Next Generation Water Resource Modeling Framework (NextGen) as a means for heterogenous model formulations to be run in a single application. NextGen is unified by the conceptual notion of a "hydro-nexus" based on the Open Geospatial Consortium (OGC) WaterML version 2.0 HY_Features international standard for representing surface hydrologic features (D. Blodgett & Dornblut, 2018; D. L. Blodgett & Johnson, 2022) and the enforcement of this conceptual model, paired with the Basic Model Interface (Peckham et al., 2013), provides an open source, standards based framework that allows modeling approaches to be regionally tailored for specific streamflow generation processes. The questions that persist are what regional traits are currently limiting model skill, what areas of the country most critically need improvements, and what processes (determined by geophysical characteristics) are driving performance and model bias in a positive or negative direction?

Here we seek to use them categorizing the performance of the NWM 2.0 across CONUS by decomposing the correlation, conditional, and unconditional bias with respect to a robust set of catchment characteristics. In doing this we seek to highlight the current state of the NWM; provide a general evaluation workflow that leverages the geospatial data fabrics that will be available for NextGen; and highlight areas and processes NextGen development can target. The discussion will outline the value of consistent catchment characteristics for this type of work and how the understanding gained in this work can be used to improve model performance both in runtime and through post-processing.

2 Data

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

2.1 Gauging Locations and streamflow records

Gage locations were selected from the Geospatial Attributes of Gages for Evaluating Streamflow (GAGES-II) dataset (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 dataset goes beyond those in the USGS Hydro-Climatic Data Network (HCDN), which focused on gages that experienced natural flow regimes at some point in the past (Slack et al., 1993). The USGS sites IDs 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 1993-01-01 and 2018-12-31; a size between 20 and 20,000 km^2; and that were completely within the USA were retained. Figure S1 shows the locations of the controlled and natural basins overlayed on a map of 26 year mean Aridity Index in CONUS.

The historical record for NWM v2.0 is approximately 40TB in size, 10TB of which is the channel output. Johnson et. al restructured this dataset to support broad scale evaluations and applications and is accessible through the nwmTools R package (Johnson, 2020; Johnson & Blodgett, 2020). Hourly records were summarized to daily averages to remain consistent with the NWIS readings, and, in total, 4,713 basins are available for analysis with natural basins making up ~21% of the dataset.

2.2 Basin Characteristics

All physical and machine learning models rely on accurate geospatial data to discretize and parameterize the models and high-quality datasets are essential for hydrological modeling and evaluation. The utility of the catchment characteristics - for a given set of areas - 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 a set of landscape, meterological, and anthropogenic characterics that we will use to characterize NWM performance. Table S1 identifies all catchment characteristics tested as well as their source data, range, and units

2.2.1 Landscape Characteristics

Noah-MP is a spatially distributed (gridded) land surface model with multiple options for land-atmosphere interaction processes (Niu et al., 2011). To determine parameter and process behavior for specific grid cells, the model relies heavily on land cover inputs. In total, forty-nine variables 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 SAI and LAI) (Barlage, 2017). 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 Dataset (NLCD) was determined (Anderson, 1976; Homer et al., n.d.; L. Yang et al., 2018) (Anderson, 1979; Homer, 2016; Yang 2018). The total impervious surface was also determined from the 2019 NLCD Impervious data product.

2.2.2 Meterological Characteristics

Following Lin's analysis of the NWM in Texas; Cai's broad evaluation of land surface models, and Peterson's evaluation of LSM models, 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; $\frac{kg}{m^2}$), Aridity Index (AI), energy correlation, snow water equivalent and snow coverage. PET and PPT were obtained from the primary forcing data of the phase 2 NLDAS for January 1993 through December 2018. For each basin the mean monthly PET and PPT was summarized over the basin areas using a method that weighted partially covered grid cells by the percentage of containment. An aridity index (AI) was calculated as the ratio of annual mean PPT to annual mean PET ($\frac{PET}{PPT}$) 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 1-2 temperate; between 2-3 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 that 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 covariability value was determined for each basin.

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

2.2.1 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 in each basin as well as the cumulative storage (NID_STORAGE). 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 historic simulation of the NWMv2.0 is able to capture the observed USGS streamflow record at a daily timescale scale. To do this,
the Nash Sutcliffe Efficiency was calculated for each location across the shared timeseries (equation 1; Nash & Sutcliffe, 1970).

$$\text{NSE} = 1 - \frac{\sum_{t=1}^{T} (Q_m^t - Q_o^t)^2}{\sum_{t=1}^{T} (Q_o^t - meanQ_o^t)^2} \qquad (1)$$

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 is a better predictor than the model.

Subjective NSE thresholds have been suggested by several authors (Criss & Winston, 2008; D. N. Moriasi et al., 2007; McCuen et al., 2006; 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). Perhaps these are too strict for the daily evaluation being performed here but they provide a general qualitative categorization. With more than 4,000 sites being evaluated, the lower NSE limit of -∞ can become problematic and in these cases, a Normalized NSE (NNSE) rescaled to the range of {0,1} can be computed (equation 2, Nossent & Bauwens, 2012).

$$\text{NNSE} = \frac{1}{2 - NSE} \qquad (2)$$

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.

NSE can also 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) (equation 3-6). This disaggregation is shown in equations 3-6.

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

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

$$B = \left(r - \frac{\sigma s}{\sigma o}\right)^2 \tag{5}$$

$$C = \left(\frac{(\mu s - \mu o)}{\sigma o}\right)^2 \tag{6}$$

Where r is the Pearson correlation coefficient (see Figure S2 for more information); 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.

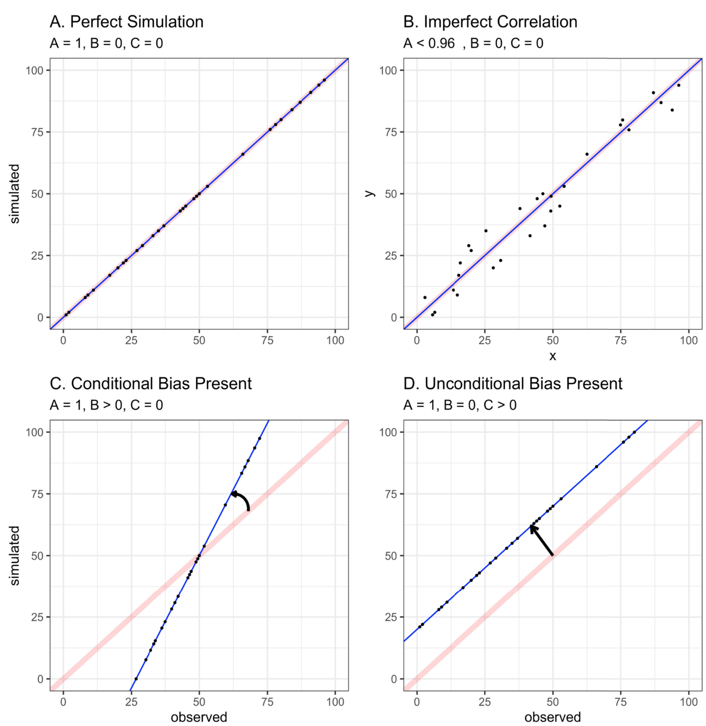

Figure 1: Conceptual diagram illustrating how NSE-A, B and C appear in a scatter plot of observed vs. simulated flows. Panel A shows a perfect simulation where A = 1 and there is no bias **(**B=C=0**)**. Panel B shows an example where there is no bias (B=C=0) and high, but imperfect correlation **(**A < 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).

3.2 Analysis of Variance ANOVA (Type II)

We used a series of ANOVA tests to find statistically significant catchment characteristics for 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 η2 squared.

$$\eta 2 = \frac{SS_{effect}}{SS_{total}} \tag{7}$$

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 η2 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 η2 states that (Cohen, 2013):

η2 > 0.01 indicates a small effect
η2 > 0.06 indicates a medium effect
η2 > 0.14 indicates a large effect

For our tests, we run independent ANOVA tests for each catchment characteristic in Table S1, on each NSE component, for natural basins and controlled basins (18 characteristics, 3 NSE metrics, 2 groups = 78 tests).

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 4 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.

4 Results

4.1 NNSE

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

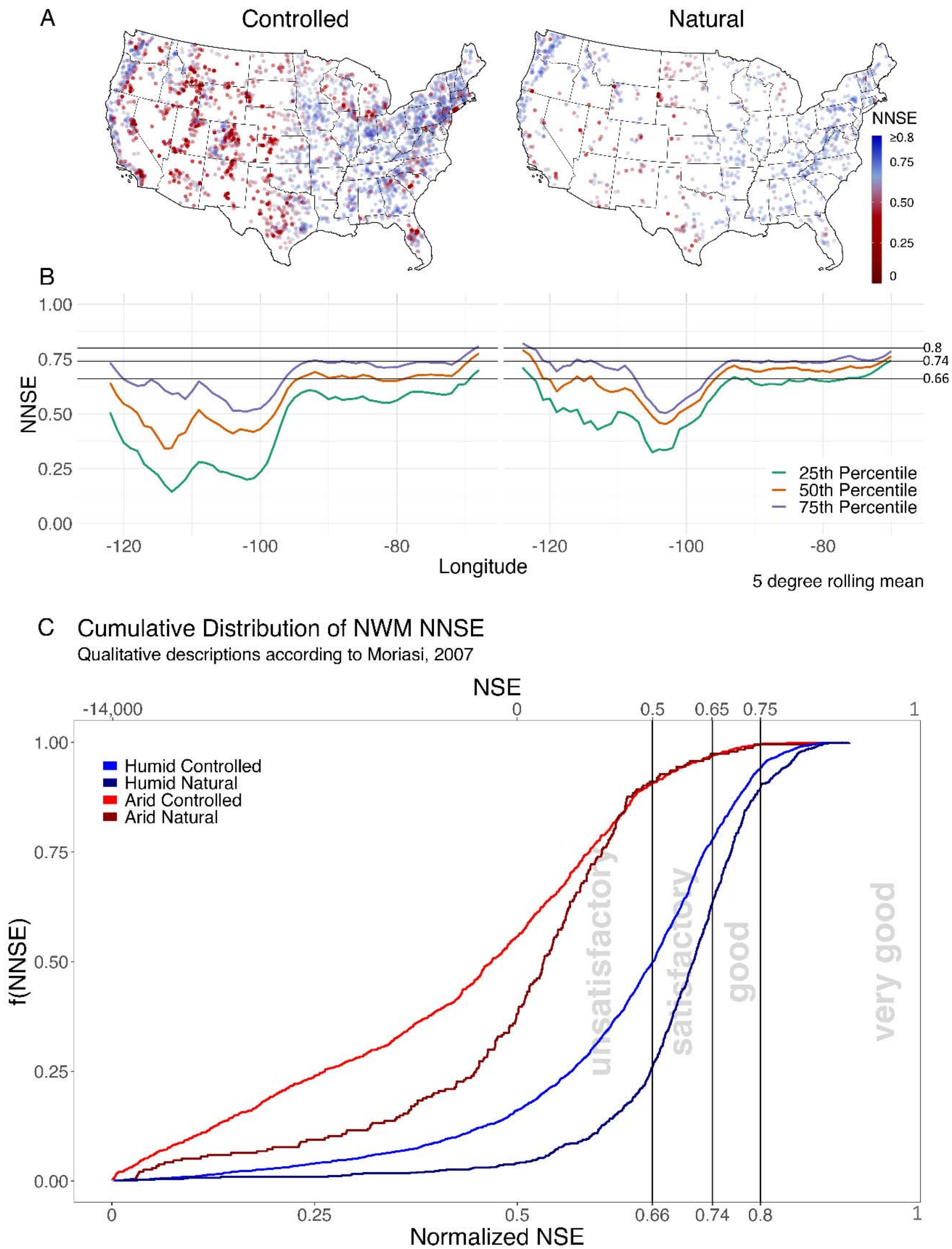

Figure 2: (A) NNSE mapped by gage location, the midway color aligns with "Satisfactory" performance. **(**B) Shows the 25th, 50th, and 75th percentile NNSE for each band of longitude smoothed with a 5-degree rolling mean. The horizontal lines at NNSE = 0.66, 0.74, and 0.80 represent the categories of "Unsatisfactory", "Satisfactory", "Good" and "Very Good". **(**C) NNSE distributions grouped by aridity and GAGES-II classification. The vertical lines represent the same qualitative groupings as panel B. Here red curves represent arid basins (AI > 2), and blue curves represent humid basins (AI < 2).

Figure 2A maps the NWIS gauges, split by classifications, and colored by NNSE. The color ramp ranges from red to blue with a midpoint at 0.66 following the mark of "satisfactory" NNSE. On the left, the control basins show strong performance in the northeast, east, and south but weak performance west of the 100th meridian. The exception to this is along the western side of the Sierra Nevada Range and the Central Valley where the Aridity Index is lower than the west at large. In the controlled basins there is a qualitative impact of cities on NWM performance with low skill surrounding the Orlando, Charlotte, New York, Detroit, Chicago, and Nashville metropolitan areas in the otherwise well performing east. In the humid west, the California Bay Area and Portland also underperform compared to their surroundings.

The natural basins show a more consistent performance east of the 100th meridian, even in the areas near large metropolitans. West of the 100th meridian, model performance begins to degrade, albeit to a lesser extent than seen in controlled basins. Other research efforts have noted the 100th meridian is a non-permanent divide splitting the continent into an "arid west" and a "humid east", defined in terms of vegetation, hydrology, crops, and farm economy (Seager et al., 2017). Our results is also constant with the evaluation of LSM driven streamflow by Cai (2015) that showed LSMs have difficulty representing streamflow in the north central region of the country but that "most models perform well east of the 95th meridian".

Figure 2B illustrates this longitudinal impact and plots the 25th, 50th, and 75th percentile NNSE, grouped by whole-degree longitude bands and smoothed with a 5-degree rolling mean. The horizontal bars illustrate the "unsatisfactory", "satisfactory", "good", and "very good" marks for NNSE according to Moraisi (2007). In all basins there are clear systematic drops in performance between the 105W and 95W meridians. When looking at just the control basins, the 50th percentile of locations achieve "satisfactory" performance while west of the 95th meridian even the 75th percentile even drops well below this mark. Not only does performance drop, but the variability increases as evident by the spread between the 25th and 75th percentiles. There is a slight recovery in performance starting around the 115th meridian, however variability remains large.

When looking at 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.

In Figure 2C, the Empirical Cumulative Distribution Function (ECDF) of NNSE grouped by basin and aridity classification is shown. In this plot, the ideal would be a curve that stays as low as possible on the y-axis for as far as possible along the x-axis. Humid basins outperform arid basins, and natural basins outperform controlled basins and the difference between controlled and natural classification is more notable 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) have unsatisfactory performance and in those with satisfactory or better performance, the distinction between natural and controlled is non-existent.

4.2 NSE-A: Relative Performance

NSE-A (r^2; Relative performance) is the correlation between the observed and simulated streamflow values. The relative performance values are mapped in Figure 3A colored by their relative performance value while panel B plots the 25th, 50th, and 75th percentile relative performance, grouped by whole-degree longitude bands and smoothed with a 5-degree rolling mean. With respect to correlation, the NWM performs better in the eastern part of the CONUS and along the west coast. The variability in relative performance is greater in the west than the east, except for natural basins in the humid west coast. Across CONUS, the variation in controlled basins is greater than in natural basins, but aside from the variability in performance, the pattern of the longitudinal profiles for natural and controlled basins are largely the same. Using the catchment characteristics identified in Table S1, 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 ($p < .05$) and practically ($\eta 2 > .01$) significant result are shown in Figure 3C.

2.2.1 Meterological Characteristics

The dominant catchment characteristic in relative performance is aridity (Figure 3Ca). As aridity increases, relative performance decreases across all basin types. The effects size suggests 45% of the variance in relative performance can be explained by the aridity of a basin. In both basin types we see relative performance decreases by a factor of 2 when comparing very arid to very humid basins. The second and fourth most dominant characteristics in understanding relative performance are PPT (Figure 3Cb), and PET (Figure 3Cd). Naturally these are highly correlated with aridity, however evaluating them independently shows that as basins experience more rainfall, 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 nearly linear pattern across groupings, the middle two sections of PET hover around the mean relative performance. This suggests that only "extreme" low PET or "extreme" high PET impact performance. In all but basins with very high PET, natural basins perform better than controlled basins. The pattern for mean correlation (Figure 3Ce) is very similar to that of PET highlighting that when the phase correlation is between (-.4 and .4) there is limited impact to relative performance but when the correlation is significantly negative (< -0.4) (out of phase), the NWM improves (particularly in natural basin). Conversely, as the correlation becomes more positive (energy and moisture in phase), NWM performance degrades. With respect to overall variance in relative performance, PPT explains 31%, PET 18%, and Mean Correlation 13%. Lastly, as mean snow coverage (Figure 3Ch) increases, so does the general performance.

When mean annual snow is between 0-10cm, the relative performance across all basins is near the overall mean. As snow increases, relative performance is seen to improve, a pattern that is more prominent in natural basins. In a broad sense, 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.

2.2.2 Landscape Characteristics

As the percentage of barren land (Figure 3Cc) increases, so does NWM performance. This is particularly true in natural basins. The effect size of 20% highlights the significance of this value. 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, performance begins to decline.

As forest (Figure 3Cf) and shrubland (Figure 3Cg) increase, the model performance decreases. Forests and shrublands are those with significant biomass that respond differently based on season and location impacting both PET and actual ET. In other words, the more a basin is covered by spatially and temporally heterogeneous processes (represented homogeneously in the model), the worse overall performance will be. This pattern is also evident in the herbaceous land cover (grasslands, Figure 3Ci) however the effect is smaller, and the pattern is quite different when looking at controlled vs natural basins.

Lastly, agriculture (Figure 3Ck) shows a counter-intuitive pattern. General theory would suspect that agriculture would increase the non-natural hydrology of a drainage basin via irrigation that brings the energy-moisture phases more into sync (trying to align PET to Actual evapotranspiration (AET) during the growing season). However, the results suggest that in controlled basins, once 15% of the basin is deemed agricultural, performance begins to improve compared to the mean relative performance.

Overall, 11 characteristics were statistically and practically significant in describing the variation in relative performance, of these Aridity, PPT, PET, an phase correlation were meteorological factors with more medium or greater effect size while barren, forest, shrubland were the landscape features with a medium or larger effect size.

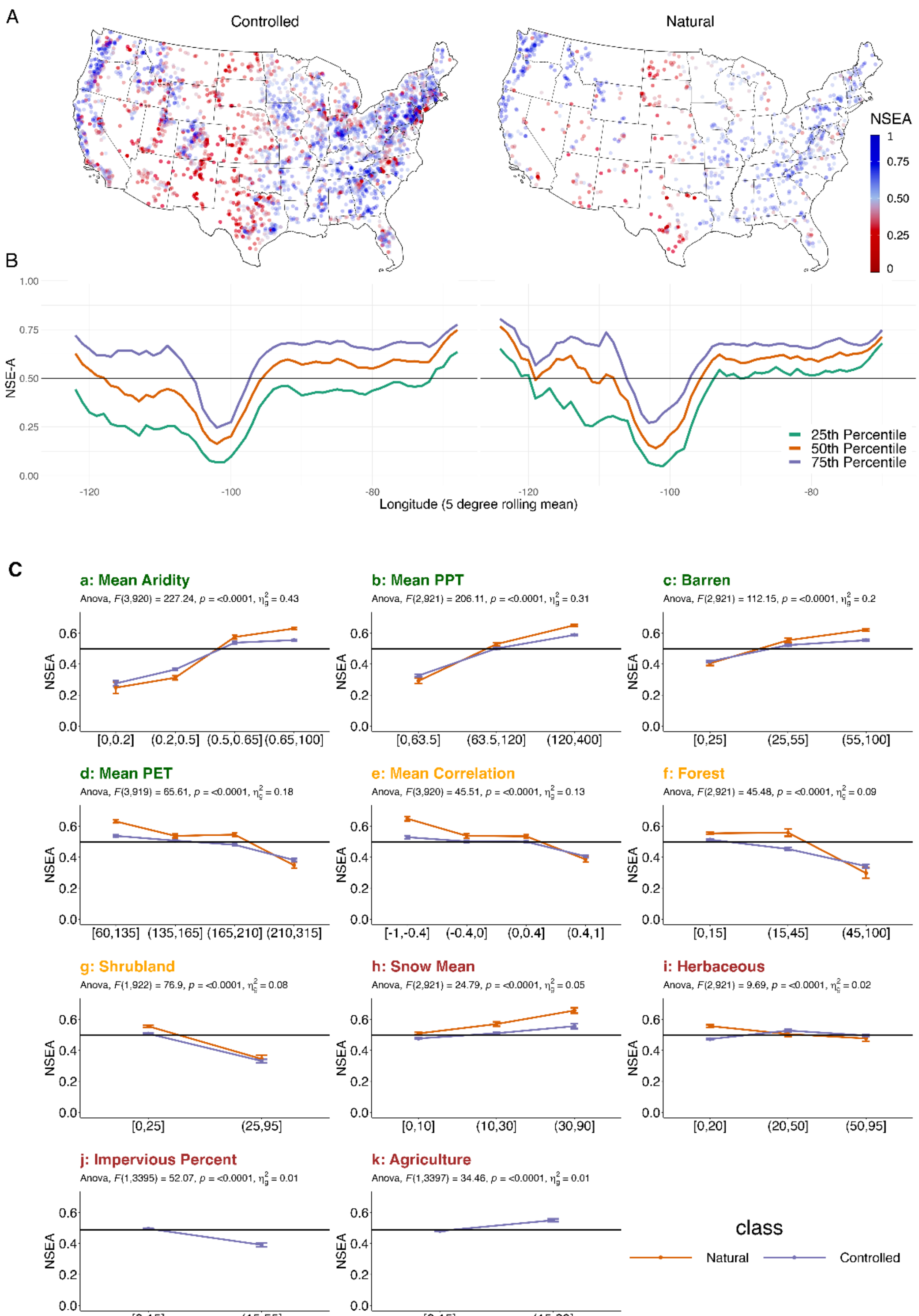

Figure 3: (A) NSE-A split by natural and controlled basins. **(B)** 25th, 50th, and 75th percentile NSE-A for each band of longitude smoothed with a 5-degree rolling mean. **(C)** Mean NSE-A is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically ($p >$ .05) and practically ($\eta 2 > .01$) 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.

4.3 NSE-B: Conditional Bias

When comparing the NNSE (Figure 2B) and NSE-A longitudinal plots (Figure 2B & 3B), NSE-A is U-shaped, showing model performance recovery west of the 100th meridian, while the NNSE plots do not. 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 (conditional bias) for the natural and controlled basins. In these, conditional bias values are truncated to 1.0, meaning anything listed as 1.0 is ≥ 1.0 and the number of truncated sites is listed in the subtitle of each plot. Beneath each map is a longitudinal average smoothed with a 5-degree rolling mean developed in the same way as section 4.1.

Larger conditional bias values occur in the arid west and the longitudinal percentile plots indicate the amount and variability of conditional bias is nearly zero in natural basins east of the 100th meridian and less than 0.15 in controlled basins. In all basins, conditional bias spikes between the 95th and 105 meridians. The natural basins show model recovery (less conditional bias) west of the 110th meridian (the Rocky Mountains). In contrast, the controlled basins don't recover - and even increase - until the humid west coast is reached. In all basins, variability and conditional bias is 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 give some caution to where the NWM can be applied in the contexts of flood forecasting (its primary, initial purpose) and speaks to the needs 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 as it indicates minimal conditional bias. Across the board, conditional bias is lower in the natural basins, but all basins demonstrate the same patterns.

4.3.1 Meterological Characteristics

Starting with 4Ca and 4Cb, dry (PPT< 63.5cm), arid (AI>3) basins have larger than average conditional bias while wet (PPT >12cm), 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. In these cases, average conditional bias almost doubles. Inversely, mean phase correlation is significant in basins that are notably out of phase (<-0.4) where conditional bias increases by a factor of 1.5. Overall PPT, AI, PET, and correlation explain 15%, 12%, 2% and 2% of variance in conditional bias respectively.

While more snow improves relative performance for all basins, it results in greater conditional bias in controlled basins highlights the challenges of modeling diverse snow processes (Figure 4Cj). This could also be a product of the primary functions of local reservoir as those in snowy basins may be designed to store runoff and snowmelt for the dry season. Snow Fraction (Figure 4Cd) also influences conditional bias suggesting that the more of a basin that is covered, the more conditional bias can be expected. There is a difference between the natural and controlled basins here in that even at high snow levels of snow coverage, natural basins exhibit average conditional bias. In contrast, conditional bias increases in an almost exponential pattern as snow coverage increases in controlled basins.

4.3.2 Landscape Characteristics

With respect to land cover, forest (Figure 4Cc) is only influential predictor in controlled basins. When forest coverage is <15% conditional bias is 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 than the other factors present in Figure 4C and is only influential in natural basins suggesting conditional bias decreases with increasing 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% respectively leading to almost 1.5 times increase in conditional bias.

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

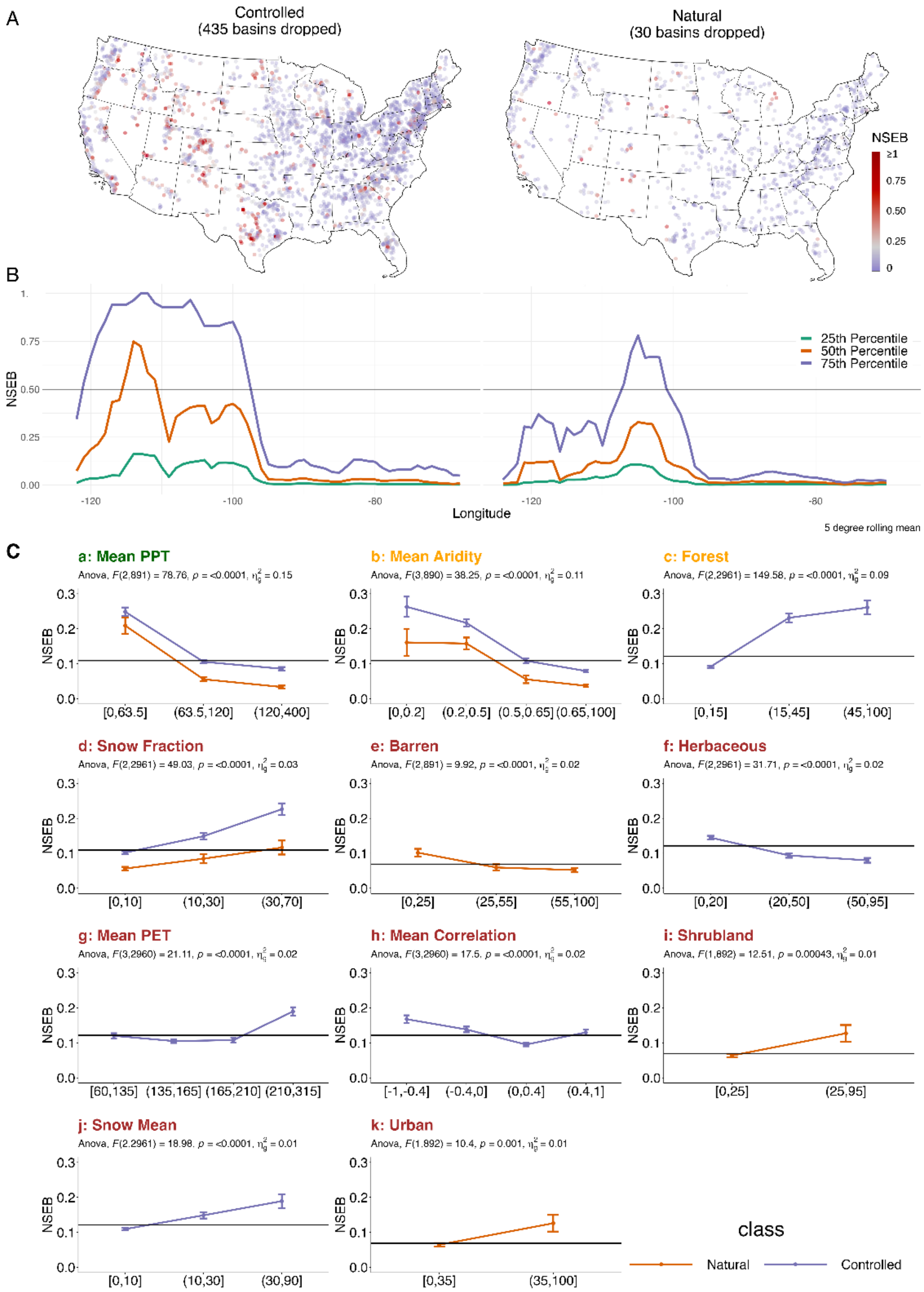

Figure 4: (A) NSE-B split by natural and controlled basins. (B) 25th, 50th, and 75th percentile NSE-B for each band of longitude smoothed with a 5-degree rolling mean. (C) Mean NSE-B is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically (p > .05) and practically (η2 > .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.

4.4 NSE-C: Unconditional Bias

Figure 5A maps NSE-C (unconditional bias) for the natural and controlled basins where unconditional bias 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-degree 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 basin is lower than in controlled basis and land cover impacts controlled basins while meteorologic properties influence all basins. When compared to the population mean (horizontal bar), natural basins exhibit significantly less unconditional bias than the controlled basins.

4.4.1 Meterological Characteristics

As aridity (Figure 5Cc) and snow fraction (Figure 5Cd) increase, so does unconditional bias. Equally as PPT (Figure 5Cb) and phase correlation (Figure 5Cj; only in controlled basins) increase, unconditional bias decreases. 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 control 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 (Figure 5Ci) where unconditional bias in controlled basins increases in a linear pattern 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 sits in controlled basins where moisture and energy cycles are out of sync, and that have low PPT, high aridity, and high snow coverage (mean and fraction).

4.4.1 Landscape Characteristics

In natural basins, urban (Figure 5Ch) and barren (Figure 5Ce) land cover are the only influential types. NSE-C associated with urban coverage increases by a factor of 2 when more than 35% of the basin is urbanized. These are likely basins that have likely urbanized post GAGES-II classification. In contrast, increasing barren land cover (Figure 5Ce) results in decreased unconditional bias in all basin types. In controlled basins, impervious surface (Figure 5Cg), and forest (Figure 5Ca) and herbaceous (Figure 5Cf) land cover are impactful. Unconditional bias is larger (factor of 2) in basins that are more than 15% impervious/forested but unconditional bias decreases when grass coverage exceeds 20%.

Overall, 10 characteristics were statistically and practically significant in describing the variation in unconditional bias, of these forest coverage was the only factor with a medium or larger effect size. In sum, 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.

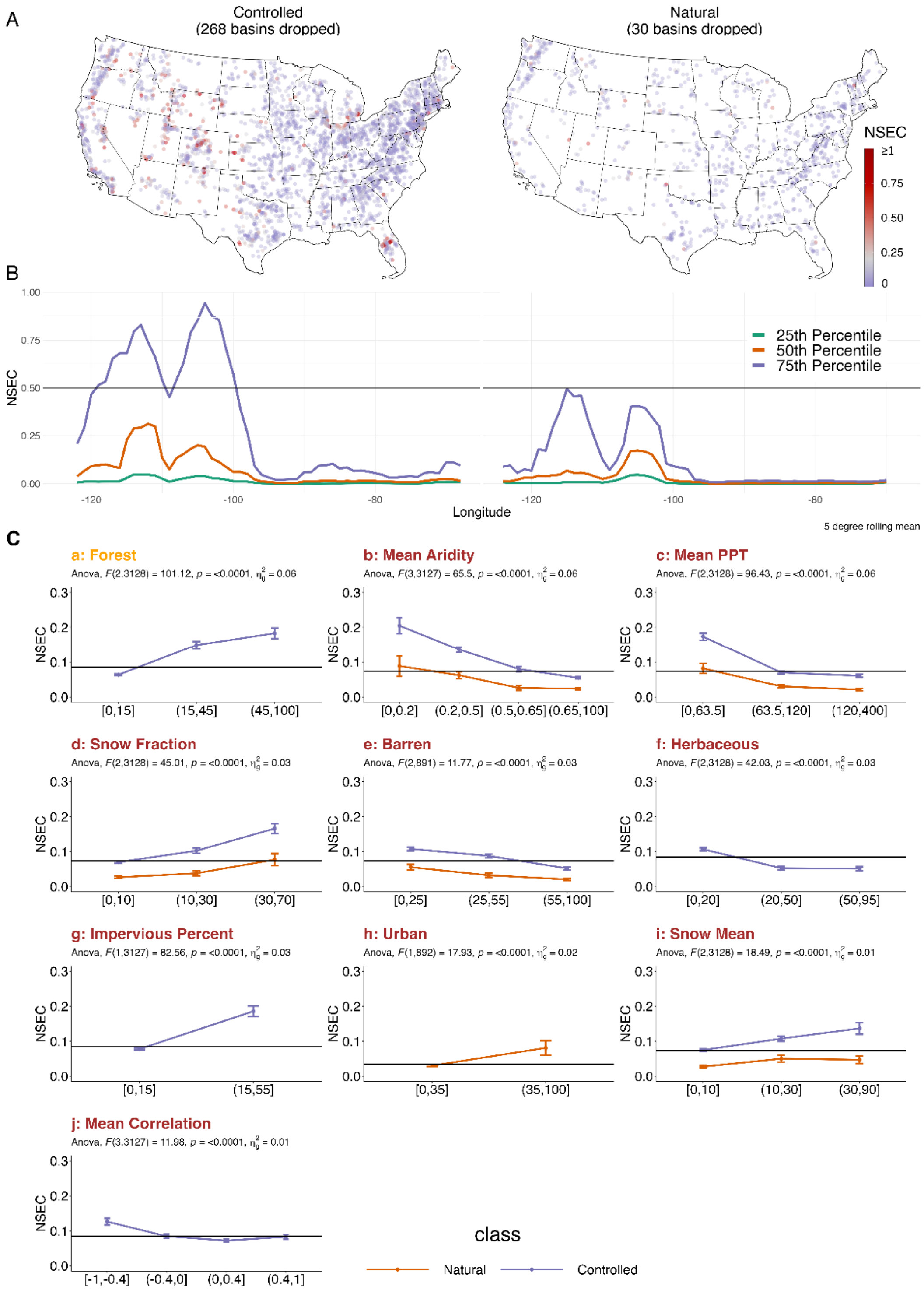

Figure 5: (A) NSE-C split by natural and controlled basins. (B) 25th, 50th, and 75th percentile NSE-C for each band of longitude smoothed with a 5-degree rolling mean. (C) Mean NSE-C is plotted by catchment characteristics grouped according to Jenks optimization and classified by basin type. Only relationships that were statistically (p > .05) and practically (η2 > .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.

5 Discussion

The WRF-Hydro based National Water Model provides a continental-scale modeling framework that integrates an operational forcing model, a high-resolution land surface model and a high-resolution overland flow and channel routing module. The resolution of each of these components, paired with the geographic extent, make this the only operational model of its class. While the NWM provides valuable information for water resources decision makers, it is important to understand its limitations in terms of performance. The historic product provides 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 the evolution of NextGen.

This research focused on evaluating the NWM 2.0 performance in controlled and natural basins, in the contexts of catchment characteristics. A broad summary of model performance and bias, as well as the catchment characteristics responsible for such performance is shown in Table 1. This table outlines the role of different catchment characteristics on model performance and bias.

Table 1: Significant catchment characteristics and their impact on model performance and bias. In each cell, the direction of influence and 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.

	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
	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
	Agriculture	↑ *controlled*			performance increases in controlled basins

Through this work, we laid the groundwork for evaluating future versions of the model, identifying regions with high error, and better understanding the catchment characteristics responsible for the degraded performance. In the remainder of this discussion we synthesize advocate for a central set of catchment characteristics aligned to national hydrofabric products; illustrate the role of these results (and those like them) in model selection within the NextGen framework to reduce NSE-A; and discuss the capacity for post processing model applications to help reduce NSE-B and NSE-C.

5.1 A Need for Catchment Characterization

In small domains, collecting, summarizing, and defining spatial data is a relatively straightforward task. However, scaling this process to a domain like CONUS, and managing the datasets and processes to achieve accurate and useful results can be challenging. A large portion of the work in this study was choosing and processing large scale (both in space and time) data products to draw conclusions about hydro-meteorological and landscape driven performance of the NWM. Providing consistent representations of different catchment characteristics can reduce the specialized geospatial expertise needed to acquire basin characteristics and expedite research into landscape function and representation in a modeling context. In the US, some efforts like the EPA STREAMCATS (Weber, 2017), and multiple USGS products, have developed reference catchment characteristics over the National Hydrography Dataset Plus V2 (NHDPlusV2, see McKay et al 2015 for more information) to provide continuous and comprehensive catchment characteristics for the USA. While these are powerful utilities when using the NHDPlus, they are one of workflows that are not updateable by the public and are tied to a single spatial representation of the landscape (hydrofabric).

There are efforts to provide authoritative continental (in the US, Bock, 2022, Johnson 2022) datasets grounded in emerging data standards for hydrologic science (Blodgett, 2018; Blodgett et al, 2020, 2022). The aim of these reference data products is to provide a consistent, high-resolution product that can be modified (upscaled) to meet the needs of different modeling applications such as NextGen and the USGS National Hydrologic Model. In the same way, a nationally consistent and comprehensive catchment characteristic dataset would aid in the ability to evaluate model performance and advance the efforts of the hydrologic community to provide more accurate and timely hydrologic prediction not only in gauged, but in ungauged locations.

5.2 A Role for Model Selection and post-processing

The results of this paper highlight areas the WRF-Hydro NWM model is underperforming as well as general reasons why. The principal driver in all basins was AI, PET, PPT, and forest coverage. While it is well known that arid regions are hard to simulate due to high non-linearities in soil dynamics that affect infiltration and evapotranspiration, the thresholds for when these processes become limiting is not fully understood. Further, while PPT, PET and AI are hydroclimatic indicators of model uncertainty they are not explicit model states.

One of the key model fluxes related to these characterstics is actual evapotranspiration (AET). In fact, within Noah-MP, water can only be removed from the system through runoff (streamflow) or AET. In basins with limited recharge to groundwater, an underestimation in runoff is directly linked to an overestimation in AET (and vice versa). Alternativly in basins with significant recharge, AET might be trying to compensate for poorly represented groundwater processes. The need to better represent this process in arid regions is a prime use case for the NextGen system.

To highlight this, we look at a natural arid basin in Nevada where total runoff is underestimated by the NWM 2.0 (Figure 6A, NWIS ID 10244950). This basin presents an aridity index of 3.86, mean annual PPT of 464.51mm, and mean snow depth of 194 mm covering 41% of the basins. For NWM 2.0, this basin presented a relative performance (NSE-A) of 0.45, a conditional bias (NSE-B) of 0.25 and an unconditional bias (NSE-C) of 7.3 indicating low accuracy and incredibly high bias.

As a first step, the NWM NextGen framework was used to simulate the basin over a five year period using the Conceptual Functional Equivalent model with the Xinanjiang rainfall-runoff partitioning module (https://github.com/NOAA-OWP/cfe). Six different PET methods were tested including: (1) Noah-OWP, which provides a pseudo-PET estimation assuming there is no moisture limitation (https://github.com/NOAA-OWP/noah-owp-modular), (2) energy balance, (3) aerodynamic, (4) combined, (5) Priestley-Taylor, and (6) Penman-Monteith methods (https://github.com/NOAA-OWP/evapotranspiration). To identify the “best” of these, the long term aridity index of the catchment was compared to the aridity index produced by each simulation.

In Figure 6B, the ratio of simulated AI to long-term AI is shown. Here, a ratio close to one indicates good agreement, while ratios smaller (larger) than one indicate NextGen CFE formulation underestimates (overestimates)

PET. For this basin, the aerodynamic method (green) produces the most consistent AI and critically we see the NOAA-OWP (effectively a modular NOAH-MP variant) significantly underpredicted PET which given the concepts of the model would explain the shortfall in streamflow seen in version 2.0.

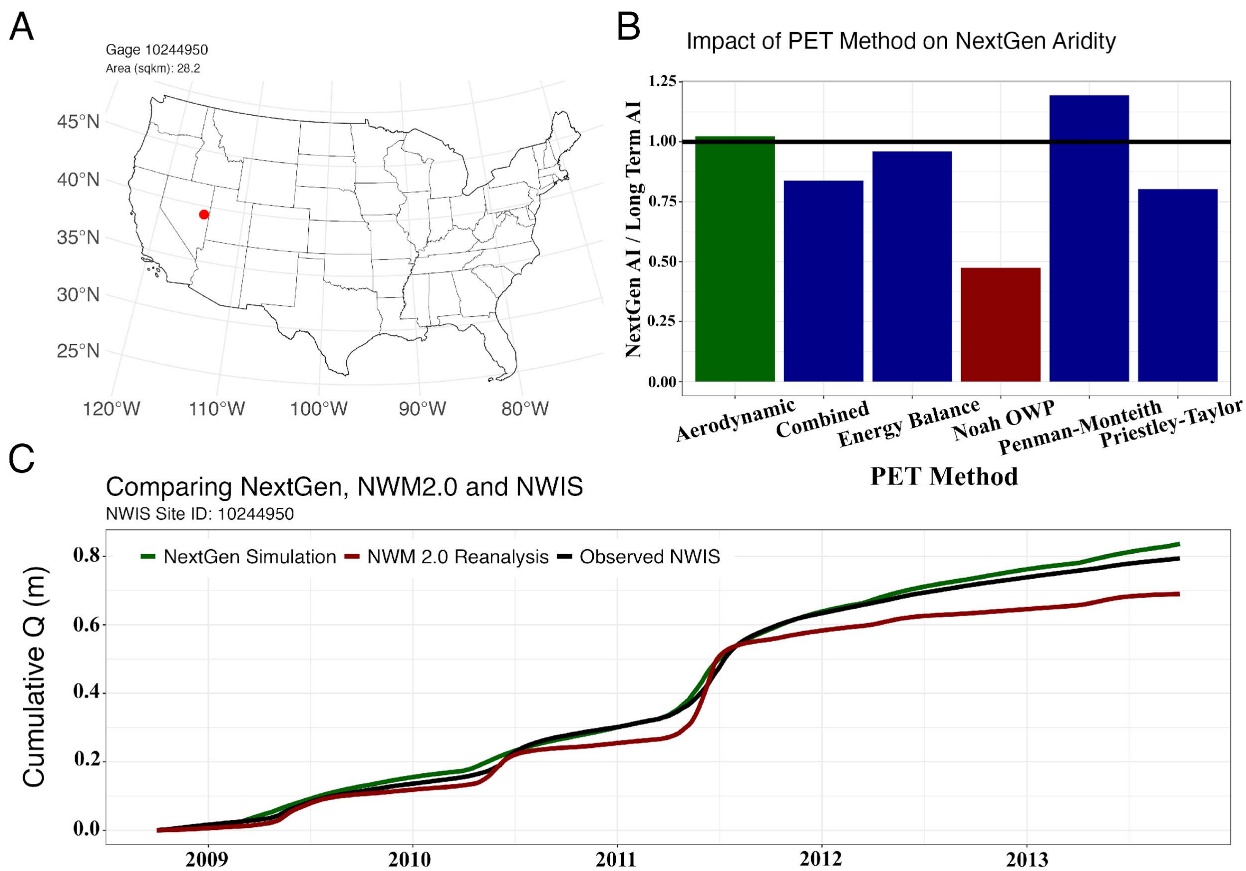

Figure 6: (A) A poor performing natural, arid basin in Nevada was selected. (B) 6 simulations were run using NextGen and the ratio of the simulated AI to the catchment AI was computed. The red bar approximates what was used in NWM2.0 while the ideal aerodynamic method (closest to 1) is in green. (C) Cumulative discharge plots of the USGS observations, NWM 2.0, and the aerodynamic NextGen simulation are shown highlighting the power of location driven processes.

Figure 6C plots the cumulative discharge over a five year period for the NWM2.0, observed USGS flows, and the location-specific PET NextGen simulation. In it we see the improved 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. Overall, this example highlights how an understanding of predominant hydroclimatic variables, paired with comprehensive catchments characteristics can support diagnostic model selection and lead to improved hydrologic prediction.

While improving model selection can improve all three elements of NSE, the maturity of NextGen (or any other heterogenous modeling system) is not fully developed meaning a single model formulation (e.g. LSM) must be used to cover continental scales. The calibrated model physics are responsible for achieving strong correlation and without refined, spatially appropriate, model formulations, it will be difficult – if not impossible – to improve poor performing areas without degrading the NSE-A in other locations.

That said, the model outputs can still be improved for applications and high quality predictions. This study showed significant conditional and unconditional biases in the NWM simulations particularly in the worst performing geographies. There are many available methods for reducing these biases without changing the NWM itself using statistical post-processing techniques (Sinha & Sankarasubramanian, 2013).

Particularly, this work identified various catchment characteristics and their role in explaining the spatio-temporal variability of streamflow at a continental scale. This understanding, and the groupings at which they are significant, provides an opportunity to apply statistical and or data-driven decision models to post process NWM output using time-varying, at-site hydrologic information (i.e., NWM predictions) alongside catchment characteristics (Frame et al., 2021; Ossandón, Rajagopalan, & Kleiber, 2021; Ossandón, Rajagopalan, Lall, et al., 2021). Equally it offers the opportunity to consider hierarchical modeling approaches based catchment traits, error characteristics, and external data sources (e.g. remote sensing) to not only improve flows in gaged basins, but to transfer that knowledge to ungaged locations. Until the NextGen reaches maturity, and likely beyond, the ability to generate hybrid approaches that take the best possible outflows and further refine them will be critical for improved prediction, modeling, and understanding.

5 Conclusions

The NWM offers an unprecedented step forward in the hydrologic forecasting capabilities of the United States. Its innovation is not only in the advancement of forecasting operations, but also in the development of an operational, near-real time, high resolution LSM with minimal lag and comparatively sophisticated routing. The implementation of this model, regardless of current results, sets the stage for achieving what has been dubbed the "grand challenge" in hydrologic modeling and, together, the infrastructure, personal, and agency forces driving the NWM means it will persist as an integral component of a national hydrologic forecasting infrastructure.

With this advancement however comes the need to evaluate and diagnose the model in ways that explain not only *how* the model is performing but *why* it is performing that way. To do this, there needs to be a comprehensive set of catchment characteristics that can be used to classify basin types in low and high dimensional space. The impacts of these types of analysis are the opportunities it offers to study the limitations of physical model process, identify better physical representations that can be applied heterogeneously, and to look for opportunities to assimilate new data sources, and postprocess output to supply better forecasts, for the appropriate reasons.

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. The approach itself is portable to other model development and intercomparison efforts and its application to the NWM v2.0 reanalysis data provides more transparency for the public and water managers who want to use NWM model outputs, as well as the research community interested in contributing to model improvement and use.

Acknowledgments

Funding provided by NSF through grants OIA #1937099 and #2033607.

Data

The GAGES-II dataset can be accessed at (https://water.usgs.gov/GIS/metadata/usgswrd/XML/gagesII_Sept2011.xml). All streamflow data can be accessed from the USGS NWIS portal (https://waterdata.usgs.gov/nwis) or the NWM reanalysis archives (Johnson et. al, 2020c). Land cover data is accessed from the Multi Resolution Land Characteristics Consortium (https://www.mrlc.gov/data) and NLDAS data by NASA EarthData GES DISC service (https://disc.gsfc.nasa.gov/datasets?keywords=NLDAS).

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.

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: Atmospheres*, *101*(D3), 7449–7459.

Adams III, T. (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). US 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.
Barlage, M. (2017, May 1). 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.
Blodgett, D., & 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.
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. *Computer Models of Watershed Hydrology*, 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.
Cai, X., Yang, Z.-L., Xia, Y., Huang, M., Wei, H., Leung, L. R., & Ek, M. B. (2015). Assessment of simulated water balance from Noah, Noah-MP, CLM, and VIC over CONUS using the NLDAS test bed, 1–20.
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.
Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Routledge.
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
D. N. Moriasi, J. G. Arnold, M. W. Van Liew, R. L. Bingner, R. D. Harmel, & T. L. Veith. (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
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). Reston, VA: U.S. Geological Survey / U.S. Geological Survey. https://doi.org/10.5066/P9X4L3GE
Falcone, J. A. (2011). *GAGES-II: Geospatial attributes of gages for evaluating streamflow*. US Geological Survey.
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. *JAWRA Journal of the American Water Resources Association*, *57*(6), 885–905.
Gochis, D. J., Yu, W., & Yates, D. (2013). The WRF-Hydro Model Technical Description and User's Guide, Version 1.0. *NCAR Tech. Doc*.
Gochis, J., & Chen, F. (2003). Hydrological Enhancements to the Community Noah Land Surface Model.
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.
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.
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.
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.
Johnson, J. M. (2020, March). nwmTools. Retrieved from https://github.com/mikejohnson51/nwmTools/

Johnson, J. M., & Blodgett, D. L. (2020). *NOAA National Water Model Reanalysis Data at RENCI*. HydroShare. Retrieved from http://www.hydroshare.org/resource/89b0952512dd4b378dc5be8d2093310f

Johnson, J. M., & Clarke, K. C. (2021). An area preserving method for improved categorical raster resampling. *Cartography and Geographic Information Science*, 1–13.

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*.

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.

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

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).

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.

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.

Liu, Y.-C., Liu, M., & Wang, X.-L. (2012). Application of self-organizing maps in text clustering: a review. In *Applications of Self-Organizing Maps*. IntechOpen.

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.

Maidment, D. R. (2016). Conceptual Framework for the National Flood Interoperability Experiment. *JAWRA Journal of the American Water Resources Association*, *53*(2), 245–257.

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: Atmospheres*, *106*(D16), 17841–17862.

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.

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), 11,809-11,825. 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.

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.

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.

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.

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), 1381–19.

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. *EGUGA*, 237.

Office of Water Prediction, N. (2022). The National Water Model. Retrieved from https://www.weather.gov/media/owp/oh/docs/2021-OWP-NWM-NextGen-Framework.pdf

Ossandón, Á., Rajagopalan, B., Lall, U., Nanditha, J., & Mishra, V. (2021). A bayesian hierarchical network model for daily streamflow ensemble forecasting. *Water Resources Research*, *57*(9), e2021WR029920.

Ossandón, Á., Rajagopalan, B., & Kleiber, W. (2021). Spatial-temporal multivariate semi-Bayesian hierarchical framework for extreme precipitation frequency analysis. *Journal of Hydrology*, *600*, 126499.

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.

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.

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.

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.

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.

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), 10078–21.

Sankarasubramanian, A., & Vogel, R. M. (2002). Annual hydroclimatology of the United States. *Water Resources Research*, *38*(6), 19–1.

Sinha, T., & Sankarasubramanian, A. (2013). Role of climate forecasts and initial conditions in developing streamflow and soil moisture forecasts in a rainfall--runoff regime. *Hydrology & Earth System Sciences*, *17*(2).

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

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.

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.

Weglarczyk, S. (1998). The Interdependence and Applicability of Some Statistical Quality Measures for Hydrological Models. *Journal of Hydrology*, *206*(1–2), 98–103.

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.

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), 54–10.

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.

Yang, Z.-L., Niu, G.-Y., Mitchell, K. E., Chen, F., Ek, M. B., Barlage, M., et al. (2011). The community Noah land surface model with multiparameterization options (Noah-MP): 2. Evaluation over global river basins. *Journal of Geophysical Research*, *116*(D12), 4257–16.