Improved National-Scale Flood Prediction for Gauged and Ungauged Basins using a Spatio-temporal Hierarchical Model

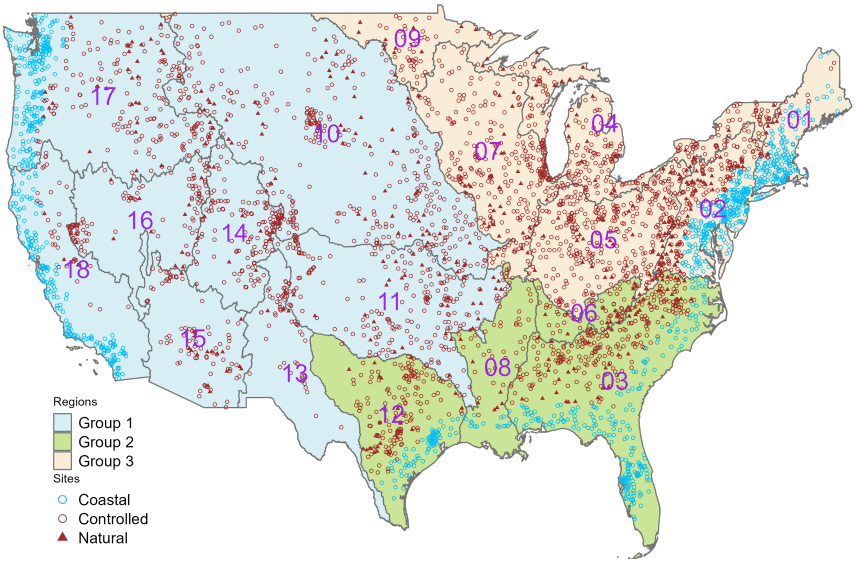
Shiqi Fang[1], J. Michael Johnson[2], Lilit Yeghiazarian[3], and Arumugam Sankarasubramanian[4]

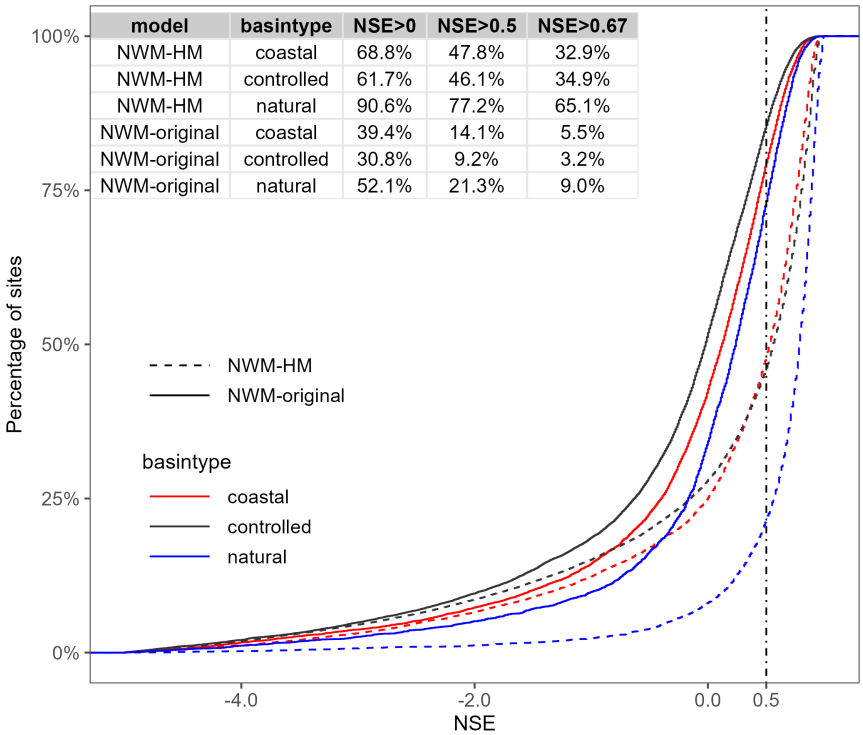
[1]North Carolina State University
[2]Lynker
[3]University of Cincinnati
[4]Department of Civil, Construction, and Environmental Engineering, North Carolina State University

February 9, 2023

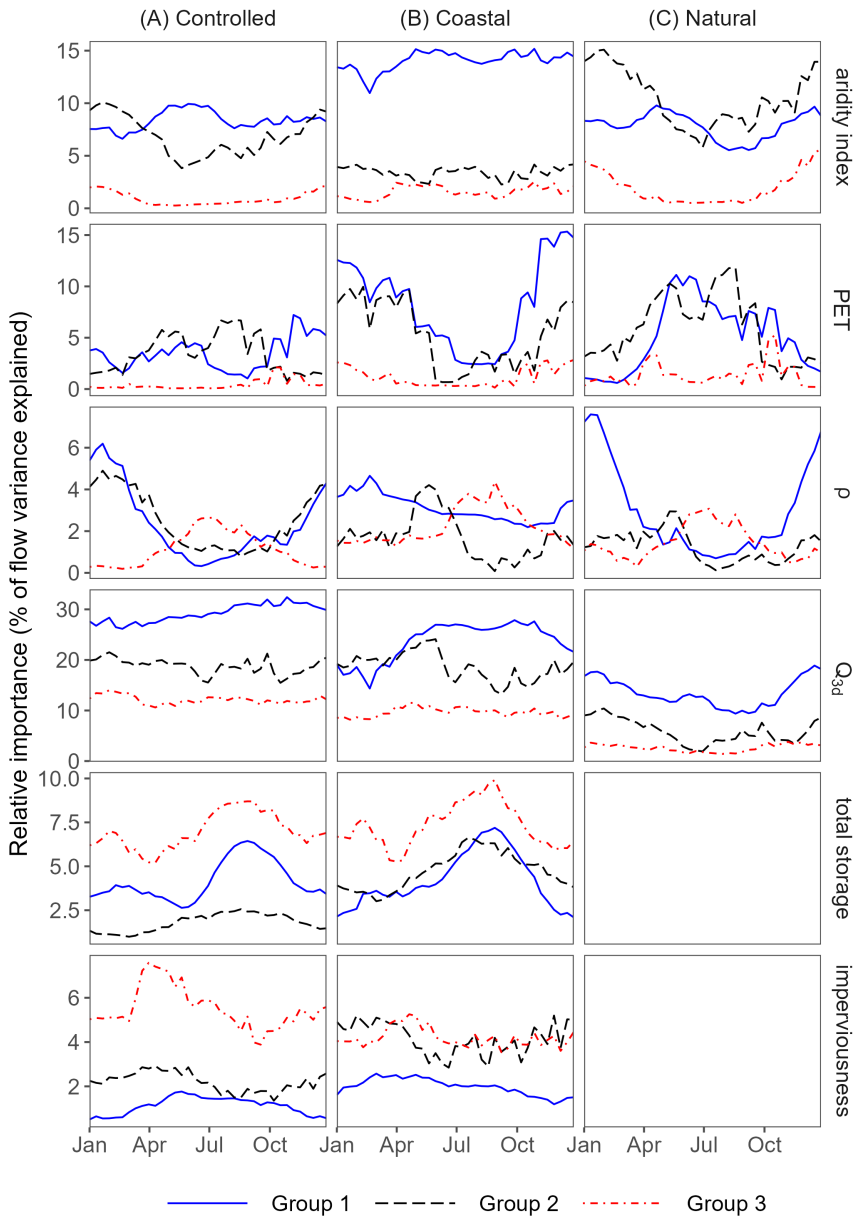
Abstract

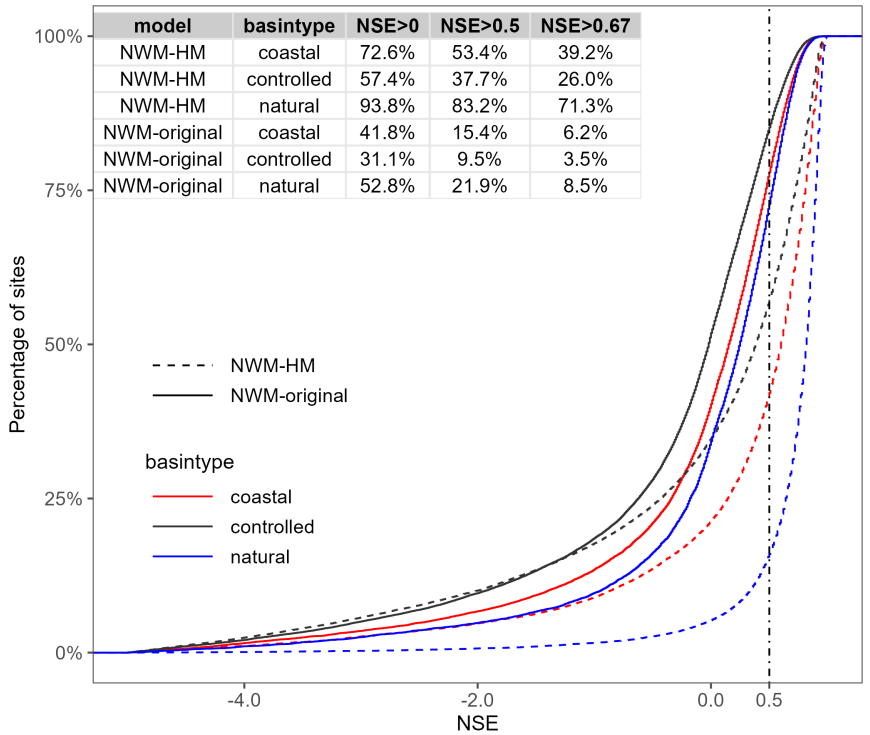
Floods cause hundreds of fatalities and billions of dollars of economic loss each year in the United States. To mitigate these damages, accurate flood prediction is needed for issuing early warnings to the public. This situation is exacerbated in larger model domains for high flows, particularly in ungauged basins. To improve flood prediction for both gauged and ungauged basins, we propose a spatio-temporal hierarchical model (STHM) to improve high flow estimation using a 10-day window of modeled National Water Model (NWM) streamflow and a variety of catchment characteristics as input. The STHM is calibrated (1993-2008) and validated (2009-2018) in controlled, natural, and coastal basins over three broad groups, and shows significant improvement for the first two basin types. A seasonal analysis shows the most influential predictors are the previous 3-day average streamflow and the aridity index for controlled and natural basins, respectively. To evaluate the STHM in improving streamflow in ungauged basins, 20-fold cross-validation is performed by leaving 5% of sites. Results show that the STHM increases predictive skill in over 50% of sites by 0.1 Nash-Sutcliffe efficiency (NSE) and improves over 65% of sites' streamflow prediction to an NSE>0.67, which demonstrates that the STHM is one of the first of its kind and could be employed for flood prediction in both gauged and ungauged basins.



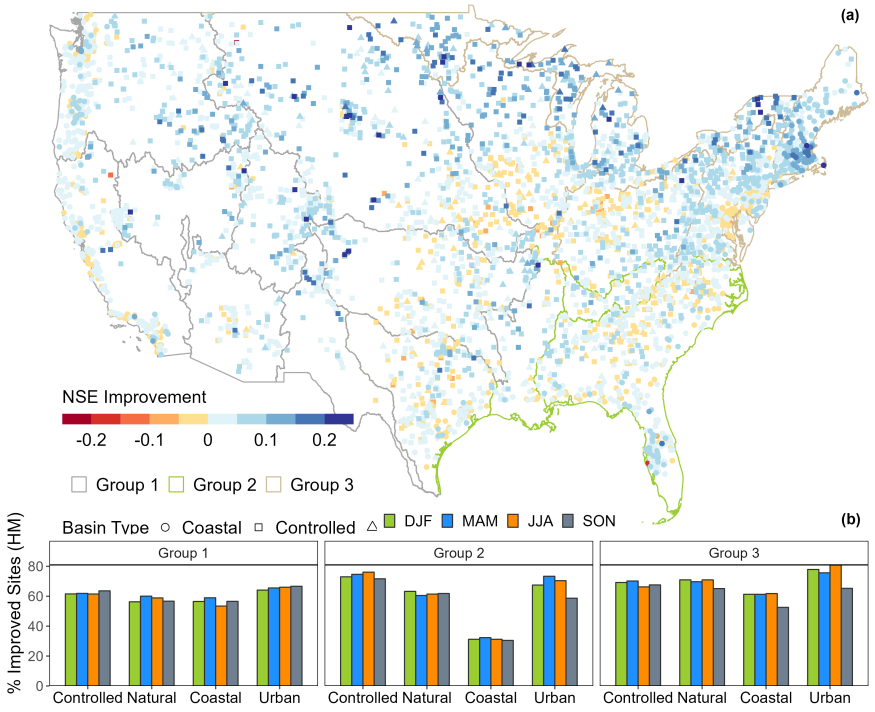


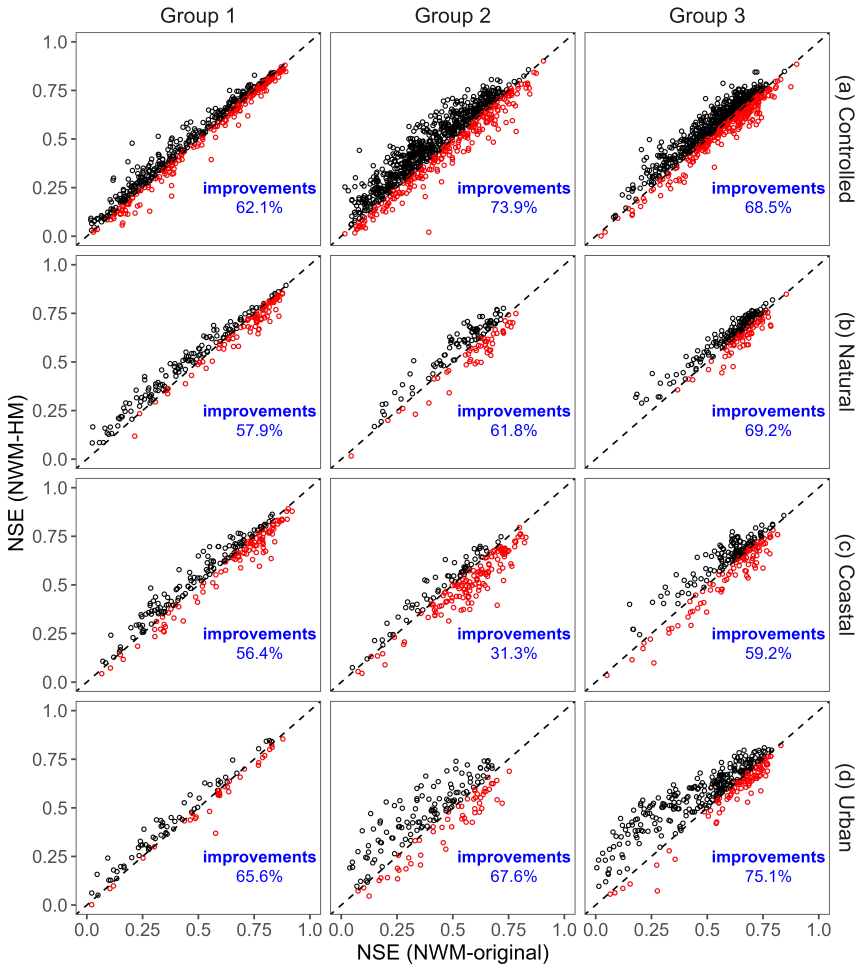
model	basintype	NSE>0	NSE>0.5	NSE>0.67
NWM-HM	coastal	68.8%	47.8%	32.9%
NWM-HM	controlled	61.7%	46.1%	34.9%
NWM-HM	natural	90.6%	77.2%	65.1%
NWM-original	coastal	39.4%	14.1%	5.5%
NWM-original	controlled	30.8%	9.2%	3.2%
NWM-original	natural	52.1%	21.3%	9.0%

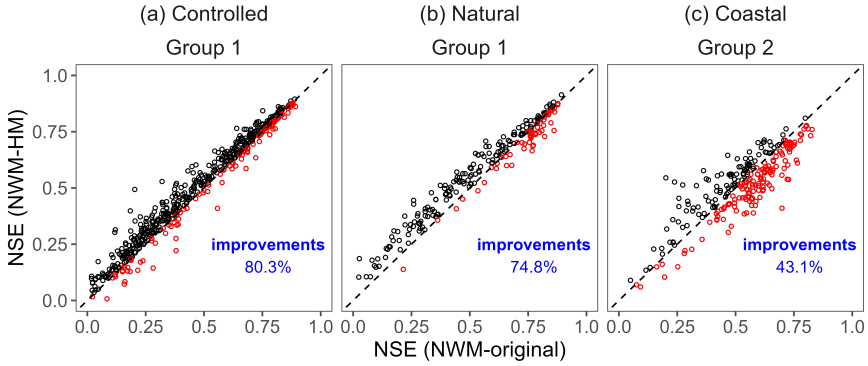


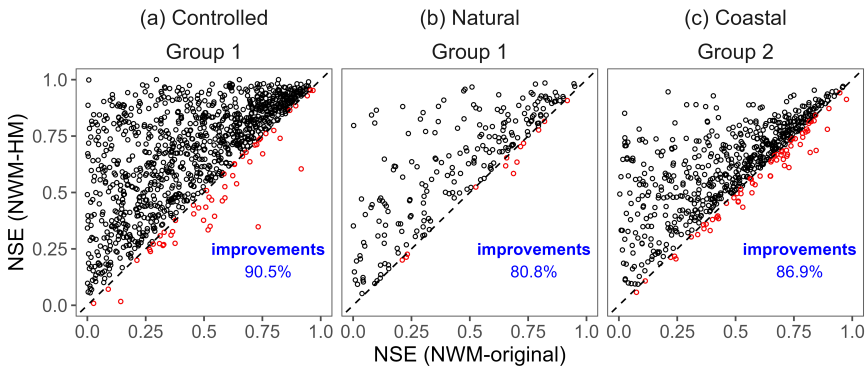


model	basintype	NSE>0	NSE>0.5	NSE>0.67
NWM-HM	coastal	72.6%	53.4%	39.2%
NWM-HM	controlled	57.4%	37.7%	26.0%
NWM-HM	natural	93.8%	83.2%	71.3%
NWM-original	coastal	41.8%	15.4%	6.2%
NWM-original	controlled	31.1%	9.5%	3.5%
NWM-original	natural	52.8%	21.9%	8.5%









Improved National-Scale Flood Prediction for Gauged and Ungauged Basins using a Spatio-temporal Hierarchical Model

Shiqi Fang[1], J. Michael Johnson[2], Lilit Yeghiazarian[3], A. Sankarasubramanian[1]

[1]Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh NC, USA
[2]Lynker, Fort Collins CO, USA
[3]Department of Biomedical Chemical and Environmental Engineering, University of Cincinnati, Cincinnati Ohio, USA

Key Points:

- We present a space-time hierarchical modeling framework to predict flood
- The model uses antecedent streamflow prediction and hydroclimatic data to show major improvements in various basin types across CONUS
- The method is shown to have good skill in making prediction in ungauged basins providing enhanced flood predictions for the country

Corresponding author: Shiqi Fang, `sfang6@ncsu.edu`

Abstract

Floods cause hundreds of fatalities and billions of dollars of economic loss each year in the United States. To mitigate these damages, accurate flood prediction is needed for issuing early warnings to the public. This situation is exacerbated in larger model domains for high flows, particularly in ungauged basins. To improve flood prediction for both gauged and ungauged basins, we propose a spatio-temporal hierarchical model (STHM) to improve high flow estimation using a 10-day window of modeled National Water Model (NWM) streamflow and a variety of catchment characteristics as input. The STHM is calibrated (1993-2008) and validated (2009-2018) in controlled, natural, and coastal basins over three broad groups, and shows significant improvement for the first two basin types. A seasonal analysis shows the most influential predictors are the previous 3-day average streamflow and the aridity index for controlled and natural basins, respectively. To evaluate the STHM in improving streamflow in ungauged basins, 20-fold cross-validation is performed by leaving 5% of sites. Results show that the STHM increases predictive skill in over 50% of sites' by 0.1 Nash-Sutcliffe efficiency (NSE) and improves over 65% of sites' streamflow prediction to an NSE>0.67, which demonstrates that the STHM is one of the first of its kind and could be employed for flood prediction in both gauged and ungauged basins.

1 Introduction

Flood, causing major disruptions in urban and rural areas and threatening lives, is one of the deadliest and costliest hazards in the United States (National Weather Service, 2022). Extreme flood events severely impact infrastructure and environment (Merz et al., 2021; Wallemacq & House, 2018). Both global climate change and local anthropogenic activities have also been exacerbating these extreme events, particularly in urbanized coastal areas (Arnone et al., 2018; Pörtner et al., 2022). Providing accurate flood predictions is one key mechanism to reduce economic impacts and loss of life by providing early warnings for the public and agencies to set up contingency measures and evacuation warnings (Emerton et al., 2016; Johnson et al., 2018). The value of flood predictions are quite significant in providing early warnings and in developing evaluation strategies, thereby providing indirect benefits to society (Pappenberger et al., 2015).

Flood predictions are typically of two types – deterministic and probabilistic – with the former providing the conditional mean (Sinha & Sankarasubramanian, 2013) and the latter providing the entire conditional distribution (Sankarasubramanian & Lall, 2003). Deterministic predictions are easier to communicate as it does not provide uncertainty in the outcome, but probabilistic predictions are more useful in relating the uncertainty to different degrees of contingency measures (Cloke & Pappenberger, 2009). Flood predictions are typically developed either physical-process-based modeling approach – forcing precipitation forecast into a hydrological model – or using a statistical model in which precipitation forecast and antecedent watershed conditions are related directly to the observed streamflow. Uncertainty in flood predictions arises from multiple sources that include input uncertainty from precipitation predictions, hydrologic model uncertainty, and uncertainty in quantifying the initial conditions (Mazrooei et al., 2021; Mendoza et al., 2012). Efforts have focused on reducing these uncertainties ranging from multi-model combination (Devineni et al., 2008) on precipitation predictions, on hydrologic models (Singh & Sankarasubramanian, 2014), and through correcting initial conditions through data assimilation (Mazrooei et al., 2020). However, most of these uncertainty reduction techniques have focused primarily on gauged basins with limited/no evaluation of these techniques for ungauged basins.

The main intent of this research is to develop a unified approach that corrects the errors in flood predictions for both gauged and ungauged locations over the Coterminous United States (CONUS). Predicting streamflow in ungauged basins (PUB) is an established area of research in hydrology (Hrachowitz et al., 2013). Correcting hydrological errors in ungauged basins is challenging (Mishra & Coulibaly, 2009) as streamflow variability is unknown. Methods to address this challenge to date include spatial proximity methods (Tamaddun et al., 2019; Y. Zhang & Chiew, 2009), physical similarity approaches (Narbondo et al., 2020), data-driven methods such as artificial neural networks (Heng & Suetsugi, 2013), and nonlinear regression models (Parajka et al., 2013). These studies mainly focused on regional scaled or seasonal streamflow prediction, but flood forecast is typically required at daily-to-weekly time scale. Further, with regard to flood prediction, most studies relate the flow attributes available at gauged sites (e.g., 25-year return period flood) with the hydroclimate and basin characteristics to develop a statistical model and then use that relationship to estimate the corresponding flood values at based on basin and climate characteristics available at ungauged locations (see Table 3 in Salinas et al., 2013). However, most of these two-step approaches have focused primarily on design flood as opposed to predicting daily flood flows, which are critically important for issuing early warnings. Further, these two-step regression modeling can be effectively integrated into a single step using a hierarchical model (Das Bhowmik et al., 2020; Devineni et al., 2013). To our knowledge, limited/no application of hierarchical model has been performed for estimating daily flows at ungauged locations over the CONUS.

Hierarchical modeling framework (aka., multilevel models) is commonly used to combine time-varying hydrologic information (e.g., observed streamflow) with the spatially varying basin and hydroclimatic characteristics (e.g., Ossandón et al., 2022). Hierarchical model frameworks have the advantage of considering both spatio-temporal predictors and categorical (i.e., spatial or temporal) predictors for estimating a predictand (Gelman & Hill, 2006). Streamflow prediction studies that used hierarchical models used the river basin dendritic structure to predict the flows at a downstream location based on predictors such as basin-level meteorological variables and observed or hydrologic model predicted streamflow (Ossandón et al., 2022; Ravindranath et al., 2019). However, these studies have focused on predicting streamflow at gauged locations. Given that the hierarchical model is a spatiotemporal model with multi-level predictors, a hierarchical model could in principle be extended for predicting streamflow at ungauged locations by considering predicted streamflow available from any hydrologic model and other basin characteristics (e.g., aridity index) that commonly influence the error structure in hydrologic model prediction. Based on these underpinnings, we propose a novel hierarchical modeling structure that uses spatio-temporally varying observed/predicted streamflow information and spatially varying basin characteristics (e.g., drainage area) and hydroclimatic information (e.g., aridity index) for estimating flood flows at ungauged basins over the CONUS.

Continental-scale hydrology studies have evaluated parsimonious mechanistic models (Archfield et al., 2015), lumped-hydrological models (Vogel & Sankarasubramanian, 2000), and hybrid (statistical-mechanistic) models for estimating streamflow at different time scales (Evenson et al., 2021). Utilizing distributed hydrological models provide a viable alternative to estimating daily streamflow at ungauged locations, but challenges remain in accurately predicting streamflow over continental scale (Johnson et al., 2023). Frame et al. (2021) evaluated the National Water Model (NWM) for selected virgin basins and found that the performance of NWM is poorer compared to the post-processing models. Post-processing can often decrease bias in hydrological model outputs and reduce systematic errors from forcing and other process representations (Li et al., 2017; Rezaie-Balf et al., 2019). Post-Processing methods can be data-driven (e.g. Johnson et al., 2023) or physically informed (Wu et al., 2019). Recently, Frame et al. (2021) applied a long short-term memory (LSTM) machine learning approach to improve daily NWM streamflow prediction across CONUS and compared its performance with streamflow prediction using LSTM with just atmospheric forcings. However, this analysis has been performed only for virgin basins and did not focus on controlled basins. Johnson et al. (2023) used random forest models to identify the basin characteristics and hydroclimatic information that influence the NWM performance over the CONUS. They found that exist in arid basins and basins with moisture and energy being out of phase exhibit significant bias and reduced Nash-Sutcliffe Coefficient (NSE) (Johnson et al., 2023). They also found basin characteristics such as total contributing drainage area and path length also influence the NWM performance. Further, variables indicating anthropogenic activities - percent imperviousness and upstream storage in dams– also influence the bias and NSE in predicting the flood flows (Johnson et al., 2023). This indicates that the NWM performance depends on basin and hydroclimatic information, thereby exhibiting a regional/spatial error structure in predicting flood flows. We intend to consider these basin characteristics and hydroclimatic information as a hierarchy in predicting flood flows in ungauged basins within the proposed hierarchical model.

In this study, we propose a spatiotemporal hierarchical modeling framework by using hydroclimatic information and basin characteristics as a hierarchy for improving flood prediction in ungauged basins under natural and controlled types over the CONUS. For demonstration, we consider NWM streamflow prediction as a basin-specific predictor, but one could replace this with meteorological forcings (see the discussion for additional details) or any other hydrologic model predictions. The proposed hierarchical model is evaluated rigorously using a spatio-temporal validation based on its ability to predict

in 2,674 gauges, which include both natural and controlled basins, over the CONUS. The manuscript is organized as follows: We first describe the datasets used and the formulation of the hierarchical model. Then the results are presented from overall model performance analyses through rigorous temporal and spatial validation, which is then followed by a seasonal analysis of the explained variance from each predictor and by a discussion. Finally, we summarize the key findings from the study along with implications for future work in improving flood prediction in ungauged basins.

2 Hydroclimatic Data and Hierarchical Model Setup

To develop a hierarchical model for predicting flood flows in ungauged basins, we have obtained several predictors that include daily streamflow from NWM, basin characteristics and hydroclimatic information. Details on the procedure for obtaining the predictors and USGS gaging stations are described below in the following sections.

2.1 National Water Model

The National Oceanic and Atmospheric Administration (NOAA) National Weather Service (NWS) Office of Water Prediction (OWP) have implemented the operational National Water Model (NWM) to support the operational activities of NWS River Forecast Centers, the Federal Emergency Management Agency and other government agencies (National Research Council, 2006). A primary goal of the NWM development is to provide flood predictions for any given riverine location within the coterminous United States (CONUS). The NWM is a continental-scale distributed high-resolution hydrologic model that produces streamflow predictions for 2.7 million stream reaches across the contiguous United States (CONUS), based on a variety of data ranging from radar-gauge observed precipitation to numerical weather prediction (National Research Council, 2006). The NWM relies on the Weather Research and Forecasting hydrologic model (WRF-Hydro) architecture (Ghotbi et al., 2020) and provides streamflow predictions extending up to 30 days in advance over the CONUS. NWM provides these predictions at gauged locations but still consists of errors, which depend on both hydrologic process representation and forcing errors (Viterbo et al., 2020). Furthermore, NWM predictions lack spatial correlation between predictions available at ungauged locations and nearby gauged locations, particularly in estimating high flows because of the spatially uninformed model parameters (Brunner et al., 2020; Tijerina et al., 2021). Johnson et al. (2023) highlighted that the NWM exhibits systematic errors across space and depends on basin characteristics and hydroclimatic information. Due to these shortcomings, studies focusing on improving the NWM forecasts have been emerging recently using various post-processing methods (Frame et al., 2021). For this study, we consider daily flows from NWM (Q^{NWM}) as a predictor in the hierarchical model.

2.2 CONUS Basin selection

Co-located NHDPlusV2 COMIDs and USGS National Water Information System (NWIS) gages are extracted from the Routelink file associated with NWM v2.0. The "dataRetrieval" R package (Hirsch & De Cicco, 2017) is then used to identify which of these basins have at least 10 years of observed daily flows between 1993 and 2018 for evaluating the proposed hierarchical model. Once identified, streamflow data are extracted from NWIS by gage ID using "dataReteival", and NWM v2.0 data are extracted by COMID using "nwmTools" (`https://mikejohnson51.github.io/nwmTools`). For the NWM estimated flows, hourly data are converted to daily mean flows. Drainage area for each basin is obtained from the GAGESII USGS database. Additionally, the GAGESII dataset classifies each USGS station's flow into controlled/natural based on the 2009 hydro-climatic network (HDCN) database and we consider that classification for developing the hierarchical modeling in the region. A coastal classification is applied to those catchments

within 150 km of the coastline, as these areas are susceptible to impacts from tides (F. Zhang et al., 2018; Ramaswamy et al., 2004). In total, we consider 2,674 controlled USGS gages, 451 natural gages, and 1,150 coastal gages spanning 1,508 basins (at HUC08 levels) and they are grouped into natural, controlled and coastal basins across the 18 HUC02 (Figure 1). As the focus of this study is on improving flood flows in ungauged basins, we consider only the above-normal streamflow condition, which is defined as the flow above the 67th percentile of daily flow in a given station. Thus, we obtain observed daily high streamflow (Q) (above 67th percentile of daily flow) from NWIS, which will be the predictand in setting up the hierarchical model. The corresponding day's daily streamflow from NWM reanalysis runs (v2.0) is considered as a predictor (Q^{NWM}) for the selected basin.

2.3 Upstream Reservoir Storage

Since streamflow is regulated by reservoirs to meet downstream water demand (Kumar et al., 2022), we consider cumulative upstream reservoir storage of a USGS gaging station as a predictor in the hierarchical model. Studies have shown that reservoir storage and their retention time significantly alter the downstream flow characteristics (Chalise et al., 2021). The dams associated with each gage are obtained from the 2019 United States Army Corp of Engineers National Inventory of Dams (NID) database (`https://nid.usace.army.mil`) and the cumulative upstream reservoir storage are obtained for each gage. We use the "Maximum Storage" from the NID database for calculating the cumulative upstream storage (S) (USACE, 2022). We also obtain the contributing area above each dam from NID.

2.4 Hydroclimate Data

In addition to dam attributes, we consider the following hydroclimatic attributes as predictors for the hierarchical model: a) aridity index b) mean monthly potential evapotranspiration and c) phase difference between moisture (precipitation) and energy (potential evapotranspiration).

Monthly potential evapotranspiration ($PET; \mathrm{kg/m^2}$) and precipitation ($P; \mathrm{kg/m^2}$) are obtained from phase 2 of NLDAS for January 1993 through December 2018 (`https://disc.gsfc.nasa.gov/datasets?keywords=NLDAS`). For the contributing area to each gage, the mean monthly PET and P are computed. The availability of moisture (i.e., precipitation) and energy (i.e., PET) together within the seasonal cycles influence the streamflow estimation (Petersen et al., 2012, 2018). The aridity index (AI) is calculated as the ratio of mean annual PET to mean annual P ($\overline{PET}/\bar{P}$) over each basin. is used since arid basins are more difficult to calibrate and estimate streamflow compared to humid basins (Sankarasubramanian & Vogel, 2002).

The phase difference between moisture and energy is computed as the Spearman correlation between the monthly precipitation and potential evaporation ($\rho(P, PET)$). The Spearman correlation coefficient is determined for each NLDAS2 cell using the mean monthly PET and P over the 26 years and the basin-wide mean ρ is computed. If ρ is negative (positive), it indicates moisture and energy are out of phase (in phase), which could result in more (less) potential for runoff generation from the basin. We also consider mean monthly PET at basin level as an additional predictor in the hierarchical model. It is important to note that AI and ρ are not time-varying predictors and quantify the climatological interaction between moisture and energy.

2.5 Land use data

Since our interest is in developing flood predictions, land use, particularly urbanization, can play an important role in generating runoff and evapotranspiration (Merz et al., 2021). Urban imperviousness represents developed surface (e.g., roads, driveways,

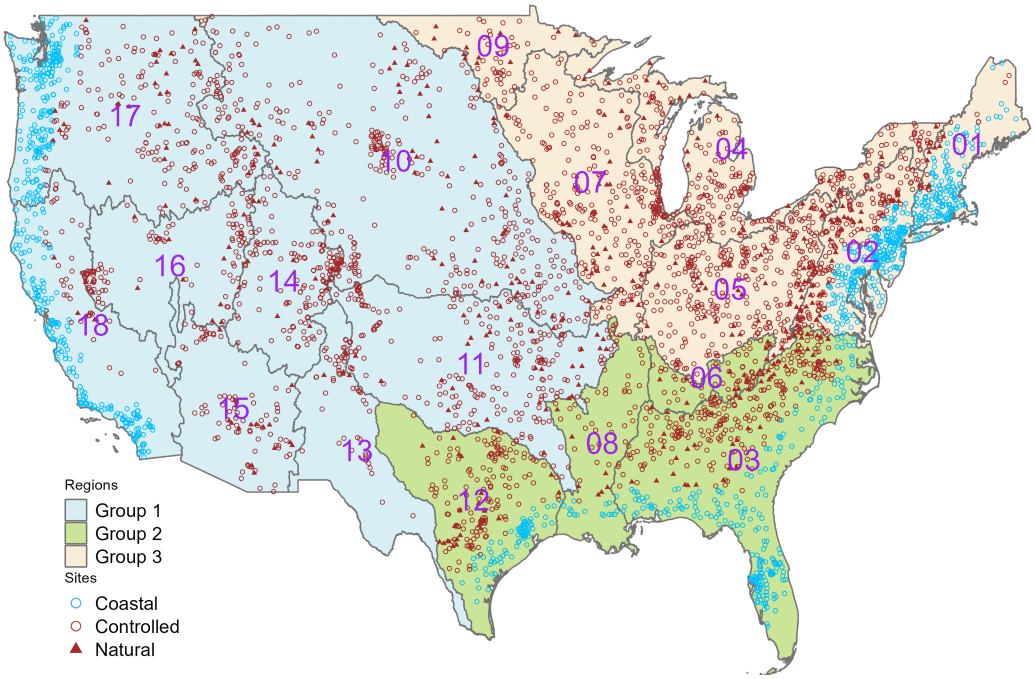


Figure 1. Spatial distribution of 3,640 USGS stream gages classified as "controlled" and "natural" based on Hydro-Climatic Data Network (HCDN). These sites are also classified as coastal sites if they are within 150 km distance of a coastline. The HUC02 regions are grouped into three regions based on regional hydroclimatology.

sidewalks, parking lots, rooftops) that limit the infiltration into the underlying soil and increase the frequency and intensity of downstream runoff (Caldwell et al., 2012). Thus, to reflect the development in the basin, urban imperviousness is derived from the U.S. Geological Survey (USGS) National Land Cover Database (NLCD) 2019 Impervious data layer that quantifies the percent developed impervious surface in each pixel (`https://www.mrlc.gov`). We also use the NLCD 2019 land cover layer to identify the percentage of each basin that is categorized as urban (Anderson level 1 value 2). Additionally, we also identify 7% of NWIS sites as urban sites based on the Census Bureau definition with densely settled urbanized areas of 50,000 or more people using the 2020 Census data (`https://www2.census.gov/geo/tiger/TIGER2020`).

2.6 Spatiotemporal Hierarchical Model (STHM) - Formulation

We use spatial and temporal hierarchies to develop the model for flood prediction in ungauged basins. We divide the day-of-year into 37 10-day windows and we denote each 10-day window, τ, for the temporal hierarchy. We categorize the 18 HUC02 regions into three groups/levels with each group, j, denoting the spatial hierarchy (Figure 1). These three groups are based on empirical hydroclimate similarity considering aridity, regional landform, climate, and ecosystems (Heidari et al., 2020). Thus, all sites are nested under each spatial group k, and each time step is nested under the spatial group k.

To predict the streamflow (Q) at a specific site in a basin i at daily time step t within each temporal cluster τ and spatial cluster k, we have basin-specific terms/coefficients and fixed terms (i.e., coefficients are common to all sites under the same spatial and temporal cluster). The fixed terms include predictors aggregated mean values at the basin (HUC08, j) level, and can be separated into two groups: (1) no variation over time, and

(2) varying over time. Predictors not varying temporally include the spearman correlation indicating moisture and energy being in-phase or out-phase (ρ), the total dam storage (S), the aridity index (AI), and the percent impervious surface (Imp). These predictors share the same coefficients with the same 10-day widow and spatial group (i.e. same i and j value). Predictors varying over time include mean potential evapotranspiration in the corresponding basin (PET) within the same 10-day windows, and HUC08 level previous 3-day area weighted observed streamflow (Q_t^{3d}). For PET, the coefficients will also vary among different months within the same 10-day window and spatial group. For Q_t^{3d}, the coefficients will also vary among different basins (HUC08).

Thus, for each τ-th 10-day window,

$$Q_{t(\tau,i,j,k)} = Qs_{t(\tau,i,j,k)} + \varepsilon_{t(\tau,i,j,k)} \tag{1}$$

at each site level (i):

$$Q_{t(\tau,i,j,k)} = \beta_{0(\tau,i,j,k)} + \beta_{1(\tau,i,j,k)} Q^{\mathrm{NWM}}_{t(\tau,i,j,k)} + \beta_{2(\tau,i,j,k)} Q^{3d}_{(t(\tau,i,j,)} \tag{2}$$

The intercept term $\beta_{0(\tau,i,j,k)}$ in equation (2) is estimated for each τ-th time window at HUC08 level (j) using mean potential evaporation in the corresponding basin (PET) within the same τ-th 10-day windows (PET). Thus, at each HUC08 level (j):

$$\beta_{0(\tau,i,j,k)} = \beta_{00(\tau,i,j,k)} + \beta_{01(\tau,j,k)} PET_{(\tau,j,k))} \tag{3}$$

The intercept term $\beta_{00(\tau,i,j,k)}$ in equation (3) is estimated at the grouped HUC02 level (k, spatial group shown in Figure 1)

$$\beta_{00(\tau,i,j,k)} = \beta_{000,\tau} + \beta_{001,\tau} AI_{i(j,k)} + \beta_{002,\tau} Imp_{i(j,k)} + \beta_{003,\tau} \rho_{i(j,k)} \tag{4}$$

Thus, the proposed spatio-temporal hierarchical model has the final form as follows:

$$\begin{aligned} Q_{t(\tau,i,j,k)} = \beta_{000,\tau} + \beta_{1(\tau,i,j,k)} Q^{\mathrm{NWM}}_{t(\tau,i,j,k)} + \beta_{2(\tau,i,j,k)} Q^{3d}_{t(\tau,i,j,k)} + \\ \beta_{01(\tau,j,k)} PET_{\tau(j,k)} + \beta_{001,\tau} AI_{i(j,k)} + \\ \beta_{002,\tau} Imp_{i(j,k)} + \beta_{003,\tau} \rho_{i(j,k)} + \varepsilon_{t(\tau,i,j,k)} \end{aligned} \tag{5}$$

Q^{NWM} is the NWM daily flow; ρ is the spearman correlation indicating moisture and energy being in-phase or out-phase (ρ); PET is the mean 10-day potential evaporation as mentioned above; S is the upstream total dam storage; AI is the aridity index, and Imp is the percent impervious and ε is the residual.

Since we are interested in estimating the flow at ungauged basins, we represent the antecedent conditions based on previous 3-day average flow at the HUC08-level. Thus, if there are 'm' gauged basins within the HUC08, for a given basin, then we obtain the depth of runoff for the previous 3 days for all the 'm' basins. This 3-day average depth of runoff is then multiplied by the drainage area of the basin to get the basin-relevant average depth of runoff, and then average them 3 days over 'm' basins to get previous average 3-day flows. For a gauged basin, of course, one can simply use the 3-day average flows as a predictor instead of the area-weighted flows. Q_t^{3d} is HUC08 level, previous 3-day area weighted observed streamflow. The moving average is calculated as:

$$Q_t^{3d} = \frac{1}{3}(Q^A_{t-3} + Q^A_{t-2} + Q^A_{t-1}) \tag{6}$$

Q_t^{3d} on the same day is calculated as:

$$Q_t^A = \sum A_i \sum \frac{Q_{t,i}}{A_i} \tag{7}$$

where $Q_{t,i}$ is the streamflow at i-th basin within the same HUC08 at date t, A is the corresponding drainage area of the i-th basin. t is the corresponding date of the prediction.

Thus, equation (5) is set up for nine groups for three basin types (i.e., natural, controlled and coastal) for three groups (Figure 1). For each of the nine models, coefficients are estimated over each 10-day window. To determine the best-fitted model for each group, we select the variables using the $\ell 1$-penalized maximum likelihood method proposed by Groll and Tutz (2014) and computed the coefficients using the R package 'glmmLasso' developed by Schelldorfer et al. (2014).

2.7 Model assessment and validation

Given that we are interested in assessing the performance of the hierarchical model for estimating flows in ungauged locations, we consider both temporal and spatial validation procedures for assessing model performance. The temporal validation is performed to evaluate the STHM performance over a period different from the calibration, whereas the spatial validation is performed to evaluate the STHM for application in ungauged basins. The temporal validation is performed by calibrating the STHM model using the data from 1993 to 2008 with the remaining data from 2009 to 2018 being considered for validation. For spatial validation, we use the $k(20)$-folder cross-validation method (Browne, 2000). We treat 5% of locations as ungauged within their hierarchical group, fit the STHM for the remaining 95% of stations for the period 1993 and 2018, and then evaluate the SHTM performance for the period 2009 to 2018 for the left-out 5% of the basins. This process of leaving out the 5% basins is repeated until all the considered basins are left out and evaluated in a cross-validation mode.

The Nash–Sutcliffe efficiency (NSE) is widely used to assess the predictive skill of hydrological models (McCuen et al., 2006). In a perfect model with an estimation error variance equal to zero, the resulting NSE equals 1; a model with an estimation error variance equal to the variance of the observed time series, the corresponding NSE equals 0. Conversely, an NSE less than zero occurs when the observed mean is a better predictor than the model. The model performance criteria recommended by Moriasi et al. (2007) are used for evaluating the improvement's performance. Model prediction is considered "acceptable" if NSE scores are greater than 0.5, and considered "good" if the NSE is above 0.67.

To evaluate the impact of each predictor in predicting streamflow, we use the relative importance estimator proposed by Grömping (2007), which decomposes the explained variance ($r^2_{y(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)}$, i.e., r^2_β) of observed streamflow to each predictor. Model performance from the temporal validation alone is considered for analyzing the significance of each selected predictor. In general, the whole model variance $R^2_{y(x_(1,)x_2,\ldots,x_p)}$ is the sum of $r^2_{y(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)}$ which is the correlation between y and that portion of x_j which is uncorrelated with the remaining predictors of j-th predictor x_j with the remaining predictors. Thus,

$$R^2_{y(x_1,x_2,\ldots,x_p)} = \sum_{j=1}^{p} r^2_{y(x_j|x_1,\ldots,x_(j-1),x_(j+1),\ldots,x_p)} \tag{8}$$

$$r^2_{y(x_j|x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)} = R^2_{y(x_1,x_2,\ldots,x_p)} - R^2_{y(x_1,\ldots,x_{j-1},x_{j+1},\ldots,x_p)} \tag{9}$$

3 Results and Analysis

We first evaluated the STHM performance from the temporal validation (Figure 2) and then provided a detailed analysis on the role of each predictor (Figure 3). Following that, we analyzed the performance of STHM in predicting flood flows in ungauged basins based on spatial validation (Figures 4-8).

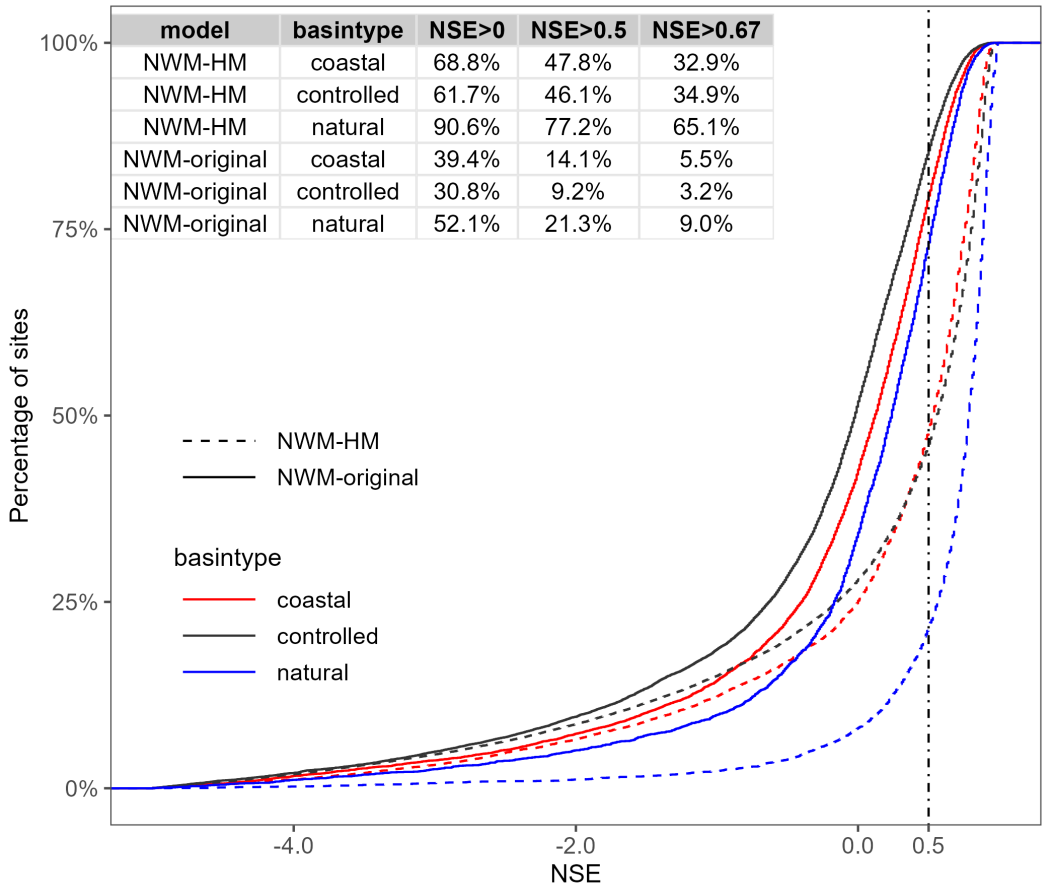


Figure 2. Cumulative distribution of high flow NSE from the NWM and NWM-HM approaches, separated by basin classification, during the validation period (2009-2018).

3.1 STHM Performance on CONUS Flood Prediction

The STHM parameters were estimated over the calibration period (1993-2008) only for high flows, which was defined as the 67th percentile of observed daily streamflow for a given day obtained from the STHM predicted flows from 2009 to 2018. Thus, all the reported results were only for high flows by considering when the observed flows are above the respective day's 67th percentile flow. The validation results, presented as the cumulative distribution of NSE, showed significant improvement in high streamflows/flood flows across the CONUS compared to the flood flows estimated from NWM streamflow reanalysis (Figure 2). Figure 2 also provided the improvements in NSE from STHM for the three groups of basins – natural, controlled and coastal – over the CONUS. Overall, STHM improved NSE by 0.1 for more than 65% of sites under temporal validation (Figure 2). This suggested that STHM not only reduced the error in NWM systematically but also improved the prediction using a limited number of parameters estimated by pooling NWM

data and other characteristics from the sites in the grouped region (Figures 2). Overall, during the validation period, 62.7% of controlled basins and 68.4% of natural basins improved better in predicting high flows compared to NWM reanalysis prediction of high flows (Figure 2). The improvement in natural basins was mainly dependent on climatic factors, hence exhibiting better performance. In contrast, controlled basins were more complex as their observed streamflow depends on the reservoir operation policies (Turner et al., 2020; Zhao & Cai, 2020). Coastal basins also showed limited improvements from the STHM as the observed high flows are influenced by high tides (Ramaswamy et al., 2004; F. Zhang et al., 2018). The performance of STHM in improving NWM reanalysis runs was summarized over the CONUS and by season (Figure S1). Overall, our model improved mean NSE by at least 0.1 during the validation period for more than 60% of the sites (Figure S1).

3.2 Importance of predictors in the STHM

Overall, the STHM model with all selected predictors (Q^{NWM}, ρ, PET, AI, Q^{3d}, Imp) explained more than 74.8% of the variance of the observed high flows during the validation period (Figure 3). To be specific, the STHM explained the observed flow variance by 74.4%, 76.3%, and 71.1% for controlled, natural, and coastal basins respectively (overall average values). The explained variance by each predictor in the STHM equation (1) in improving NWM streamflow can be decomposed using the relative importance estimator proposed by Grömping (2007). Based on the decomposed explained variance (r_β^2), NWM reanalysis streamflow accounts for more than 55% of the variance of the observed streamflow on average across the CONUS. Other predictors at the HUC08 level explained an additional 18% to 35% of observed streamflow variance (Figure 3).

Previous three-day areal-weighted streamflow (Q^{3d})

Previous 3-day streamflow (Q^{3d}) was the most important predictor (excluding Q^{NWM}) in controlled and coastal basins, explaining an average 18% and 16% variability of streamflow, respectively (Figure 3). In natural basins, Q^{3d} is the second most important predictor, and explained 12% of the variability in streamflow but showed strong seasonality over all the regions (Figure 3). Q^{3d} showed higher importance in controlled basins than in natural basins as dam operations highly regulate the observed flow (Gierszewski et al., 2020). Based on the variance explained by Q^{3d} , coastal basins showed pronounced seasonality particularly in the West region. Further, Q^{3d} was more important in warmer regions than in colder regions as higher evapotranspiration results in a more varying antecedent conditions. The coefficient of Q^{3d} showed strong seasonal changes in the warmer regions, but did not have clear seasonality in colder regions since they did not experience much change in antecedent conditions due to reduced evapotranspiration.

Aridity index

Aridity index (AI) proved to be an important variable in improving the model predictions in warmer regions (e.g., Group 1, Figure 3). AI represents the long-term balance between water and energy and showed significant seasonality in explaining the variance of high flows for all three regions (Figure 3). Among the natural basins, Group 3 (Group 2) basins had the least (most) improvement when accounting for aridity index, since most basins were humid and had little (significant) spatial variability in the aridity index. AI also played an important role in the west, particularly during the spring, which reflects the seasonal water availability from snowmelt (Gudmundsson et al., 2016). In the case of controlled basins, a similar seasonality pattern was observed in all three regions, but the explained variance was relatively less, which was to be expected as the controlled basins dampen the natural hydroclimatic variability (Figure 3). In coastal basins, the seasonal variation in AI was minimal. Overall, AI explained NWM streamflow vari-

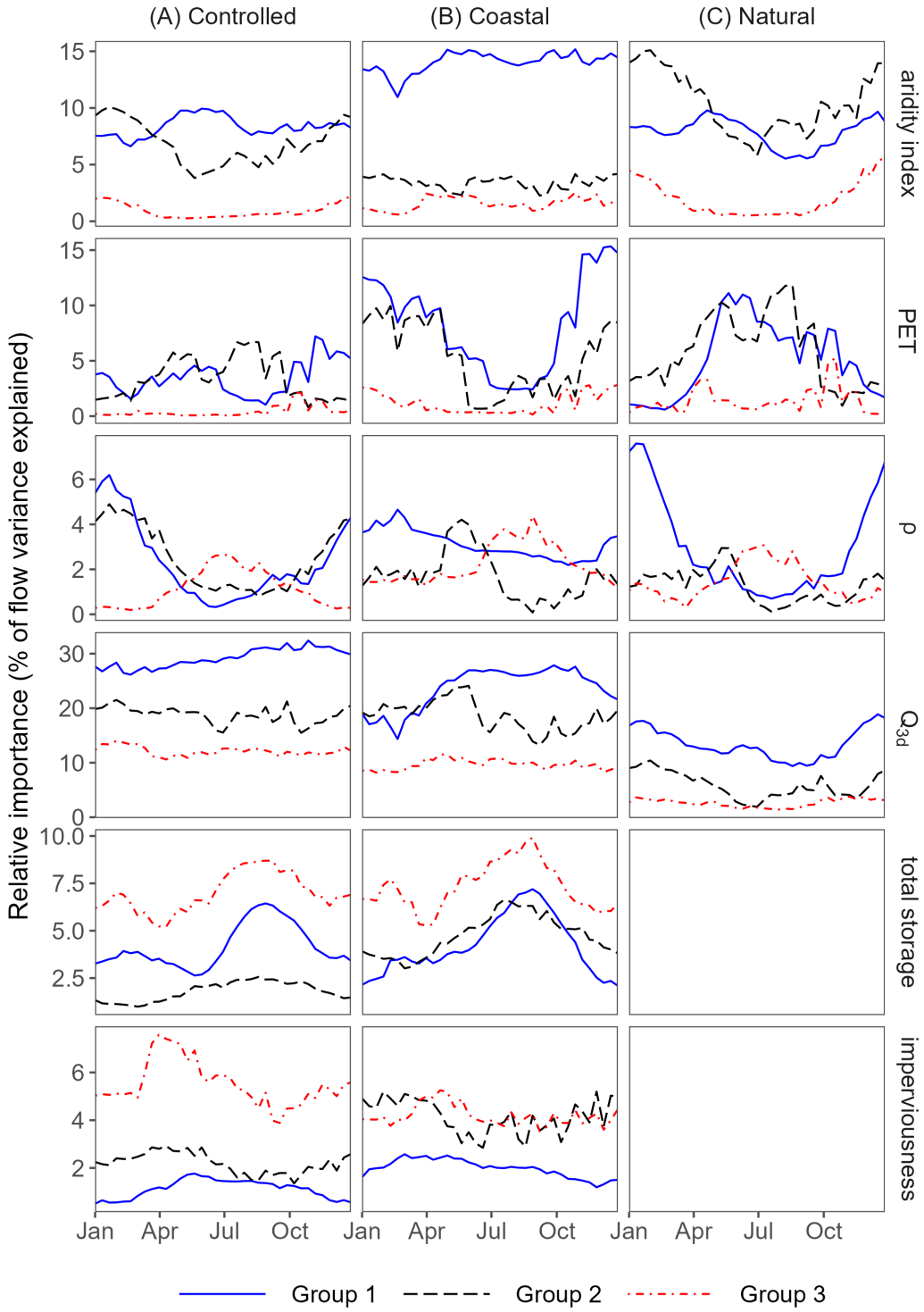


Figure 3. Relative importance of selected predictor variables, expressed as % variance of streamflow explained by the hierarchical model, for each month for (A) controlled, (B) coastal, and (C) natural basins over the grouped hydroclimatic regions (Group 1, 2, 3). Note that NWM flow was not included here.

ability as around 12-15% for the Group 1 basins all through the year and 2-4% variability for the other two groups.

Phase correlation between moisture (precipitation) and energy (PET)

The phase correlation between energy and moisture (ρ) explained a smaller amount of streamflow variability but showed strong seasonal dynamics across the CONUS. For controlled and natural basins in Group 1 and 2, the phase correlation was less important in summer months, while in Group 3, the phase correlation was more important in summer months. Since ρ was estimated by the correlation between monthly PPT and PET, it represents the co-occurrence of energy (PET) and moisture (precipitation) and the explained variance by it indicated their role in influencing the runoff. The ρ was generally negative during the summer months, and it was positive over the Southeast during the winter months (Petersen et al., 2018, 2012). Basins having negative correlation (i.e., moisture and energy being out of phase) exhibit strong seasonality in streamflow with increased potential for runoff and they were difficult to high streamflows. Further, the spatial variability in phase correlation was largest (least) during the winter and fall months over Group 1 and 2 (Group 3). On the other hand, phase correlation variability was the largest during the summer months over the Group 3 basins. Hence, it exhibits higher explained variance in improving the NWM streamflow during the summer. Explained variance by ρ over the coastal basins was around 2-4% across all the regions and does not seem to play a significant role in improving the high flow prediction. One potential reason for this was that the most coastal basins are controlled, hence they had limited role due to phase correlation.

Mean 10-day PET

Mean 10-day PET (PET) represented the amount of energy available at a given time and it displays significant seasonality in the variance explained in improving the NWM high flows (Figure 3). As expected, for controlled basins, (PET) had a minimal role in improving the NWM prediction seasonally over all three regions. In Group 3, using the mean monthly PET in the STHM had minimal impact on improving NWM streamflow, only explaining 2% of the variance for all three different types of basins. This is expected since Group 3 covers the northern regions, which exhibited minimal spatial (PET) variability in natural basins. The other two regions (1 and 2) exhibited significant spatial variability in PET, and as a result, including it in the STHM explained around 12% of the variance of observed streamflow for natural basins with significant seasonality, particularly during the summer. In the case of coastal basins, PET explained the observed high flows better during the winter and in the fall over Group 1 and 2. This was primarily due to the latitudinal gradient in (PET) over these two regions during those seasons.

Total storage and impervious surface

As expected, the total upstream artificial/reservoir storage in the basin played an important role in controlled and coastal basins, but not in natural ones. Regression coefficients of total upstream storage showed a positive correlation with streamflow throughout the year for the three grouped regions (Figure 3). In controlled basins, total storage in both Group 1 and 2 showed significant seasonality in explaining observed streamflow particularly during the summer months as these are the months significant floods occur due to in-phase seasonality (Midwest), snowmelt and hurricanes. However, in the coastal basins, explained variance from total upstream storage indicated strong seasonality in explaining NWM streamflow, particularly in the summer, explaining around 3-9% variance. Explained variance by the basin imperviousness did not show strong seasonal variability except in the Group 3 controlled basins. Overall, imperviousness accounted for 2-4% variance in explaining high flows for coastal basins.

3.3 Potential for STHM in predicting high flows in ungauged basins

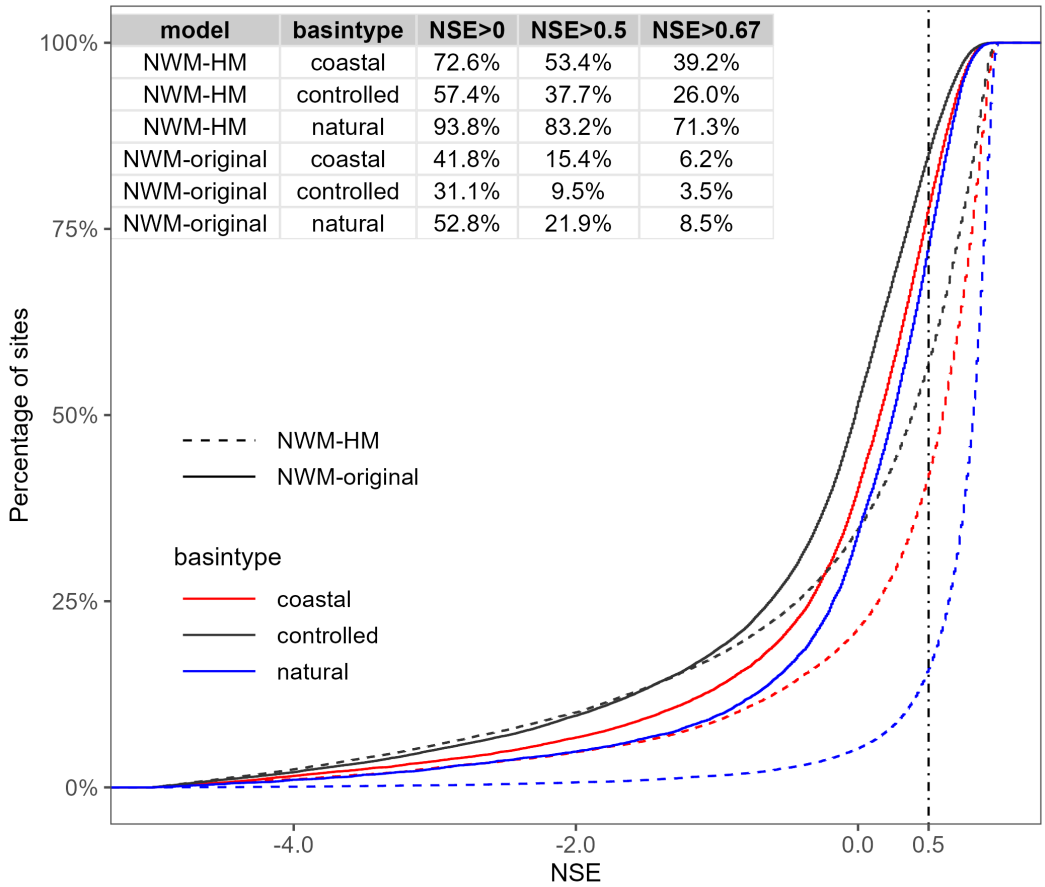


Figure 4. Cumulative distribution of high flow NSE from the NWM and NWM-HM approaches, separated by basin classification, over the period from 1993-2018. These were computed by treating each basin as an ungauged basin under k-fold cross-validation for each basin type.

Given NWM daily streamflows are available for any ungauged locations within the CONUS and other predictors of the STHM (i.e., basin characteristics and hydroclimatic information) could be estimated for any location based on openly available data sources, we evaluated the potential for flood prediction for any ungauged basins using STHM. Since STHM could not be evaluated for ungauged basins with no streamflow data, we performed $k(20)$-fold cross-validation under which 5% of the basins were left out and the remaining 95% of the basins were used for parameter estimation of the STHM. This process was repeated until all the basins are evaluated at least once in the "ungauged" prediction mode. Thus, this spatial cross-validation experiment was similar to evaluating the STHM in an "ungauged" prediction mode. The cross-validation experiment showed promising results for the "ungauged" basin, with 61% of the natural basins improved to an "acceptable streamflow prediction" (NSE>0.5), and 63% of natural basins improved to a "good streamflow prediction" (NSE>0.67) (Figure 4).

Seasonal performance of STHM under "ungauged" prediction mode

Spatially, STHM improved the majority of "ungauged" basins across all groups (Figure 5). The biggest improvement occurred in colder regions, with improvements in av-

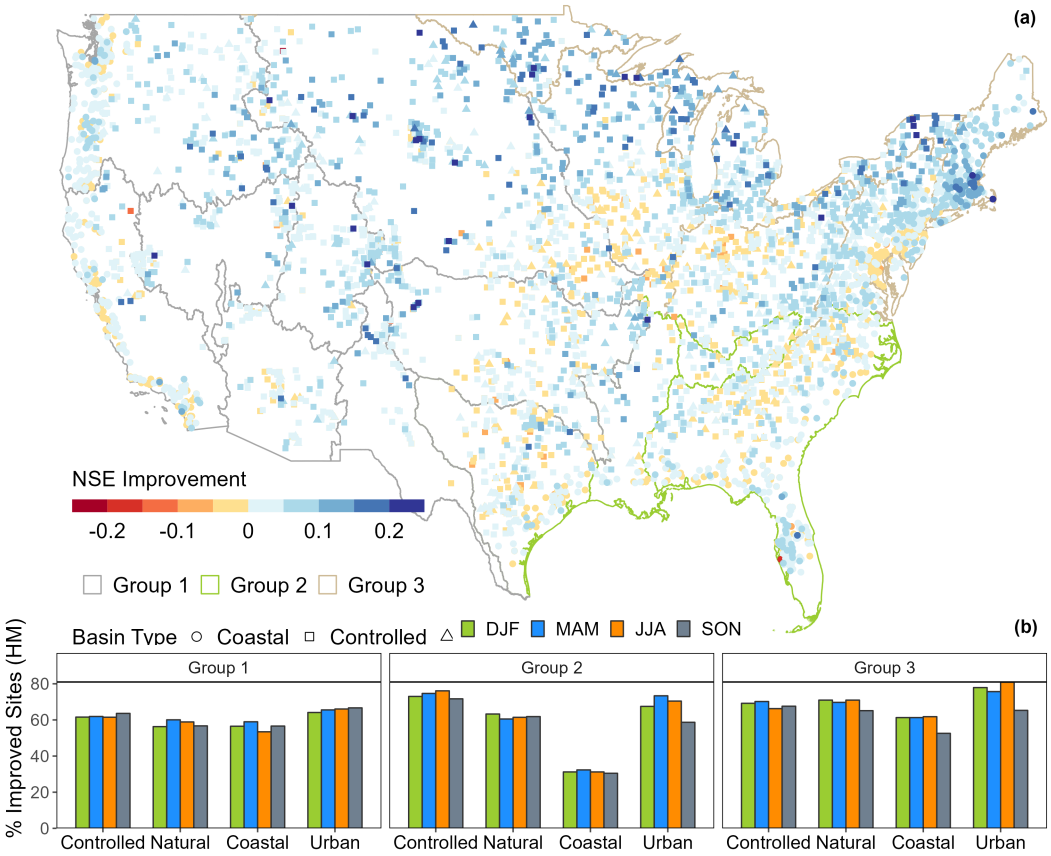


Figure 5. (a) maps the difference in NSE between HM and NWM daily flow by treating 5% sites as "ungauged" using k(20)-folder cross-validation from 1993-2018. Positive (Negative) values indicate the hierarchical model (NWM) better predicted high flows. (b) plots seasonal performance, indicated as % of sites with improved NSE, shown in for the three regions over the four basin classifications. Note that DJF, MAM, JJA, and SON are the initials for month, representing spring, summer, fall, and winter.

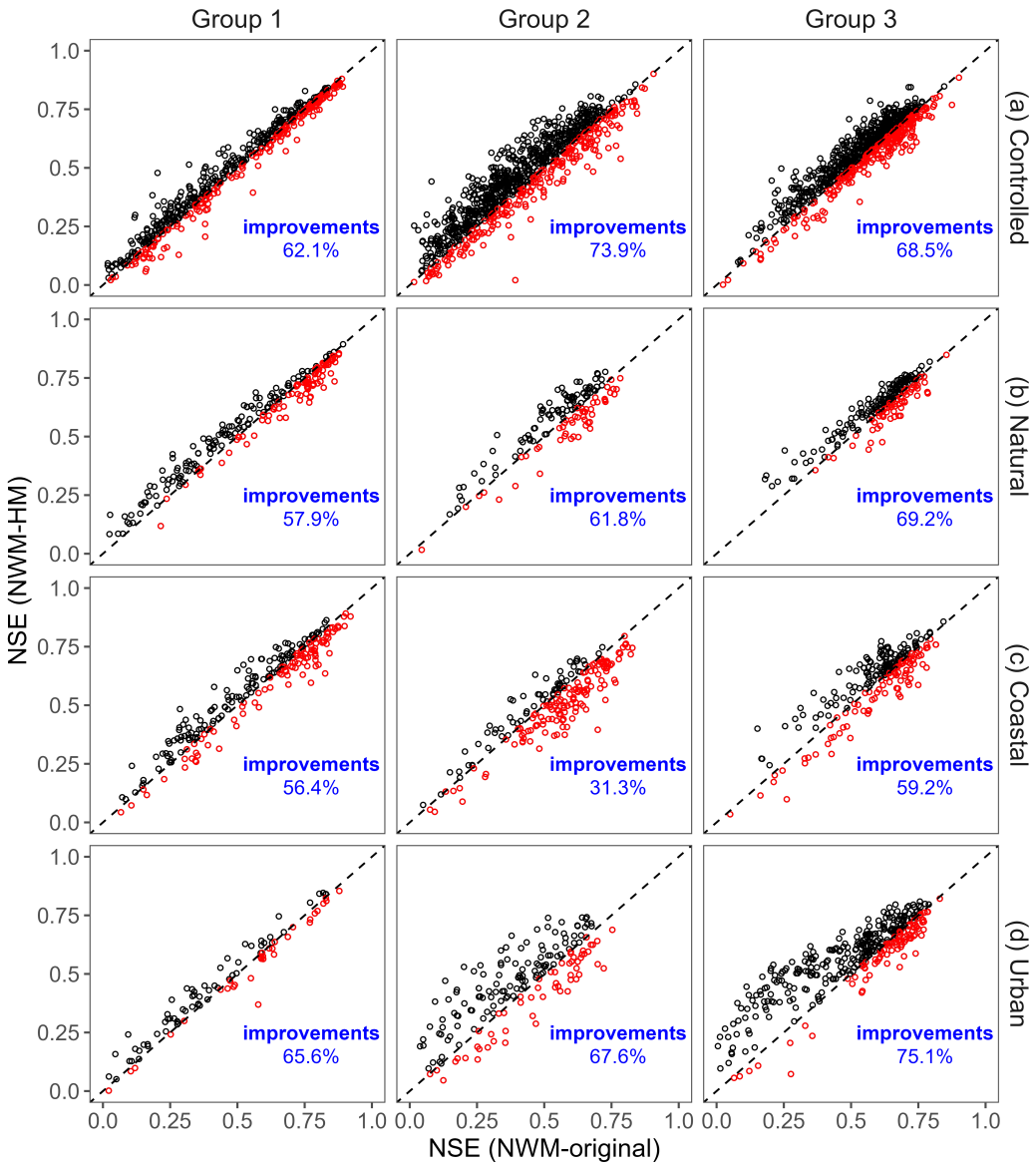


Figure 6. Performance of the NWM v2.0 reanalysis data compared to the hierarchical model for daily high flows by treating 5% sites as "ungauged" using k(20)-folder cross-validation from 1993-2018. in (a) controlled, (b) natural, (c) coastal, and (d) urban basins in each region. Urban basins are defined as basins containing 50,000 or more people based on the 2020 U.S. Census. High flows are defined when the observed daily flows are greater than the 67th percentile of daily flow.

erage NSE by at least 0.2 for more than 30% of the sites. The largest improvements occurred in the northern basins and along the Appalachian Mountains, however significant systematic error persisted in the southeastern basins (Figure 5a). The differences in the hierarchical model and the NWM performance also showed seasonal differences (Figure 5b, Figure S2). The highest overall improvement of the STHM occurred in the winter (December/January/February, or DJF) and spring (March/April/May, or MAM) seasons across the CONUS, accounting for 71% of the improvement throughout the year. HUC02-regions 14 and 17 (Upper Colorado and Pacific Northwest) had shown the smallest springtime improvement (MAM, Figure S2), while basins in HUC02-regions 7 and 9 (Upper Mississippi and Souris-Red-Rainy) had the worst winter performance (DJF, Figure S2). The largest improvement of NSE in Group 3 occurred in summer (June/July/August or JJA, Figure S2). The largest improvement of NSE in Group 1 occurred in fall (September/October/November, SON, Figure S2) and the coastal basins had the highest NSE improvement overall. However, Southeast coastal basins showed limited improvement, as most basins did not improve over all the seasons. Furthermore, there was no significant improvement in skill between controlled and natural basins among different seasons (Figure S2). But, in Group 3, urban and coastal basins showed significant improvement in skill in all seasons but fall (SON, Figure 5b).

Performance of STHM for different basins types under "ungauged" mode

Among different regions, Group 3 had the highest percentages of basins with significant NSE improvement; Group 1 had the least percentage of sites improved for controlled and natural basins (Figure 5). The primary reason for limited improvement over the west (except Region 17 - Pacific Northwest) was that most basins are arid, and the runoff has strong seasonality, hence they are difficult to model (Sankarasubramanian and Vogel, 2003). Group 2 had the minimum improvement among coastal and urban basins. Contrastingly, Group 1 outperformed other regions under coastal and urban basins, respectively (Figure 5). Overall, except for Group 2 coastal basins, the STHM improved the NWM flood prediction for over 56% of the basins in that category (Figure 6). Basically, STHM performance improved NWM prediction from 56% to 75% of the basins under each category. It was important that STHM did not improve NWM performance in the remaining 25-45% of the basins. In the discussion section, we provided details and experiments on how to improve STHM in those basins along with challenges in improving the performance of STHM in coastal and urban basins.

Cross-validating the STHM in predicting flood flows in "ungauged" mode showed 63% of natural basins, 39% of coastal basins, and 26% of the controlled sites have NSE above 0.67 from the STHM (Figure 6). This was a significant improvement in comparison to the NWM reanalysis runs. The cross-validation results also showed a similar spatial pattern (Figure 5) with NSE improving in 63% of overall basins (Figure 6). The highest NSE improvement was along the Appalachian Mountain range (Figure 5). The hierarchical model improved the overall NSE for more than 68%, 63%, and 49% of sites in controlled, natural, and coastal basins respectively (Figure 6b). This shows potential in utilizing STHM for flood prediction in ungauged locations as the hierarchical model uses both NWM streamflow, basin characteristics, and hydroclimatic information by leaving out the basin in the parameter estimation process. It was important to note that the previous 3-day streamflow (Q^{3d}) was the only predictor that depends on observed streamflow. We had considered a simple drainage-area method to estimate the previous 3-day streamflow for an ungauged location for quantifying antecedent conditions. We discussed alternate approaches for improving that estimate in the next section.

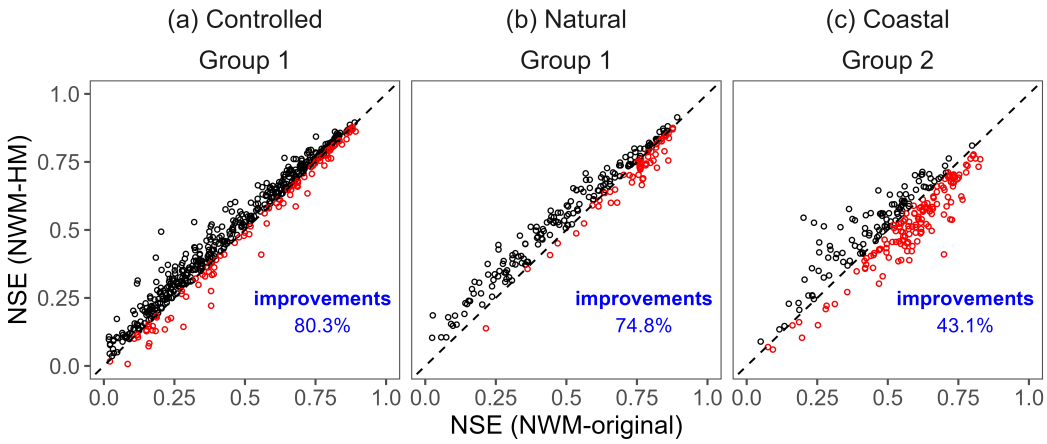


Figure 7. Performance of the NWM v2.0 reanalysis data compared to the hierarchical model for daily high flows based on HUC02 levels (instead of spatial groups shown in Figure1) for selected groups. Data defined by treating 5% of the locations sites as "ungauged"and using a k(20)-folder cross-validation during the period (1993-2018). The results are grouped to be comparable with Figure 6.

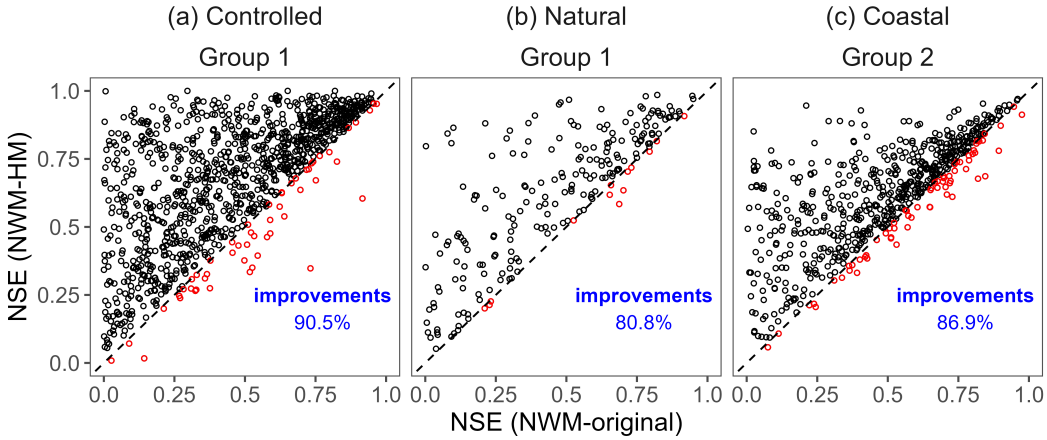


Figure 8. Performance of the NWM v2.0 reanalysis data compared to the hierarchical model for daily high flows using a site-specific model (as a proof of concept) by treating 5% sites as "ungauged" locations using k(20)-folder cross-validation during the period (1993-2018), the results are grouped to be comparable with Figure 6.

Potential for improving the performance of STHM for gauged and ungauged basins

The temporal (Figures 2-3) and spatial (Figures 4-**??**fig:fig6) validation results shown from the application of STHM predicted the floods using higher-level three spatial groups as shown in Figure 1, which was implemented primarily to support a continental scale flood prediction and to reduce the required computation time. (Note that computing all the groups' cross validation costs approximately 6 hours.) Even though the STHM improved the NWM prediction for more than 55% of the basins, it was important to note that the STHM did not improve the flood prediction for the remaining 45% of basins in few groups (e.g., Group 1 natural basins, and Group 2 coastal basins) (Figure 6). This was primarily because the temporal and spatial hierarchies defined under those groups did not aid in improving the model performance as the spatially and temporally varying intercept term ($\beta_{0(\tau,i,j,k)}$), in equation (1) was not explained by the predictors defined in next two-level hierarchies (equations (2) and (3)). This implies that the spatio-temporal variability of $\beta_{0(\tau,i,j,k)}$ was too large or the predictors in the next hierarchies do not correspond to that variability. To demonstrate this point, we considered two experiments for three moderately performing categories – Group-1 controlled and natural and Group 2 coastal – under $k(20)$ spatial cross-validation. The first experiment was performed by fitting the STHM across basin of similar type (i.e., natural/ controlled/ coastal) in a given HUC02 (eight HUC02s in Group 1 and four HUC02s in Group 2) under spatial validation and the results were aggregated to the group level (Figure 7). The second experiment was performed by fitting the STHM for each basin under spatial validation and the results were aggregated to the group level (Figure 8). Thus, in the second experiment, there won't be any hierarchies as defined by the equations (2) and (3) and the intercept term, $\beta_{0(\tau,i,j,k)}$, was simply left as a basin specific-term varying every 10 days.

From Figure 7, the performance of STHM fitted at HUC02 levels significantly improved model prediction performance as a percentage of improved sites for the three selected groups compared to the performance shown in Figure 6. One could argue this comes from the increased number of parameters fitted in explaining the spatio-temporal variability of $\beta_{0(\tau,i,j,k)}$. However, this was daily high streamflow prediction, the number of data points (i.e., daily high streamflow) available for fitting the STHM across the fitting period (1990-2008) was quite large, hence we didn't interpret this as a result of overfitting. It implied that the spatio-temporal hierarchies defined in equations (2) and (3) are not explained by the selected predictors at the broad group level under that category. Under the second experiment, the performance of STHM further improved when the STHM was fitted with no hierarchies, but the performance of the basin was still evaluated under $k(20)$-folder spatial cross-validation (Figure 8). This implied that observed streamflow at a particular site was not used for fitting the parameters in equation (1), only the remaining 80% of those basins in that category was used for fitting the STHM for evaluating the model at a given site. From Figure 8, it was clear that the performance of STHM further improved compared to Figure 7. Figure 8 also could also be improved further by fitting the STHM directly at each site without spatial validation and the hierarchies in equations (2) and (3). We did not perform that experiment as that model will not have applicability for ungauged basins. Thus, by redefining the group or regions in fitting the STHM, the performance of STHM could be improved further. However, this improvement also comes with additional limitations. The computational time for running the STHM also increased by 6 times and 15 times for obtaining the NSE in Figures 7 and 8 respectively. Another limitation of at-site evaluation of STHM (Figure 8) was that it limits the model applicability for ungauged basins as it will require obtaining the parameters by grouping of basins that are similar to the ungauged basin and developing such grouped basin to estimate the STHM for any ungauged basin will be a humongous computational task. Overall, the improved performance of STHM that we observed in Figures 7 and 8 come as a trade-off in fitting the model for regional performance versus at-

site performance. Such issues had been addressed in the context of regional versus at-site calibration of hydrological models (Fernandez et al., 2000). We discussed these issues further in the Discussion.

4 Discussion

This study focused on developing a spatio-temporal hierarchical model (STHM) for flood prediction in gauged and ungauged basins across the CONUS. For this purpose, the study used hydroclimatic (e.g., AI, PET) information and basin characteristics (e.g., imperviousness) along with NWM predicted streamflow to estimate the floods in natural, controlled and coastal basins. The proposed STHM was evaluated under split-sample temporal validation (Figure 2) to understand the role of different multi-level predictors (Figure 3) and using $k(20)$-fold spatial validation (Figure 4) for understanding the utility in flood prediction in ungauged basins over three grouped regions. Both temporal and spatial validation indicated the STHM ability to improve NWM streamflow prediction among different basin types (Figure 2, Figure 4). The spatial cross-validation results indicated the robustness of the model in predicting ungauged basins. The STHM greatly improved the performance of NWM predicted high streamflow for more than 52% of basins, resulting in a 0.1 improvement in NSE. This improvement is important for flood forecast systems looking to provide accurate and reliable information to the public. The model improved most in control and natural basins, particularly, in Group 2 and 3 during colder seasons (SON, DJF). The underperformance in coastal basins could be influenced by lunar tides forcing a lagged runoff, particularly on the east Coast (Cerveny et al., 2010). However, compared to previous studies focusing on improving long-term mean annual streamflow predictions (Alexander et al., 2019a, 2019b) and natural basin alone (Frame et al., 2021), our model showed strong performance in improving finer-scale, daily streamflow across the CONUS. Instead of including many lagged variables from NWM as predictors (e.g., Woznicki et al., 2019), our model only selected a few key drivers of hydroclimate that are well founded in the literature on streamflow prediction and considers the concurrent NWM streamflow prediction. Further, the proposed STHM also provides improved predictions for both controlled and coastal basins. This gives our model the flexibility to be easily expanded to predict floods at the CONUS scale for both gauged and ungauged basins.

It is important to note that STHM predictions performed worse than the original NWM streamflow in 25%-45% of basins (Figure 4). To address concerns regarding this, we performed two experiments that fitted the STHM at each HUC02 (Figure 7) and for individual basins (Figure 8) for Group 1 (controlled and natural) and Group 2 coastal basins. This resulted in a significant improvement with most basins performing better than NWM performance. This indicates that poor performance of STHM in Figure 4 is primarily due to the trade-off in improving the regional performance of the model at the cost of at-site performance (Fernandez et al., 2000). However, fitting the STHM purely for at-site would limit its ability to predict in ungauged predictions. In a traditional hierarchical modeling approach, this would be considered as an "unpooled regression" model (Das Bhowmik et al., 2020; Devineni et al., 2013) as such a model will result with no regional modeling terms. It can be easily understood that post-processing a model's flow would naturally result in improved model performance as regression is expected to reduce the model bias. Thus, our proposed STHM could be fitted after a reasonable grouping of basins or HUC02- regions so that the resulting regional model parameters (i.e., equations 2 and 3) would provide useful information for ungauged basins prediction.

It is important to note that our model mainly relies on basin characteristics (e.g., imperviousness) and hydroclimatic information (e.g., AI, phase correlation), which could be obtained based on widely available database mentioned in the data section for any ungauged basin. However, obtaining antecedent streamflow conditions, Q^{3d}, of the basin is difficult for ungauged basins. Hence, in our study, we obtained Q^{3d} purely based on

gauged basins available at the HUC08 level to get the depth of runoff and convert it to runoff based on the ungauged basins' drainage area (equations 5 and 6). However, this step could be eliminated for gauged basins as one could use the observed 3-day streamflow itself for estimating the antecedent conditions. We also would like to mention that this 3-day could be improved using the stage information available from remote-sensing satellites, e.g., the Global Flood Detection System (`https://www.gdacs.org/flooddetection/`). This also could also potentially extend the STHM into a forecasting model if one were to use the real-time NWM forecasts. Thus, STHM could be utilized for real-time flood forecasting for both ungauged and gauged basins.

We utilized NOAA's NWM reanalysis runs for evaluating the proposed STHM ability in predicting the ungauged basins as these are immediately available over the entire CONUS. However, in principle, STHM could be fitted with any other hydrologic model outputs such as Variable Infiltration Capacity model (Liang et al., 1996) or SWAT (Arnold et al., 2012). Similarly, the NWM prediction could also be replaced with basin-level precipitation. Recently, Frame et al. (2021) utilized atmospheric forcings alone, instead of NWM streamflow predictions, as a predictor for predicting streamflow and found that the Long Short-Term memory model (LSTM) performed equally well as that of LSTM trained with NWM streamflow. The proposed STHM modeling structure is also hierarchical and semi-parametric as its parameters vary over 10-day moving window, which makes it to estimate the non-linear dependencies between streamflow and the relevant predictors consisting of basin characteristics and hydroclimatic information. This indicates that there is potential for extending the STHM with other distributed hydrologic model outputs and/or with atmospheric forcings that drive the hydrologic models.

5 Summary and Conclusions

We describe a hierarchical spatial-temporal post-processing model for improving flood prediction in both gauged and ungauged basins across the CONUS. The proposed STHM is hierarchical and semi-parametric, thereby having the ability to predict non-linear dependencies between streamflow and the predictors – NWM streamflow, basin characteristics, upstream reservoir storage and hydroclimatic information – for estimating floods in natural, controlled and coastal basins over the CONUS. Performance evaluation of the hierarchical model showed that increased predictive skill in over 50% of sites' by 0.1 NSE, and improved over 65% of sites' streamflow prediction to "good" (NSE>0.67). For controlled basins, the primary improvement was due to the inclusion of areal averaged previous 3-day flow, which accounts for 18% of the variability of high streamflow over all regions. But the explained variability of high streamflow for coastal basins are only limited to 10% due to other unconsidered factors, e.g., tidal influence. For natural basins, the biggest improvement by the models is due to the inclusion of predictors such as aridity index and phase correlation for extending the STHM for ungauged prediction. We also demonstrated that the reduced performance of STHM in several basins also stem the trade-off in parameter estimation between at-site improvement versus the regional performance, which is required particularly for ungauged basins prediction. Performance evaluation of the STHM under temporal and spatial the cross-validation results has shown robustness in predicting floods under "ungauged" prediction mode.

In addition to improved flood prediction, the developed model was also rigorously evaluated in predicting floods for ungauged basins through k-fold spatial validation. Even though the STHM used NWM streamflow as a predictor, the model could be recalibrated with any other hydrologic model outputs or with precipitation and relevant atmospheric forcings. Further, the proposed STHM also post-processes the NWM prediction, thereby reducing the systematic biases in the model prediction. Since the STHM predictors are widely available for both any given site (e.g., NWM prediction and previous 3-day streamflow) with spatially and temporally varying predictors, we can apply the estimated model coefficients to any ungauged site using the corresponding HUC08 level parameters.Given

that the NWM for real-time streamflow forecasts are available for any locations within the US, the proposed STHM could be employed for real-time streamflow forecasts for both gauged and ungauged basins. The proposed modeling approach is also hybrid as it combines physical modeling outputs with statistical modeling for developing streamflow prediction across the CONUS. These hybrid approaches are essential as real-time weather forecasts always have considered both dynamical model predictions with statistical correction scheme, which is popularly known as Model Output Statistics, in the weather forecasting community (Antolik, 2000). Thus, the STHM could be eventually employed for both ungauged flood prediction as well as for issuing real-time flood forecasts.

Open Research Section

The data for this paper are available in the following Zenodo repository: `https://doi.org/10.5281/zenodo.7574439`

Acknowledgments

This research was primarily funded by the National Science Foundation Convergence Accelerator (No. 2033607)

References

Alexander, R. B., Schwarz, G. E., & Boyer, E. W. (2019a). Advances in quantifying streamflow variability across continental scales: 1. identifying natural and anthropogenic controlling factors in the usa using a spatially explicit modeling method [Journal Article]. *Water Resources Research*, *55*(12), 10893-10917.

Alexander, R. B., Schwarz, G. E., & Boyer, E. W. (2019b). Advances in quantifying streamflow variability across continental scales: 2. improved model regionalization and prediction uncertainties using hierarchical bayesian methods [Journal Article]. *Water Resources Research*, *55*(12), 11061-11087. doi: https://doi.org/10.1029/2019WR025037

Antolik, M. S. (2000). An overview of the national weather service's centralized statistical quantitative precipitation forecasts [Journal Article]. *Journal of Hydrology*, *239*(1-4), 306-337.

Archfield, S. A., Clark, M., Arheimer, B., Hay, L. E., McMillan, H., Kiang, J. E., ... Wagener, T. (2015). Accelerating advances in continental domain hydrologic modeling [Journal Article]. *Water Resources Research*, *51*(12), 10078-10091.

Arnold, J. G., Moriasi, D. N., Gassman, P. W., Abbaspour, K. C., White, M. J., Srinivasan, R., ... Van Liew, M. W. (2012). Swat: Model use, calibration, and validation [Journal Article]. *Transactions of the ASABE*, *55*(4), 1491-1508.

Arnone, E., Pumo, D., Francipane, A., La Loggia, G., & Noto, L. V. (2018). The role of urban growth, climate change, and their interplay in altering runoff extremes [Journal Article]. *Hydrological Processes*, *32*(12), 1755-1770.

Browne, M. W. (2000). Cross-validation methods [Journal Article]. *Journal of mathematical psychology*, *44*(1), 108-132.

Brunner, M. I., Melsen, L. A., Newman, A. J., Wood, A. W., & Clark, M. P. (2020). Future streamflow regime changes in the united states: assessment using functional classification [Journal Article]. *Hydrology and Earth System Sciences*, *24*(8), 3951-3966. doi: 10.5194/hess-24-3951-2020

Caldwell, P., Sun, G., McNulty, S., Cohen, E., & Moore Myers, J. (2012). Impacts of impervious cover, water withdrawals, and climate change on river flows in the conterminous us [Journal Article]. *Hydrology and Earth System Sciences*, *16*(8), 2839-2857. doi: https://doi.org/10.1029/2006WR005855

Cerveny, R. S., Svoma, B. M., & Vose, R. S. (2010). Lunar tidal influence on

inland river streamflow across the conterminous united states [Journal Article]. *Geophysical Research Letters*, *37*(22). doi: https://doi.org/10.1029/2010GL045564

Chalise, D. R., Sankarasubramanian, A., & Ruhi, A. (2021). Dams and climate interact to alter river flow regimes across the united states [Journal Article]. *Earth's Future*, *9*(4), e2020EF001816. doi: https://doi.org/10.1029/2020EF001816

Cloke, H. L., & Pappenberger, F. (2009). Ensemble flood forecasting: A review [Journal Article]. *Journal of Hydrology*, *375*(3), 613-626. doi: https://doi.org/10.1016/j.jhydrol.2009.06.005

Das Bhowmik, R., Ng, T. L., & Wang, J.-P. (2020). Understanding the impact of observation data uncertainty on probabilistic streamflow forecasts using a dynamic hierarchical model [Journal Article]. *Water Resources Research*, *56*(4), e2019WR025463. doi: https://doi.org/10.1029/2019WR025463

Devineni, N., Lall, U., Pederson, N., & Cook, E. (2013). A tree-ring-based reconstruction of delaware river basin streamflow using hierarchical bayesian regression [Journal Article]. *Journal of Climate*, *26*(12), 4357-4374. doi: 10.1175/jcli-d-11-00675.1

Devineni, N., Sankarasubramanian, A., & Ghosh, S. (2008). Multimodel ensembles of streamflow forecasts: Role of predictor state in developing optimal combinations [Journal Article]. *Water resources research*, *44*(9).

Emerton, R. E., Stephens, E. M., Pappenberger, F., Pagano, T. C., Weerts, A. H., Wood, A. W., ... Cloke, H. L. (2016). Continental and global scale flood forecasting systems [Journal Article]. *WIREs Water*, *3*(3), 391-418. doi: https://doi.org/10.1002/wat2.1137

Evenson, G. R., Kalcic, M., Wang, Y.-C., Robertson, D., Scavia, D., Martin, J., ... Brooker, M. (2021). Uncertainty in critical source area predictions from watershed-scale hydrologic models [Journal Article]. *Journal of Environmental Management*, *279*, 111506.

Fernandez, W., Vogel, R., & Sankarasubramanian, A. (2000). Regional calibration of a watershed model [Journal Article]. *Hydrological sciences journal*, *45*(5), 689-707.

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 [Journal Article]. *JAWRA Journal of the American Water Resources Association*.

Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models* [Book]. Cambridge university press.

Ghotbi, S., Wang, D., Singh, A., Mayo, T., & Sivapalan, M. (2020). Climate and landscape controls of regional patterns of flow duration curves across the continental united states: Statistical approach [Journal Article]. *Water Resources Research*, *56*(11), e2020WR028041. doi: https://doi.org/10.1029/2020WR028041

Gierszewski, P. J., Habel, M., Szmańda, J., & Luc, M. (2020). Evaluating effects of dam operation on flow regimes and riverbed adaptation to those changes [Journal Article]. *Science of The Total Environment*, *710*, 136202. doi: https://doi.org/10.1016/j.scitotenv.2019.136202

Groll, A., & Tutz, G. (2014). Variable selection for generalized linear mixed models by l 1-penalized estimation [Journal Article]. *Statistics and Computing*, *24*(2), 137-154.

Grömping, U. (2007). Estimators of relative importance in linear regression based on variance decomposition [Journal Article]. *The American Statistician*, *61*(2), 139-147.

Gudmundsson, L., Greve, P., & Seneviratne, S. I. (2016). The sensitivity of water availability to changes in the aridity index and other factors—a probabilistic

analysis in the budyko space [Journal Article]. *Geophysical Research Letters*, *43*(13), 6985-6994. doi: https://doi.org/10.1002/2016GL069763

Heidari, H., Arabi, M., Warziniack, T., & Kao, S.-C. (2020). Assessing shifts in regional hydroclimatic conditions of u.s. river basins in response to climate change over the 21st century [Journal Article]. *Earth's Future*, *8*(10), e2020EF001657. doi: https://doi.org/10.1029/2020EF001657

Heng, S., & Suetsugi, T. (2013). Using artificial neural network to estimate sediment load in ungauged catchments of the tonle sap river basin, cambodia [Journal Article]. *Journal of Water Resource and Protection*.

Hirsch, R., & De Cicco, L. (2017). *dataretrieval-retrieval functions for usgs and epa hydrologic and water quality data* [Generic].

Hrachowitz, M., Savenije, H. H. G., Blöschl, G., McDonnell, J. J., Sivapalan, M., Pomeroy, J. W., ... Cudennec, C. (2013). A decade of predictions in ungauged basins (pub)—a review [Journal Article]. *Hydrological Sciences Journal*, *58*(6), 1198-1255. doi: 10.1080/02626667.2013.803183

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 Article]. *JAWRA Journal of the American Water Resources Association*, *54*(4), 872-881. doi: https://doi.org/10.1111/1752-1688.12645

Johnson, J. M., Fang, S., Sankarasubramanian, A., Rad, A. M., da Cunha, L. K., Clarke, K. C., ... Yeghiazarian, L. (2023). Comprehensive analysis of the noaa national water model: A call for heterogeneous formulations and diagnostic model selection. *ESS Open Archive*. doi: DOI:10.22541/essoar.167415214.45806648/v1

Kumar, H., Hwang, J., Devineni, N., & Sankarasubramanian, A. (2022). Dynamic flow alteration index for complex river networks with cascading reservoir systems [Journal Article]. *Water Resources Research*, *58*(1), e2021WR030491. doi: https://doi.org/10.1029/2021WR030491

Li, W., Duan, Q., Miao, C., Ye, A., Gong, W., & Di, Z. (2017). A review on statistical postprocessing methods for hydrometeorological ensemble forecasting [Journal Article]. *WIREs Water*, *4*(6), e1246. doi: https://doi.org/10.1002/wat2.1246

Liang, X., Lettenmaier, D. P., & Wood, E. F. (1996). One-dimensional statistical dynamic representation of subgrid spatial variability of precipitation in the two-layer variable infiltration capacity model [Journal Article]. *Journal of Geophysical Research: Atmospheres*, *101*(D16), 21403-21422.

Mazrooei, A., Sankarasubramanian, A., & Lakshmi, V. (2020). Technical note: Evaluation of the skill in monthly-to-seasonal soil moisture forecasting based on smap satellite observations over the southeastern us [Journal Article]. *Hydrol. Earth Syst. Sci.*, *24*(3), 1073-1079. doi: 10.5194/hess-24-1073-2020

Mazrooei, A., Sankarasubramanian, A., & Wood, A. W. (2021). Potential in improving monthly streamflow forecasting through variational assimilation of observed streamflow [Journal Article]. *Journal of Hydrology*, *600*, 126559. doi: https://doi.org/10.1016/j.jhydrol.2021.126559

McCuen, R. H., Knight, Z., & Cutter, A. G. (2006). Evaluation of the nash–sutcliffe efficiency index [Journal Article]. *Journal of hydrologic engineering*, *11*(6), 597-602.

Mendoza, P. A., McPhee, J., & Vargas, X. (2012). Uncertainty in flood forecasting: A distributed modeling approach in a sparse data catchment [Journal Article]. *Water Resources Research*, *48*(9). doi: https://doi.org/10.1029/2011WR011089

Merz, B., Blöschl, G., Vorogushyn, S., Dottori, F., Aerts, J. C. J. H., Bates, P., ... Macdonald, E. (2021). Causes, impacts and patterns of disastrous river floods [Journal Article]. *Nature Reviews Earth & Environment*, *2*(9), 592-609. doi: 10.1038/s43017-021-00195-3

Mishra, A. K., & Coulibaly, P. (2009). Developments in hydrometric network design: A review [Journal Article]. *Reviews of Geophysics*, *47*(2).

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 [Journal Article]. *Transactions of the ASABE*, *50*(3), 885-900.

Narbondo, S., Gorgoglione, A., Crisci, M., & Chreties, C. (2020). Enhancing physical similarity approach to predict runoff in ungauged watersheds in subtropical regions [Journal Article]. *Water*, *12*(2), 528.

National Research Council. (2006). *Toward a new advanced hydrologic prediction service (ahps)* [Book]. National Academies Press.

National Weather Service. (2022). *Us flood fatality statistics* (Tech. Rep.). National Oceanic and Atmospheric Administration. Retrieved from `https://www.weather.gov/arx/usflood`

Ossandón, A., Rajagopalan, B., Tiwari, A. D., Thomas, T., & Mishra, V. (2022). A bayesian hierarchical model combination framework for real-time daily ensemble streamflow forecasting across a rainfed river basin [Journal Article]. *Earth's Future*, e2022EF002958.

Pappenberger, F., Cloke, H. L., Parker, D. J., Wetterhall, F., Richardson, D. S., & Thielen, J. (2015). The monetary benefit of early flood warnings in europe [Journal Article]. *Environmental Science & Policy*, *51*, 278-291. doi: https://doi.org/10.1016/j.envsci.2015.04.016

Parajka, J., Viglione, A., Rogger, M., Salinas, J., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins–part 1: Runoff-hydrograph studies [Journal Article]. *Hydrology and Earth System Sciences*, *17*(5), 1783-1795.

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 Article]. *Journal of hydrology*, *468*, 139-150.

Petersen, T., Devineni, N., & Sankarasubramanian, A. (2018). Monthly hydroclimatology of the continental united states [Journal Article]. *Advances in Water Resources*, *114*, 180-195.

Pörtner, H. O., Roberts, D. C., Adams, H., Adler, C., Aldunce, P., Ali, E., ... Biesbroek, R. (2022). Climate change 2022: impacts, adaptation and vulnerability [Journal Article]. *IPCC Sixth Assessment Report*.

Ramaswamy, V., Rao, P. S., Rao, K. H., Thwin, S., Rao, N. S., & Raiker, V. (2004). Tidal influence on suspended sediment distribution and dispersal in the northern andaman sea and gulf of martaban [Journal Article]. *Marine Geology*, *208*(1), 33-42. doi: https://doi.org/10.1016/j.margeo.2004.04.019

Ravindranath, A., Devineni, N., Lall, U., Cook, E. R., Pederson, G., Martin, J., & Woodhouse, C. (2019). Streamflow reconstruction in the upper missouri river basin using a novel bayesian network model [Journal Article]. *Water Resources Research*, *55*(9), 7694-7716.

Rezaie-Balf, M., Fani Nowbandegani, S., Samadi, S. Z., Fallah, H., & Alaghmand, S. (2019). An ensemble decomposition-based artificial intelligence approach for daily streamflow prediction [Journal Article]. *Water*, *11*(4), 709.

Salinas, J. L., Laaha, G., Rogger, M., Parajka, J., Viglione, A., Sivapalan, M., & Blöschl, G. (2013). Comparative assessment of predictions in ungauged basins - part 2: Flood and low flow studies [Journal Article]. *Hydrology and Earth System Sciences*, *17*(7), 2637-2652. doi: 10.5194/hess-17-2637-2013

Sankarasubramanian, A., & Lall, U. (2003). Flood quantiles in a changing climate: Seasonal forecasts and causal relations [Journal Article]. *Water Resources Research*, *39*(5).

Sankarasubramanian, A., & Vogel, R. M. (2002). Annual hydroclimatology of the

united states [Journal Article]. *Water Resources Research*, *38*(6), 19-1-19-12. doi: https://doi.org/10.1029/2001WR000619

Schelldorfer, J., Meier, L., & Bühlmann, P. (2014). Glmmlasso: an algorithm for high-dimensional generalized linear mixed models using $\ell 1$-penalization [Journal Article]. *Journal of Computational and Graphical Statistics*, *23*(2), 460-477.

Singh, H., & Sankarasubramanian, A. (2014). Systematic uncertainty reduction strategies for developing streamflow forecasts utilizing multiple climate models and hydrologic models [Journal Article]. *Water Resources Research*, *50*(2), 1288-1307. doi: https://doi.org/10.1002/2013WR013855

Sinha, T., & Sankarasubramanian, A. (2013). Role of climate forecasts and initial conditions in developing streamflow and soil moisture forecasts in a rainfall–runoff regime [Journal Article]. *Hydrology and Earth System Sciences*, *17*(2), 721-733.

Tamaddun, K. A., Kalra, A., & Ahmad, S. (2019). Spatiotemporal variation in the continental us streamflow in association with large-scale climate signals across multiple spectral bands [Journal Article]. *Water Resources Management*, *33*(6), 1947-1968. doi: 10.1007/s11269-019-02217-8

Tijerina, D., Condon, L., FitzGerald, K., Dugger, A., O'Neill, M. M., Sampson, K., ... Maxwell, R. (2021). Continental hydrologic intercomparison project, phase 1: A large-scale hydrologic model comparison over the continental united states [Journal Article]. *Water Resources Research*, *57*(7), e2020WR028931.

Turner, S. W., Doering, K., & Voisin, N. (2020). Data-driven reservoir simulation in a large-scale hydrological and water resource model [Journal Article]. *Water Resources Research*, *56*(10), e2020WR027902.

USACE. (2022). *Us army corps of engineers national inventory of dams* [Generic]. Retrieved from `https://nid.sec.usace.army.mil/#/`

Viterbo, F., Mahoney, K., Read, L., Salas, F., Bates, B., Elliott, J., ... Cifelli, R. (2020). A multiscale, hydrometeorological forecast evaluation of national water model forecasts of the may 2018 ellicott city, maryland, flood [Journal Article]. *Journal of Hydrometeorology*, *21*(3), 475-499. doi: 10.1175/jhm-d-19-0125.1

Vogel, R. M., & Sankarasubramanian, A. (2000). Spatial scaling properties of annual streamflow in the united states [Journal Article]. *Hydrological sciences journal*, *45*(3), 465-476.

Wallemacq, P., & House, R. (2018). *Economic losses, poverty & disasters 1998-2017* [techreport]. Centre for Research on the Epidemiology of Disasters and United Nations Office for Disaster Risk Reduction.

Woznicki, S. A., Baynes, J., Panlasigui, S., Mehaffey, M., & Neale, A. (2019). Development of a spatially complete floodplain map of the conterminous united states using random forest [Journal Article]. *Science of The Total Environment*, *647*, 942-953. doi: https://doi.org/10.1016/j.scitotenv.2018.07.353

Wu, R., Yang, L., Chen, C., Ahmad, S., Dascalu, S. M., & Harris Jr, F. C. (2019). Melpf version 1: Modeling error learning based post-processor framework for hydrologic models accuracy improvement [Journal Article]. *Geoscientific Model Development*, *12*(9), 4115-4131. doi: 10.5194/gmd-12-4115-2019

Zhang, F., Sun, J., Lin, B., & Huang, G. (2018). Seasonal hydrodynamic interactions between tidal waves and river flows in the yangtze estuary [Journal Article]. *Journal of Marine Systems*, *186*, 17-28. doi: https://doi.org/10.1016/j.jmarsys.2018.05.005

Zhang, Y., & Chiew, F. H. (2009). Relative merits of different methods for runoff predictions in ungauged catchments [Journal Article]. *Water Resources Research*, *45*(7).

Zhao, Q., & Cai, X. (2020). Deriving representative reservoir operation rules using a hidden markov-decision tree model [Journal Article]. *Advances in Water Resources*, *146*, 103753. doi: https://doi.org/10.1016/j.advwatres.2020.103753

Figure 1.

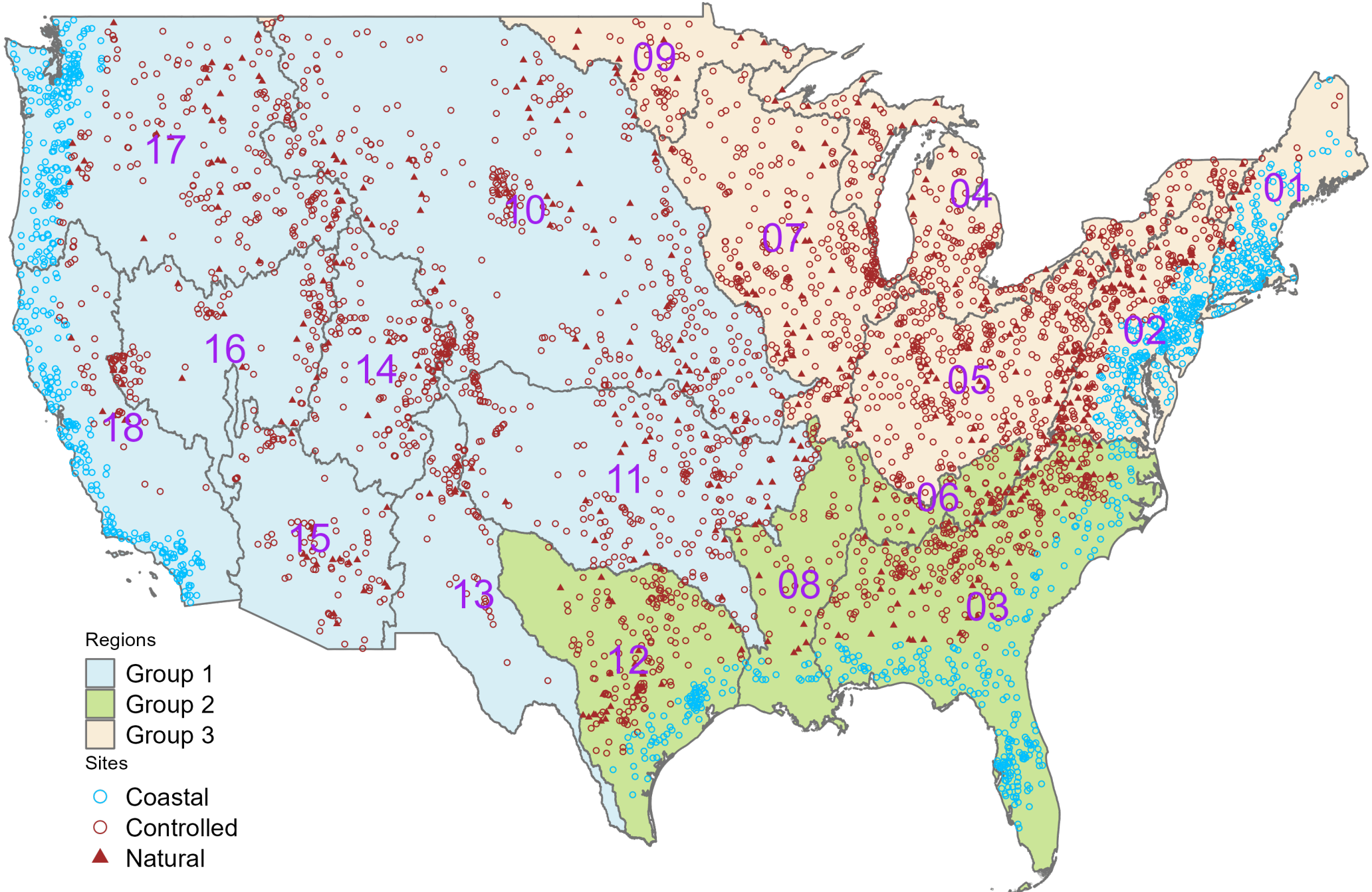

09
17
10
04
01
07
02
16
14
05
18
11
06
15
08
13
03
12
Regions
Group 1
Group 2
Group 3
Sites
Coastal
Controlled
Natural

Figure 2.

model	basintype	NSE>0	NSE>0.5	NSE>0.67
NWM-HM	coastal	68.8%	47.8%	32.9%
NWM-HM	controlled	61.7%	46.1%	34.9%
NWM-HM	natural	90.6%	77.2%	65.1%
NWM-original	coastal	39.4%	14.1%	5.5%
NWM-original	controlled	30.8%	9.2%	3.2%
NWM-original	natural	52.1%	21.3%	9.0%

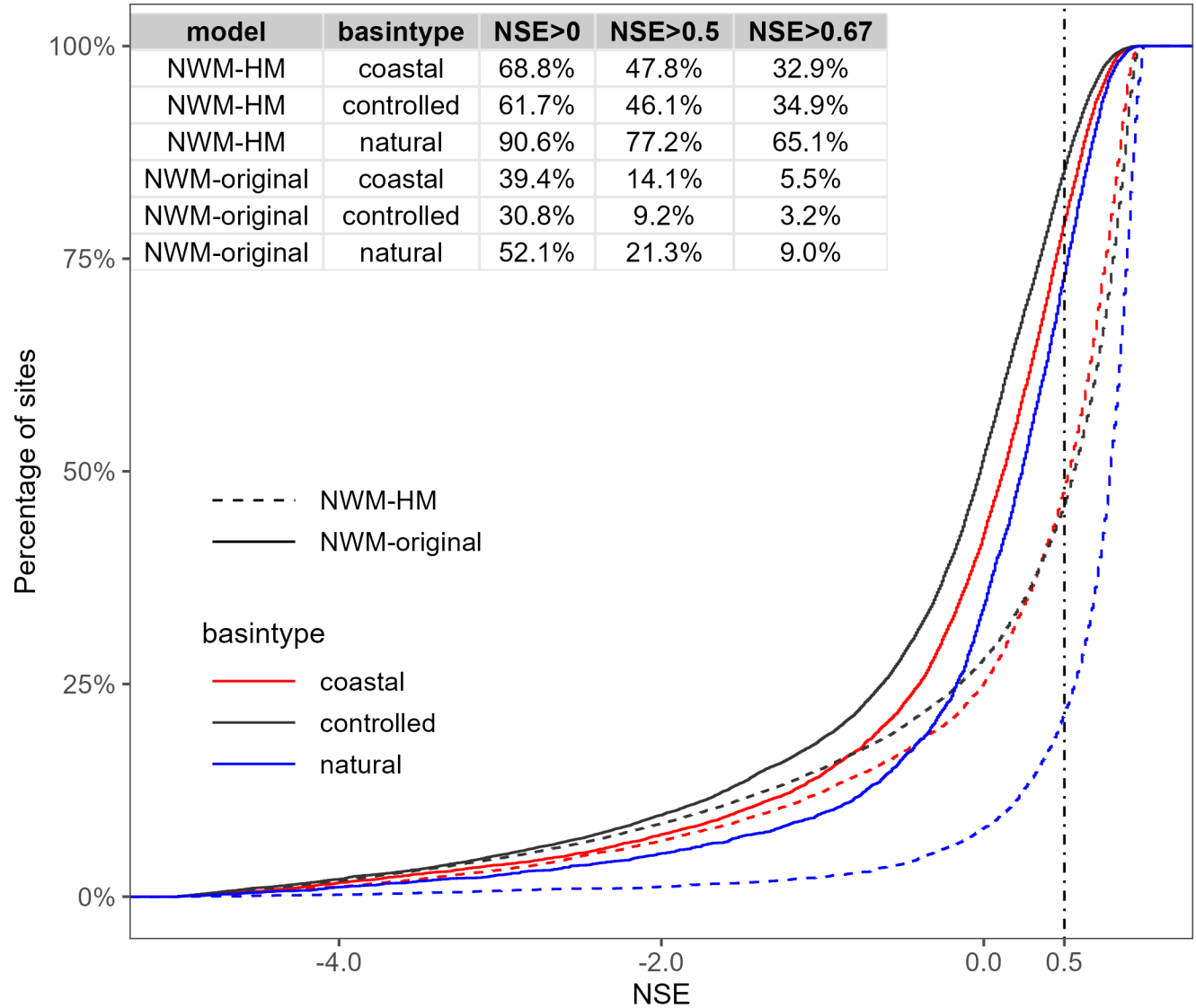


Figure 3.

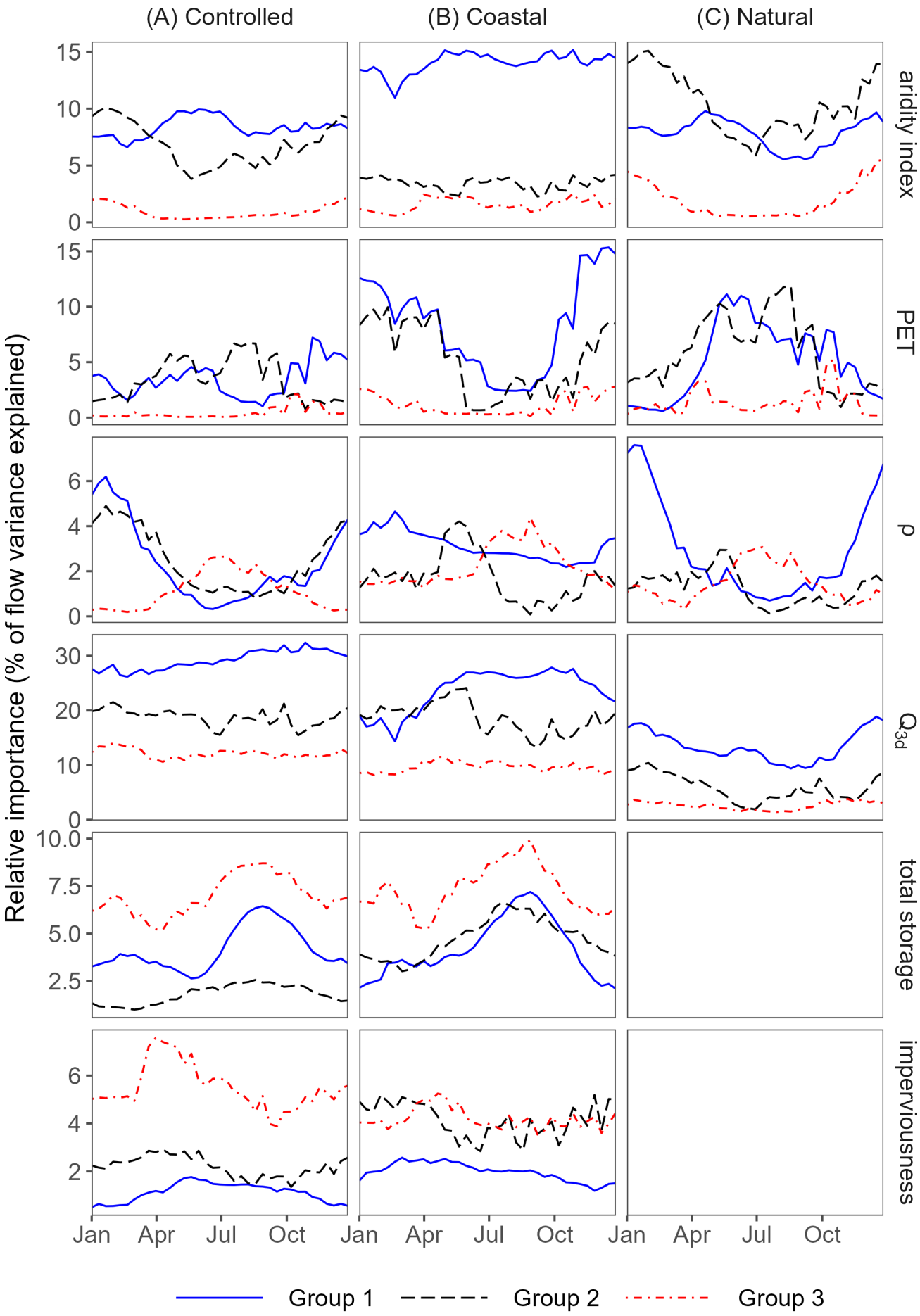

(A) Controlled
(B) Coastal
(C) Natural
aridity index
PET
P
Q_{3d}
total storage
imperviousness
Relative importance (% of flow variance explained)
Jan
Apr
Jul
Oct
Group 1
Group 2
Group 3

Figure 4.

model	basintype	NSE>0	NSE>0.5	NSE>0.67
NWM-HM	coastal	72.6%	53.4%	39.2%
NWM-HM	controlled	57.4%	37.7%	26.0%
NWM-HM	natural	93.8%	83.2%	71.3%
NWM-original	coastal	41.8%	15.4%	6.2%
NWM-original	controlled	31.1%	9.5%	3.5%
NWM-original	natural	52.8%	21.9%	8.5%

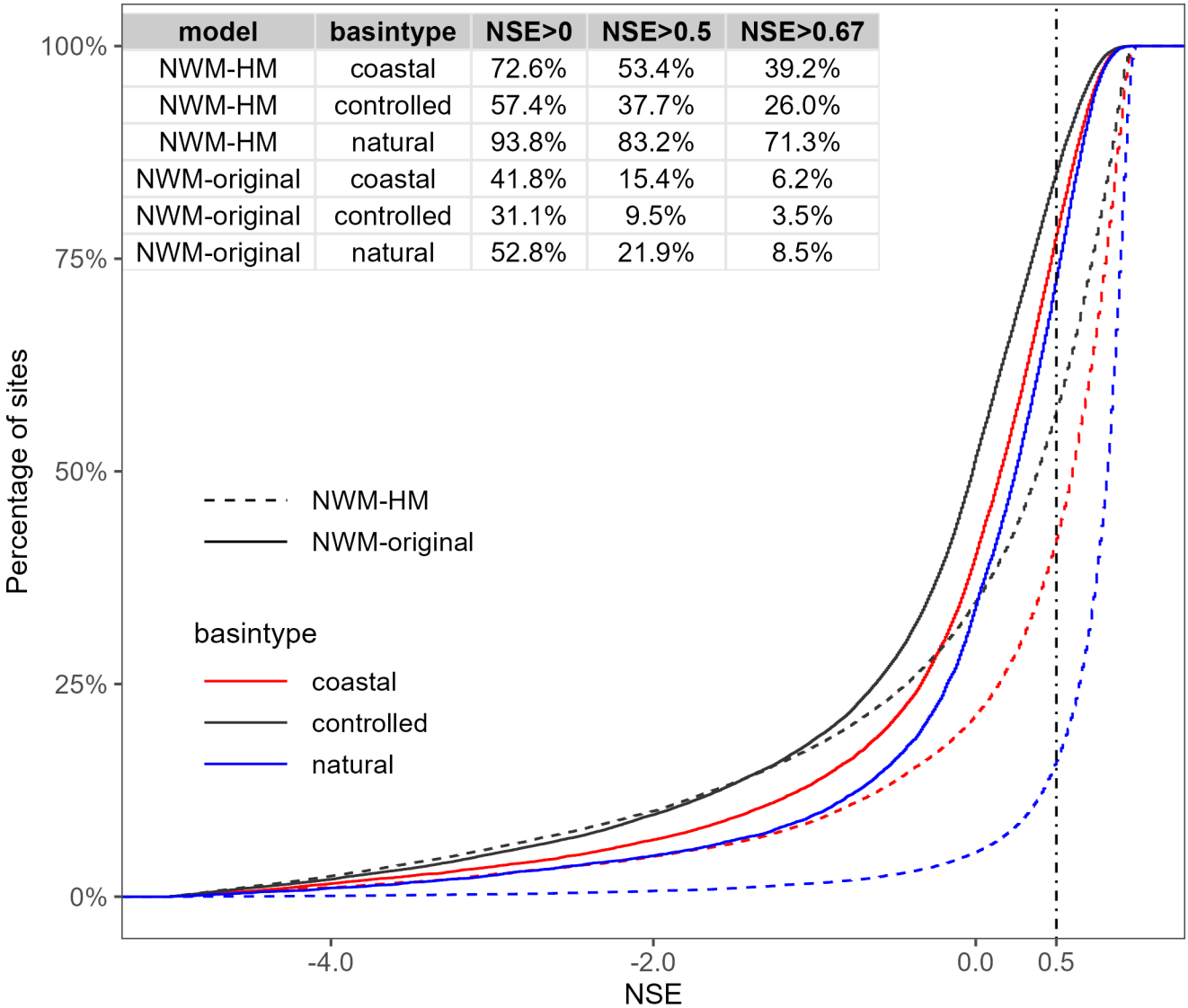


Figure 5.

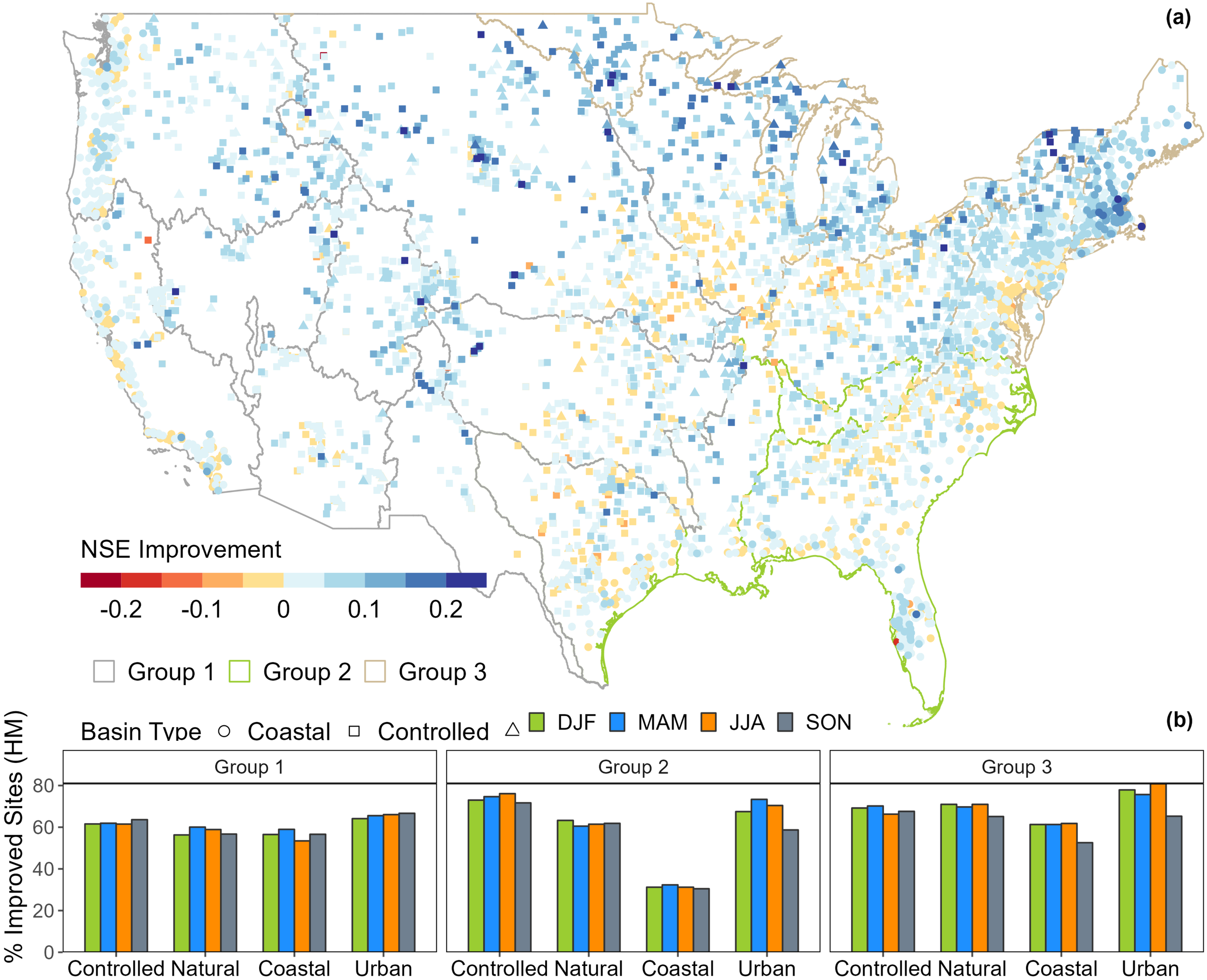

(a)
NSE Improvement
-0.2
-0.1
0
0.1
0.2
Group 1
Group 2
Group 3
(b)
Basin Type
Coastal
Controlled
DJF
MAM
JJA
SON
% Improved Sites (HM)
0
20
40
60
80
Group 1
Group 2
Group 3
Controlled
Natural
Coastal
Urban

Figure 6.

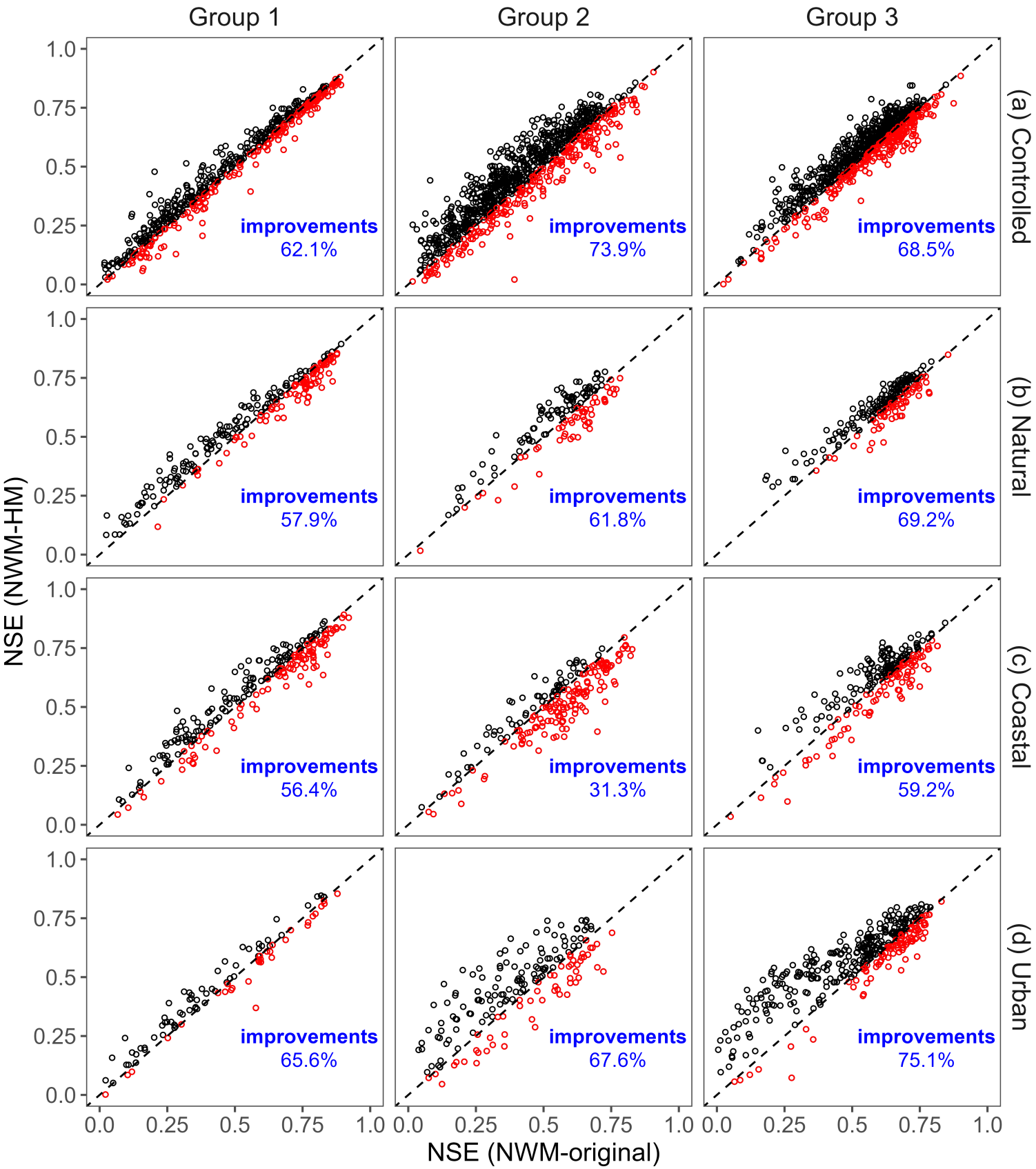

Group 1
Group 2
Group 3
(a) Controlled
(b) Natural
(c) Coastal
(d) Urban
improvements
62.1%
73.9%
68.5%
57.9%
61.8%
69.2%
56.4%
31.3%
59.2%
65.6%
67.6%
75.1%
NSE (NWM-HM)
NSE (NWM-original)
0.0
0.25
0.5
0.75
1.0

Figure 7.

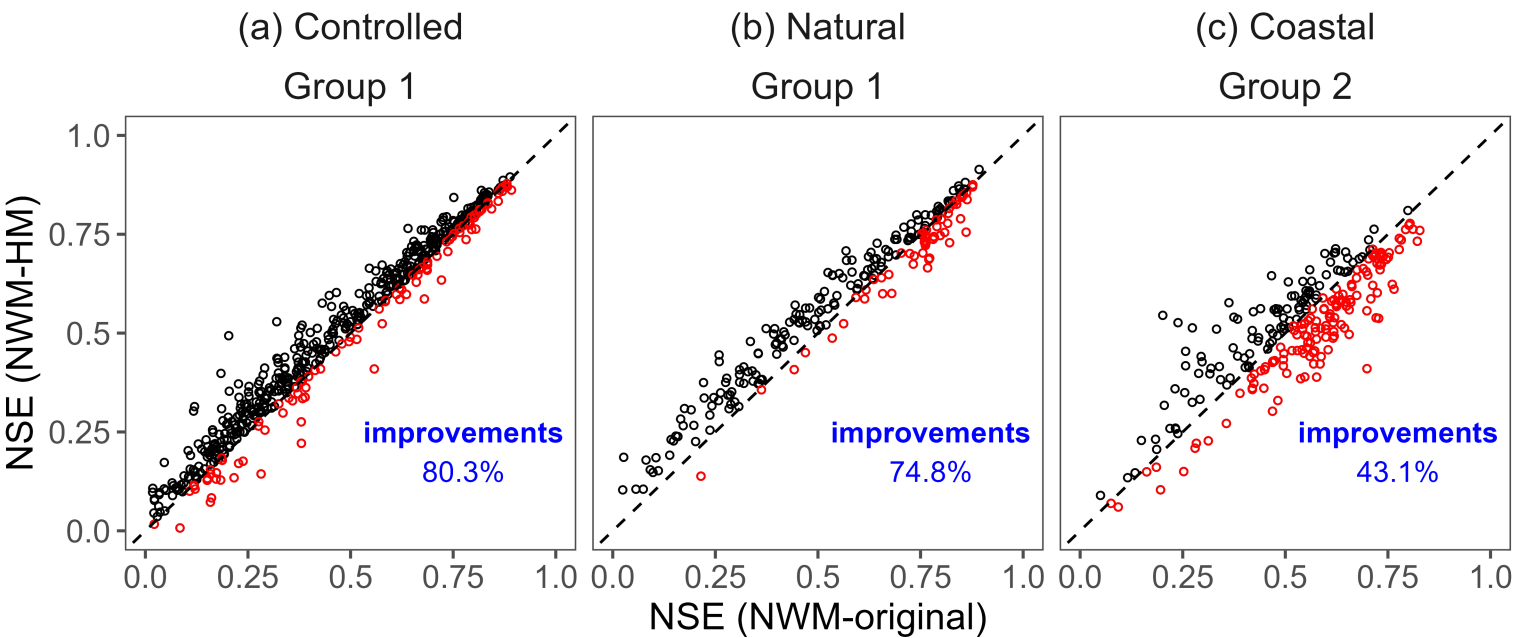
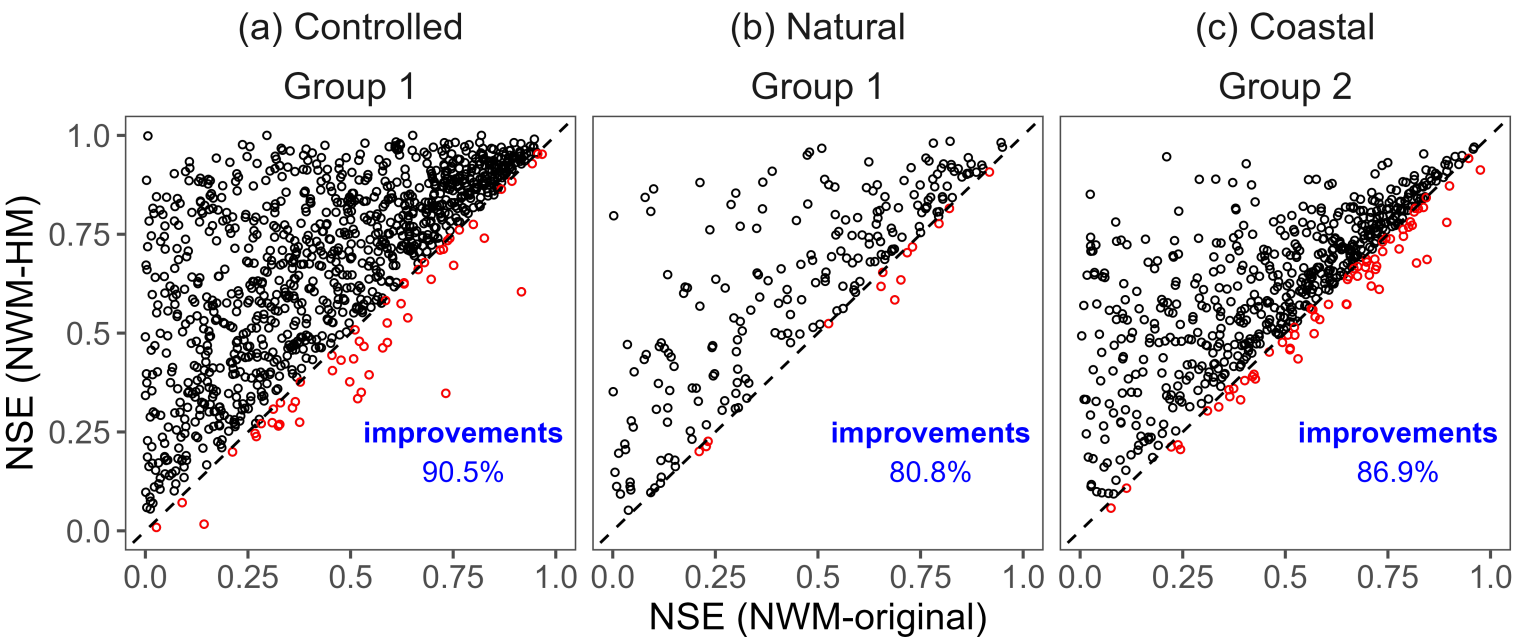

(a) Controlled
Group 1
(b) Natural
Group 1
(c) Coastal
Group 2
NSE (NWM-HM)
NSE (NWM-original)
0.0
0.25
0.5
0.75
1.0
improvements
80.3%
improvements
74.8%
improvements
43.1%

Figure 8.


(a) Controlled
Group 1
(b) Natural
Group 1
(c) Coastal
Group 2
NSE (NWM-HM)
NSE (NWM-original)
0.0
0.25
0.5
0.75
1.0
improvements
90.5%
improvements
80.8%
improvements
86.9%

Supporting Information for "Improved National-Scale Flood Prediction for Gauged and Ungauged Basins using a Spatio-temporal Hierarchical Model"

Shiqi Fang[1], J. Michael Johnson[2], Lilit Yeghiazarian[3], A. Sankarasubramanian[1]

[1]Department of Civil, Construction and Environmental Engineering, North Carolina State University, Raleigh NC, USA

[2]Lynker, Fort Collins CO, USA

[3]Department of Biomedical Chemical and Environmental Engineering, University of Cincinnati, Cincinnati Ohio, USA

January 27, 2023, 10:51am

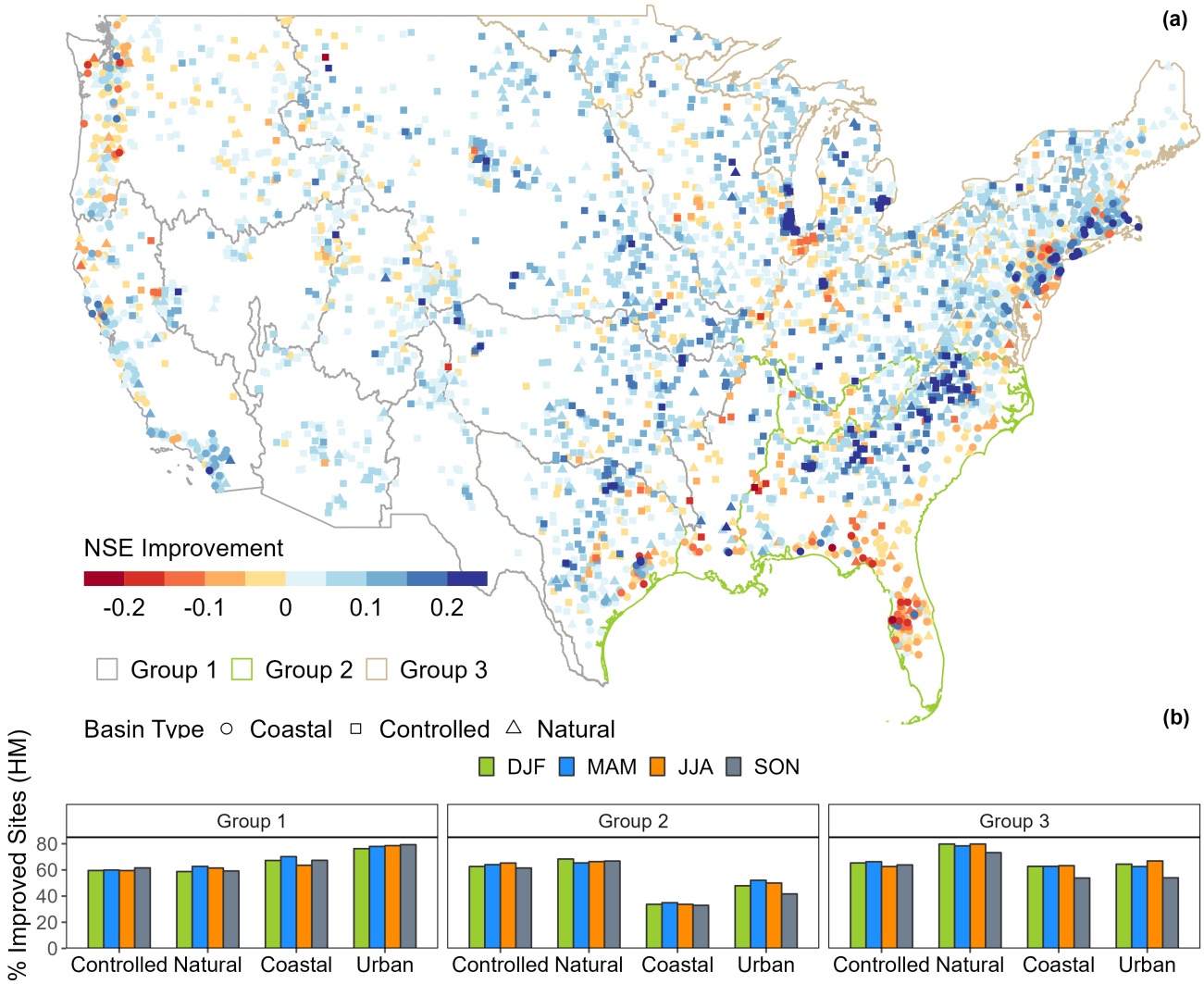


Figure S1. (a) maps the difference in NSE between the HM predicted daily high flows and NWM v2.0 over the validation period (2009-2018). Positive (Negative) values indicate that the hierarchical model (NWM) performed better (worse) in predicting high flows. (b) plots seasonal performance, indicated as 5% of sites with improved NSE during the validation period for the three regions over the four basin classifications.

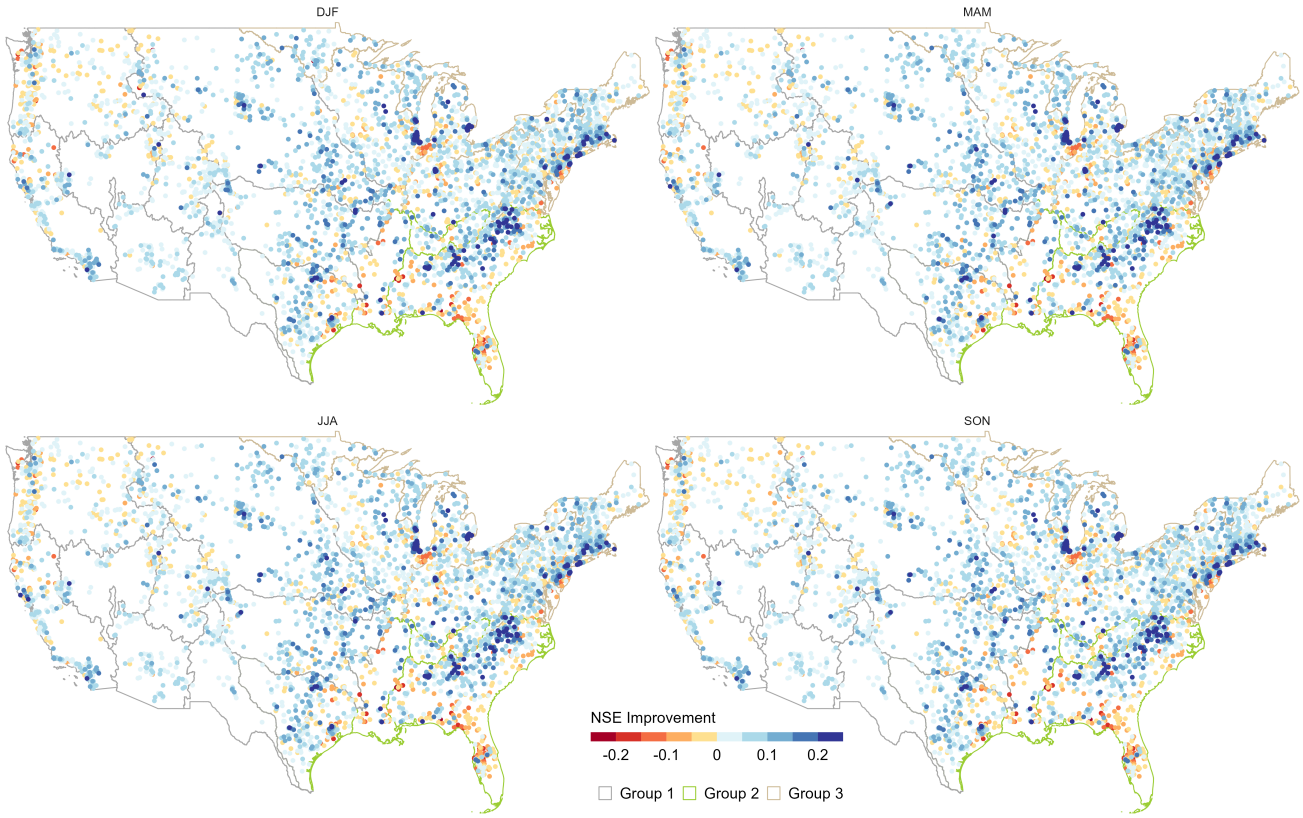


Figure S2. Differences in NSE between the hierarchical model and NWM v2.0 for four seasons (DJF, MAM, JJA and SON) in natural and controlled basins. Data was generated by treating 5% sites as "ungauged" using k(20)-folder cross-validation and NWM daily flows over the period 1993-2018

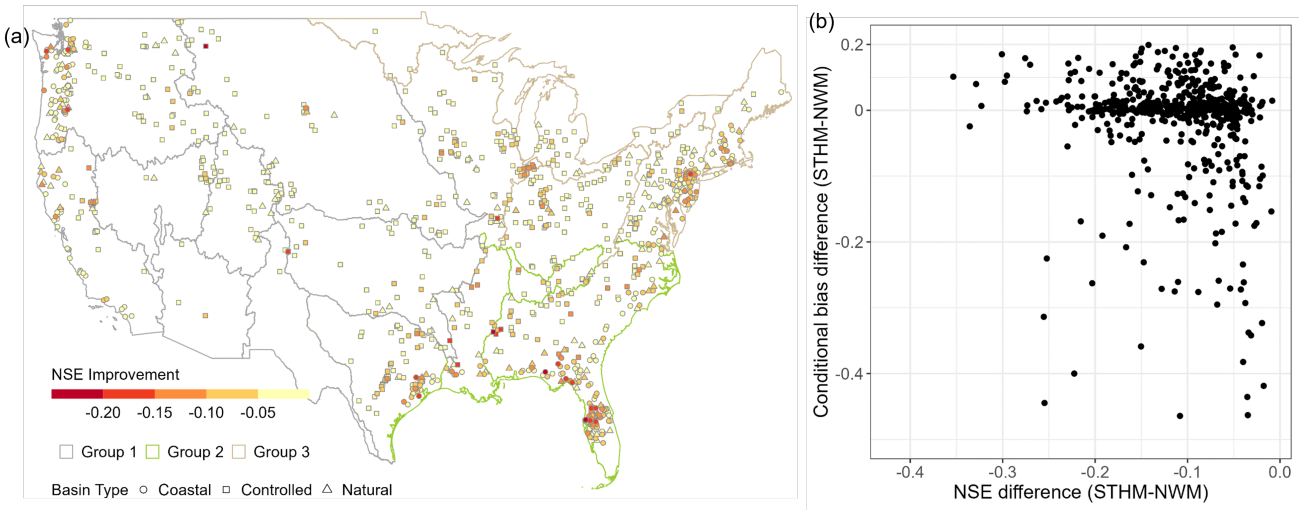


Figure S3. (a) Sites where the HM predicted daily high flows performed worse (in terms of NSE) than NWM high. (b) Unconditional bias difference between STHM and NWM relationship with NSE difference between STHM and NWM for the corresponding sites in figure (a).