General Response for All Reviewers:

We thank all three Referees for their reviews, which have been addressed to improve the paper. Though we provide a point-by-point response to each reviewer, we also have put together this General Response, articulating the main changes that were made to the manuscript, which should be read first. We focus on main changes in four of the sections, including: (i) Introduction, (ii) Evaluation Approach, (iii) Discussion, and (iv) Results. We refer to this in our point-by-point responses as needed. Further, although we have not finalized our manuscript revision, we have added relevant drafted excerpts here, followed by the updated Tables and Figures, to give the Referees a better sense how these changes were integrated.

(i) **Introduction**: *Added citations and revised to better articulate our contribution*. We thank the reviewers for their suggestions from the literature to provide more context for our work. In addition, we have edited the final paragraph to better highlight our contribution. We provide the relevant revised draft excerpts from the last 3 paragraphs of the Introduction here:

Hydrologic catchment modelling has begun to move towards large-sample hydrology, an extension of comparative hydrology, where model performance is evaluated for a large sample of catchments, rather than focusing solely on individual watersheds. This is appealing since evaluating hydrologic models across a wide variety of hydrologic regimes facilitates more robust regional generalizations and comparisons (Gupta et al., 2014). As such, many hydrologic modelling evaluation efforts have begun to encompass larger spatial scales. As part of the North American Land Data Assimilation System project phase 2, Xia et al. 2012 evaluate simulated streamflow for four land surface models, focusing mostly on 961 small basins, as well as 8 major river basins in the contiguous US (CONUS), finding that the ensemble mean performs better than the individual models. Further, several large-sample datasets have been developed for community use. The Model Parameter Estimation Experiment (MOPEX) includes hydrometeorological time series and land surface attributes for hydrological basins in the US and globally that have minimal human impacts (Duan et al. 2006). The more recent CAMELS dataset (Catchment Attributes and Meteorology for Large-sample Studies) includes hydrometeorological data and catchment attributes for 600+ small- to medium-sized basins in the contiguous US (CONUS) (Addor et al. 2017). By using CAMELS basins that are minimally disturbed by human activities, Newman et al. (2015, 2017) and Addor et al. (2018) are able to attribute regional variations in model performance to continental-scale factors. Knoben et al. (2020) also use CAMELS with 36 lumped conceptual models, finding that model performance is more strongly linked to streamflow signatures than to climate or catchment characteristics.

While these efforts are useful towards evaluating smaller, minimally-impacted basins, there is also a need to benchmark model performance for larger basins, including those impacted by human activities. On the global scale, catchment techniques have been applied to global hydrologic modelling, and have been shown to outperform traditional gridded global models of river flow (Arheimer et al. 2020). On the regional scale, Lane et al. (2019) benchmark the predictive capability of river flow for over 1,000 catchments in Great Britain by using four lumped hydrological models; Lane et al. (2019) include both natural and human-impacted catchments, finding poor performance when the water budget is not closed, such as due to non-modeled human impacts. Mai et al. (2022) conducted a systematic intercomparison study over the Great Lakes Region, finding that regionally calibrated models suffer from poor performance in urban, managed, and agricultural areas. Tijerina et al. (2021) compared performance of two high-resolution models that incorporate lateral subsurface flow at 2,200 streamflow gages; they found poor performance in the Central US, potentially due to non-modeled groundwater abstraction and irrigation. As hydrologic model development moves to include human systems, these studies provide important baselines.

This study builds on previous large-sample studies by benchmarking long-term retrospective streamflow simulations over the CONUS. Specifically, we evaluate two high-resolution, process-oriented models that have been developed to address water issues nationally: the National Water Model v2.1 application of WRF-Hydro (NWM v2.1; Gochis et al., 2020a) and the National Hydrologic Model v1 application of the Precipitation-Runoff Modeling System

(NHM v1; Regan et al., 2018). The evaluation is performed on daily streamflow for 5,390 streamflow gages from 1983-2016 (~33 years), including both natural and human-impacted catchments, representing one of the most comprehensive evaluations over the CONUS to date. The model performance is compared against a climatological benchmark that accounts for seasonality, and results are examined in terms of spatial patterns and human influences. The climatological seasonal benchmark is used as a threshold to screen the sites for each model application, offering a way to target the results for model diagnostics and development.

(ii) **Evaluation Approach:** *Reduced number of evaluation metrics and focus on KGE, added climatological benchmarks, used climatological benchmark as threshold to screen results for more targeted analysis.*
Based on the Referee suggestions, we focus on KGE, removing the other efficiency metrics (NSE and logNSE), and include the components of KGE (we replaced Spearman's rank with linear r), and focused on only two of the hydrologic signatures (we removed PBIAS_FDC). See Table 1 for updated metrics. We also provide performance context by using climatological benchmarks outlined in Knoben et al. (2019) and Knoben et al. (2020). As suggested by the Referees, we also provide context for the interpretation of the KGE scores; we now note that a lower benchmark must be specified (Pappenberger et al., 2015; Schaefli and Gupta, 2007; Seibert, 2001; Seibert et al., 2018). Further, as pointed out by the Referees, the KGE does not include a built-in lower benchmark in its formulation, but Knoben et al. (2019) show that models with KGE scores higher than −0.41 contribute more information than the mean flow benchmark. We also point to Knoben et al. (2020), who show that it is more robust to define a lower benchmark that considers seasonality. Hence, a reference time series based on the average and median flows for each day-of-year is used to calculate a lower KGE value which serves as a climatological (lower) benchmark; these are referred to as AvgDOY and MedDOY, respectively. We used this threshold to further target out analysis (see Results, revised below).

(iii) **Discussion**: *Revised discussion to address several points raised by Referees*.
The Referees raised several useful points, many of which we address in the Discussion now; these are pointed out in the point-by-point responses. But in general, we tightened up our discussion of using the benchmark to determine what a "good" model is, and discussed some of the updated results and what they might mean for model diagnostics and development.

(iv) **Results:** *Replaced tables with cumulative density functions, anchor KGE results based on climatological benchmarks, use climatological KGE benchmark to focus on underperforming sites*.
As indicated above, the Referees offered several suggestions that helped to re-shape the manuscript and its results. We provide the updated Results section here, followed by the draft Figures and Tables to show the Referees how these changes manifested in the updated manuscript.

4 Results

Using daily observations and model simulations, the evaluation metrics from Table 1 are calculated for each of the gages for the NWMv2.1 (Towler et al., 2022a) and NHMv1.0 (Towler et al., 2022b) hydrologic modelling applications. KGE is also calculated using daily observations and day-of-year averages (AvgDOY) and medians day-of-year (MedDOY) to produce a seasonal KGE benchmark for each site.

KGE scores for the benchmarks and models can be seen as a cumulative density functions (CDFs; Figure 2), and Table 2 quantifies the percent of sites less than or greater than select KGE scores. First, the seasonal benchmarks and model KGE scores can be compared to the mean flow benchmark (i.e., KGE <-0.41; Knoben et al. 2019): for the MedDOY benchmark, 18% of sites have lower scores, and using the AvgDOY benchmark is always better than using the mean flow. For the models, at 14% of the sites the NWMv2.1 simulations do not provide more information than the mean flow benchmark, similar to 12% of sites using NHMv1.0. The CDFs for the models intersect with the AvgDOY curve at a KGE score of about -0.06; at this value, 19%-20% of the sites perform worse in terms of KGE using the model simulation, whereas above this value the model simulations perform better than AvgDOY. In terms of median values, the AvgDOY (MedDOY) has a median KGE of 0.08 (-0.1), while the NWMv2.1 has a median of 0.53 and the NHMv1.0 median is 0.46. Given the better performance of AvgDOY in comparison to MedDOY, only AvgDOY is used for benchmarking the forthcoming analyses.

KGE performance is also examined by whether it has been classified as Reference or Non-Reference. Reference gages indicate less-disturbed watersheds, whereas observations associated with Non-Reference gages have some level of anthropogenic influence (Falcone, 2011). Figure 3 shows KGE scores as CDFs for the models and the AvgDOY benchmark broken out by this classification. As expected, the AvgDOY curves are virtually identical regardless of classification. However, for both models, the Reference gages are outperforming the Non-Reference gages. Table 3 shows the median values for the models: for the NHMv1.0, the KGE is 0.67 (0.38) for the Reference (Non-Reference), and for NWMv2.1 it is 0.65 for the Reference versus 0.49 for the Non-Reference. Looking at the components, the r values are the same for both model Reference sites (0.78). For the PBIAS, the NHMv1.0 shows underestimation for both Reference and Non-Reference sites (-4.1% and -5.7%, respectively), but the NWMv2.1 underestimates (-4.0%) at the Reference sites and overestimates (5.3%) at the Non-Reference sites.

Figure 4 shows KGE scores as CDFs for the models broken out by region. Here it can be seen that the model applications are fairly similar, but that there are notable differences by region. In general, performance is best for the Northeast, followed by the Southeast. Central and West perform the worst, although West exhibits some high KGE values. Table 4 shows the median KGE, r, rSD, and PBIAS values broken out by region, showing the biggest differences coming from PBIAS. Regional variability can be further examined by the KGE maps for the models: in the West, more of the poor performing sites are in the arid Southwest and the lower elevation basins in the intermountain West; better performance is seen in the higher elevations in the intermountain West and West Coast, including the Pacific Northwest (Figure 5A for NWMv2.1 and Figure 5B for NHMv1.0). Figure 5 shows that for both models in the Central region, relatively poor performance is concentrated along the plains areas that span from the high plains (i.e., North Dakota) vertically down through the center of the CONUS (i.e., South Dakota, Nebraska, Kansas, Texas). Performance is more mixed as one moves further east in the Central region (e.g., around the Great Lakes). Relatively uniform good performance is seen in the Southeast. However, as previously mentioned, the model results need to be placed into context by comparing with a climatological benchmark. Figure 6 shows the KGE map for the AvgDOY, which has relatively higher KGE values mostly in parts of the western CONUS, where there are notable seasonal signatures (e.g., snowmelt runoff, etc.), and relatively lower KGE values in the most other regions. By taking KGE differences by site, it is easier to examine where the model applications are doing relatively better and worse than the seasonal benchmark. Figure 7 shows the spatial distribution of the KGE differences, where the model with the maximum KGE value is used (i.e., maximum between the $KGE_{NWMv2.1}$ and $KGE_{NHMv1.0}$). Overall, the model applications tend to outperform the AvgDOY benchmark, except in the West & western Central regions. Supplemental Figure 1 shows that if the AvgDOY benchmark is outperformed, it is usually by both models (at 63% of sites); this is similar to the findings of Knoben et al. (2020). KGE difference maps for each individual model can be seen in Supplemental Figures 2 and 3, but follow the same general spatial pattern.

Basins that do not exceed the climatological benchmark are further scrutinized for each model application to offer insights toward model diagnostics and development; that is, only sites that have KGE scores worse than the AvgDOY benchmark are examined from here forward. In this section, these are called "underperforming sites". By

classification, most underperforming sites are human impacted (Non-Ref 90-93%, see Table 5). By region, most underperforming sites are in the West (55-67%) or Central (23-28%) regions (Table 6). Next, the bias metrics can be examined to try to get at why these sites are not able to beat the climatological benchmark. Spatial maps of PBIAS shows that the NWMv2.1 (Figure 8A) generally overestimates volume; NHMv1.0 (Figure 8B) is more mixed with underestimation in Central. Both models overestimate water volumes in the West. This could be because neither model is capturing active reservoir operations or water extractions (e.g., for irrigation), which is important since water is heavily managed in the West. This is different than the overall distribution of PBIAS for the modelling applications, where if you look at all the gages (n=5390), PBIAS for both models is centered around zero (Supplemental Figure 4). Another interesting feature of the PBIAS maps is the area of underestimation in Central for the NHMv1.0, which is absent in NWMv2.1. This could be due to the different time steps of the models, where NWMv2.1 is run hourly and NHMv1.0 is run daily; this hypothesis is expanded upon in the Discussion section. Maps for PBIAS_HF show a similar pattern (Supplemental Figure 5). However, for PBIAS_HF, the overall distribution of PBIAS_HFs is centered below zero, indicating that the models tend to underestimate high flows, but for the underperforming gages this is more pronounced in the NHMv1.0 than then NWMv2.1 (Supplemental Figure 6). Results for rSD paint a similar picture: both models tend to underestimate variability, but the under-estimation is more pronounced in NHMv1.0 (Supplemental Figures 8 and 9). Figure 9 shows PBIAS_LF for both model applications: the NWMv2.1 tends to overestimate the low flows, whereas the NHMv1.0 is more mixed and the over- or under-estimation is less severe. This can also be seen in the histograms for PBIAS_LF (Supplemental Figure 7).

Tables

Table 1. Standard metric suite included in the daily streamflow evaluation. KGE = Kling–Gupta efficiency; Pearson's r = linear correlation; rSD = ratio of standard deviations between simulations and observed; PBIAS = percent bias; HF = high flows; LF = low flows.

Statistic	Description	Range (Perfect)	Comments
KGE	Kling–Gupta efficiency (Gupta et al., 2009)	-Inf to 1 (1)	Normalized hydrologic metric of overall performance geared towards high flows (sensitive to outliers); calculated from KGE in R package hydroGOF.
Pearson's r	Pearson's correlation coefficient	-1 to 1 (1)	Pearson (linear estimator) of correlation; calculated from rPearson in R Package hydroGOF.
rSD	Ratio of standard deviations	0 to Inf (1)	Indicates if flow variability is being over- or under-estimated; calculated from rSD in R Package hydroGOF.
PBIAS	Percent bias	-100 to Inf (0)	Indicates if total streamflow volume is being over- or under-estimated; calculated from pbias in R Package hydroGOF.
PBIAS_HF	Percent bias of flows >=Q98 (Yilmaz et al. 2008)	-100 to Inf (0)	Characterizes response to large precipitation events; calculated using flows >= the 98th percentile flow using pbias in R Package hydroGOF.
PBIAS_LF	Percent bias of flows <=Q30 (Yilmaz et al. 2008)	-Inf to 100 (0)	Characterizes baseflow; calculated following equations in Yilmaz et al. (2008) using logged flows <= the 30th percentile (zeros are set to USGS observational threshold of 0.01 cfs).

Table 2. Median Kling-Gupta efficiency (KGE) scores and percent of sites (p) less than or greater than given KGE scores for seasonal benchmarks based on the median day-of-year flows (MedDOY) and average day-of-year flows (AvgDOY), and the models: National Water Model v2.1 (NWMv2.1) and National Hydrologic Model v1.0 (NHMv1.0).

	Median KGE	p(KGE<-0.41)	p(KGE<-0.06)	p(KGE>0.50)	p(KGE>0.75)
MedDOY	-0.13	18%	59%	5.7%	0.2%
AvgDOY	0.08	0%	19%	8.4%	1.5%
NHMv1.0	0.46	12%	20%	46%	15%
NWMv2.1	0.53	14%	19%	54%	16%

Table 3. Median values broken out by Reference (Ref, n= 1,115) and Non-Reference (Non-ref, n= 4,274) gages (one gage was not designated as Ref or Non-ref and is therefore not included). KGE = Kling–Gupta efficiency; r = corelation coefficient, rSD = ratio of standard deviations between simulations and observed; PBIAS = percent bias; NHMv1.0=National Hydrologic Model v1.0; NWMv2.1 = National Water Model v2.1.

		KGE	r	rSD	PBIAS
NHMv1.0	Non-ref	0.38	0.72	0.86	-5.7
	Ref	0.67	0.78	0.84	-4.1
NWMv2.1	Non-ref	0.49	0.75	0.92	5.3
	Ref	0.65	0.78	0.87	-4.0

Table 4. Median values for each region. KGE = Kling–Gupta efficiency; r = correlation coefficient, rSD = ratio of standard deviations between simulations and observed; PBIAS = percent bias; NHMv1.0=National Hydrologic Model v1.0; NWMv2.1 = National Water Model v2.1.

	Model	KGE	r	rSD	PBIAS
West	NHMv1.0	0.29	0.74	0.98	9.3
	NWMv2.1	0.32	0.75	1.17	27
Central	NHMv1.0	0.33	0.68	0.78	-18
	NWMv2.1	0.45	0.71	0.87	4.4
Southeast	NHMv1.0	0.48	0.73	0.78	-11
	NWMv2.1	0.56	0.77	0.85	-1.1
Northeast	NHMv1.0	0.63	0.78	0.86	-3.0
	NWMv2.1	0.65	0.79	0.82	-7.8

Table 5. The number (percent) of sites in each classification for each hydrologic model application where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark (underperforming sites); KGE = Kling–Gupta efficiency; NHMv1.0=National Hydrologic Model v1.0; NWMv2.1 = National Water Model v2.1; max(Model) = model with maximum KGE value from NHMv1.0 or NWMv2.1; Ref = Reference (minimal human impacts); Non-Ref = Non-Reference (influenced by human activities).

Model	Class	
NHMv1.0	Ref	137 (9.4%)
	Non-Ref	1319 (91%)
NWMv2.1	Ref	136 (9.5%)
	Non-Ref	1302 (90%)
max(Model)	Ref	60 (7%)
	Non-Ref	850 (93%)

Table 6. The number (percent) of sites in each region for each hydrologic model application where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark (underperforming sites); KGE = Kling–Gupta efficiency; NHMv1.0=National Hydrologic Model v1.0; NWMv2.1 = National Water Model v2.1; max(Model) = model with maximum KGE value from NHMv1.0 or NWMv2.1.

Model	West	Central	Southeast	Northeast
NHMv1.0	795 (55%)	412 (28%)	159 (11%)	91 (6%)
NWMv2.1	842 (59%)	370 (26%)	173 (12%)	54 (4%)
max(Model)	610 (67%)	213 (23%)	61 (7%)	27 (3%)

Figures

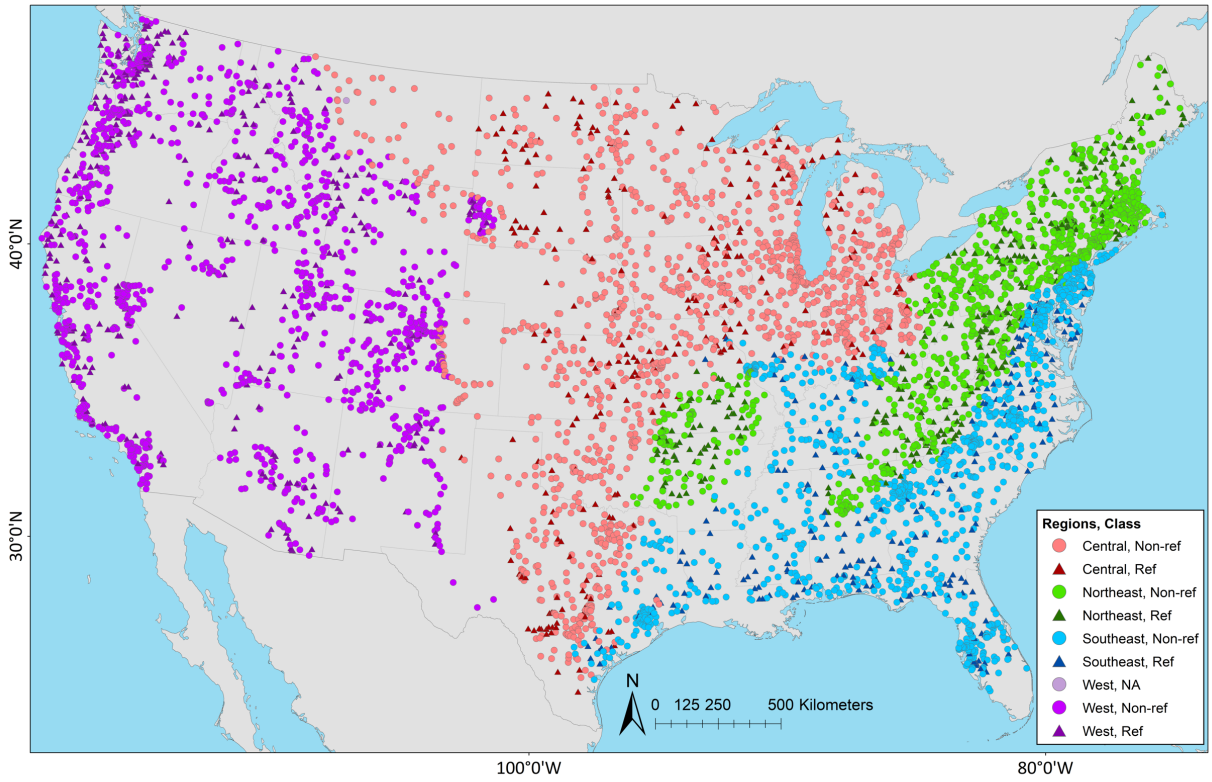


Figure 1: Site locations used in evaluation (n=5,390), including regions and classification. Regions were further combinations of aggerated ecoregions defined by Falcone (2010): Central (n=1,450) includes Central Plains, Western Plains, and Mixed Wood Shield; Northeast (n=1,218) includes Northeast and Eastern Highlands; Southeast (n=1,212) includes South East Plains and South East Coastal Plains; and West (n=1,510) includes Western Mountains and West Xeric. Classifications are from Falcone (2010): Reference (Ref, n= 1,115) and Non-Reference (Non-ref, n= 4,274); one gage was not designated (NA, n=1).

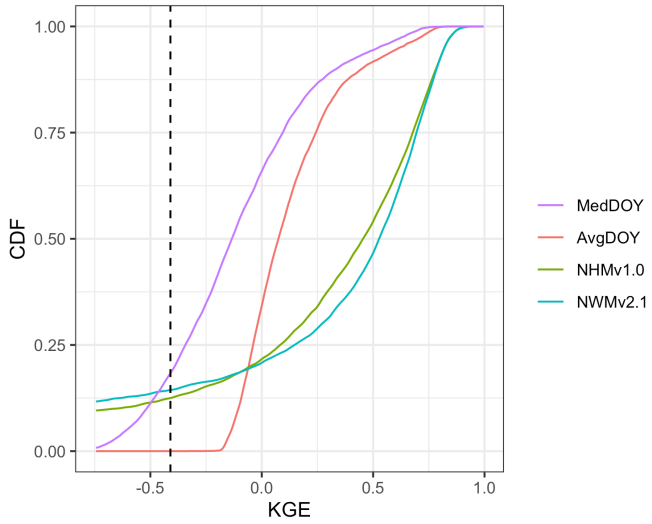


Figure 2: Cumulative density function (CDF) for Kling-Gupta efficiency (KGE) scores based on daily streamflow at U.S. Geological Survey (USGS) gages for seasonal benchmarks based on the median day-of-year flows (MedDOY) and average day-of-year flows (AvgDOY) and models: National Water Model v2.1 (NWMv2.1) and National Hydrologic Model v1.0 (NHMv1.0). Dotted vertical line is KGE mean flow benchmark (=-0.41). For sites (n=1 for NWMv2.1 and n=16 for NHMv1.0) for which a KGE could not be calculated (i.e., the modeled timeseries had all zero values for the entire timeseries), these are included as -Inf in the CDFs.

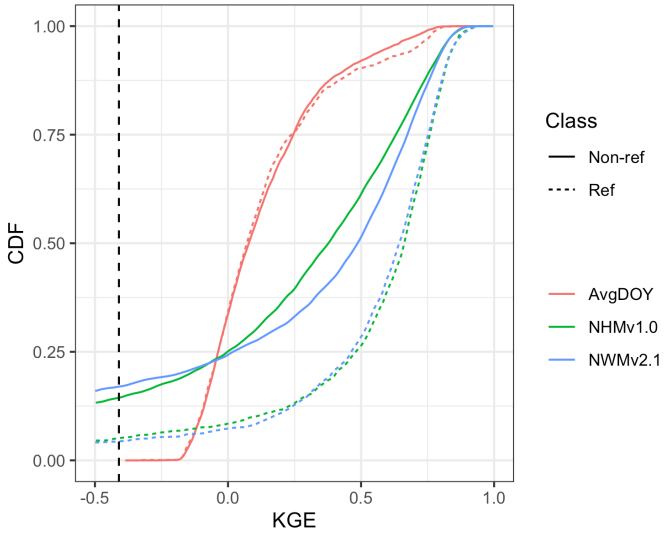


Figure 3: Cumulative density function (CDF) for Kling-Gupta efficiency (KGE) scores based on daily streamflow at U.S. Geological Survey (USGS) gages for seasonal benchmark based on average day-of-year flows (AvgDOY) and models: National Water Model v2.1 (NWMv2.1) and National Hydrologic Model v1.0 (NHMv1.0). Dotted vertical line is KGE mean flow benchmark (=-0.41). Reference (Ref, n= 1,115) and Non-Reference (Non-ref, n= 4,274) classifications are from Falcone (2010).

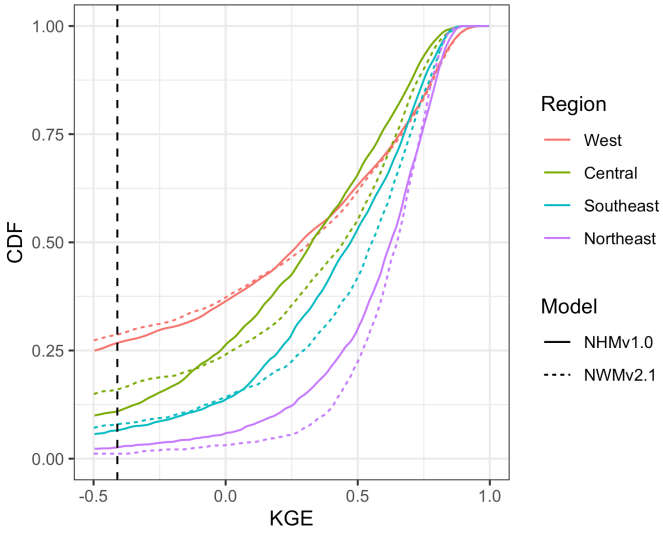


Figure 4: Cumulative density function (CDF) for Kling-Gupta efficiency (KGE) scores based on daily streamflow at U.S. Geological Survey (USGS) gages for models: National Water Model v2.1 (NWMv2.1) and National Hydrologic Model v1.0 (NHMv1.0). Dotted vertical line is KGE mean flow benchmark (=-0.41). Regions are further combinations of aggerated ecoregions defined by Falcone (2010): Central (n=1,450) includes Central Plains, Western Plains, and Mixed Wood Shield; Northeast (n=1,218) includes Northeast and Eastern Highlands; Southeast (n=1,212) includes South East Plains and South East Coastal Plains; and West (n=1,510) includes Western Mountains and West Xeric.

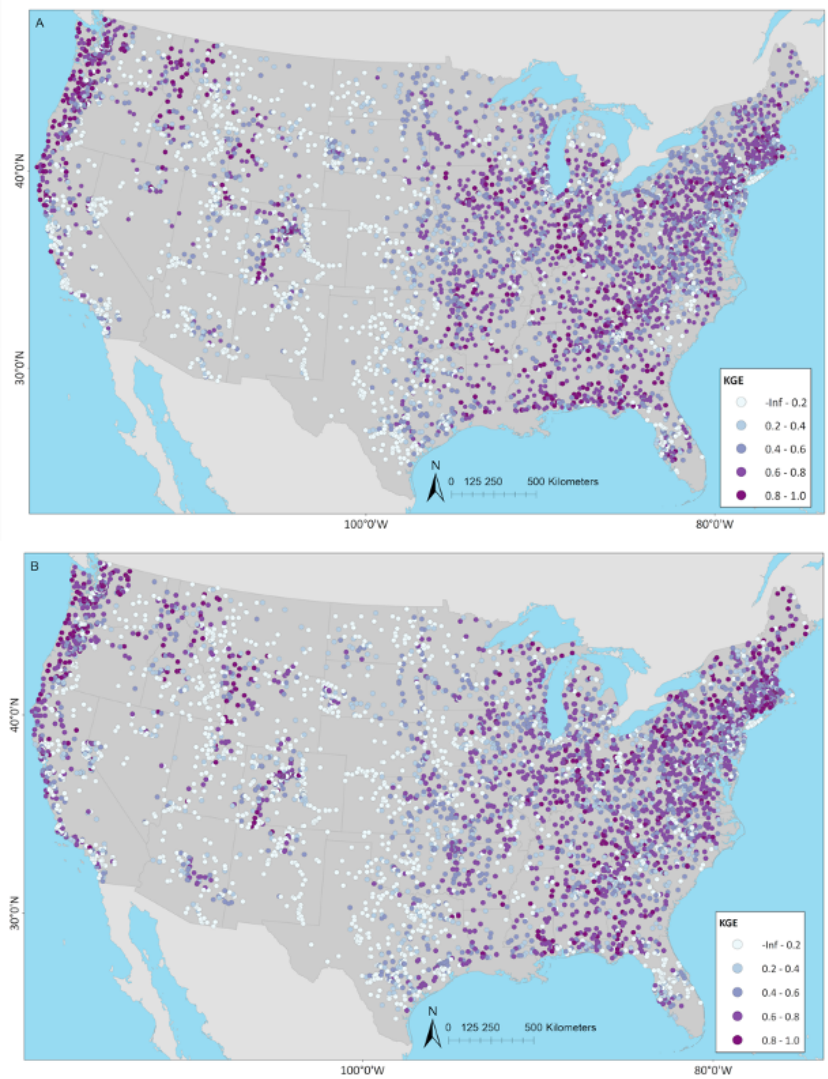


Figure 5: Kling–Gupta efficiency (KGE) based on daily streamflow at U.S. Geological Survey (USGS) gages for (A) National Water Model v2.1 (NWMv2.1) and (B) National Hydrologic Model v1.0 (NHMv1.0).

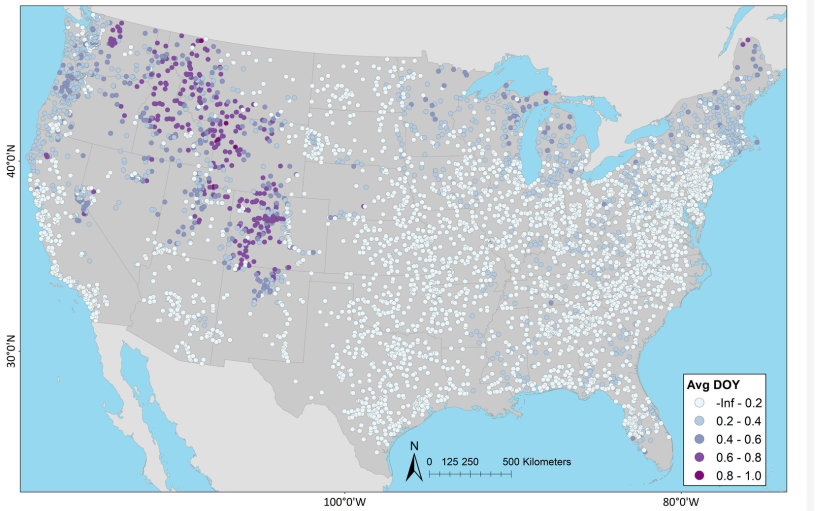


Figure 6: Kling–Gupta efficiency (KGE) based on daily streamflow at U.S. Geological Survey (USGS) gages for seasonal benchmark based on average day-of-year flows (AvgDOY).

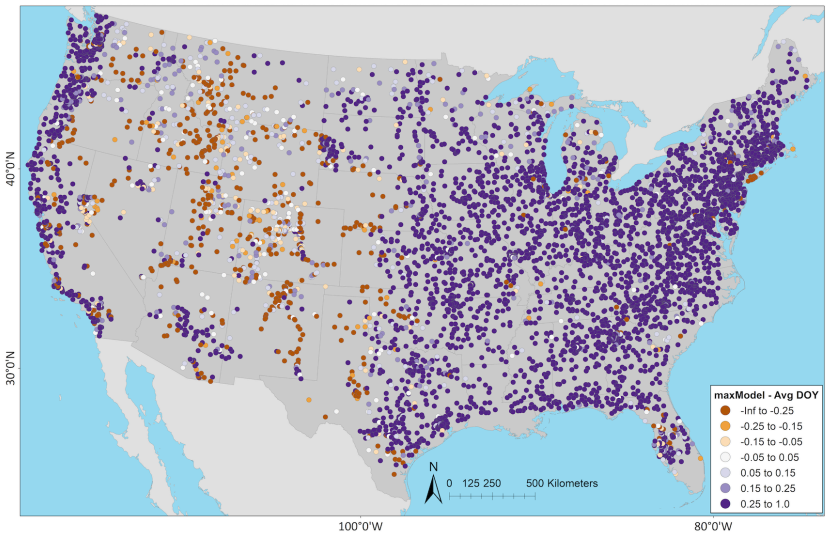


Figure 7: Difference between the Kling–Gupta efficiency (KGE) from the maximum model (maxModel) (i.e., the maximum KGE value from the National Water Model v2.1, NWMv2.1, or the National Hydrologic Model v1.0, NHMv1.0) minus the seasonal benchmark based on the average day-of-year flows (AvgDOY); negative (orange) indicates where AvgDOY has a higher (better) KGE, positive (purple) indicates that at least one of the models has a higher (better) KGE.

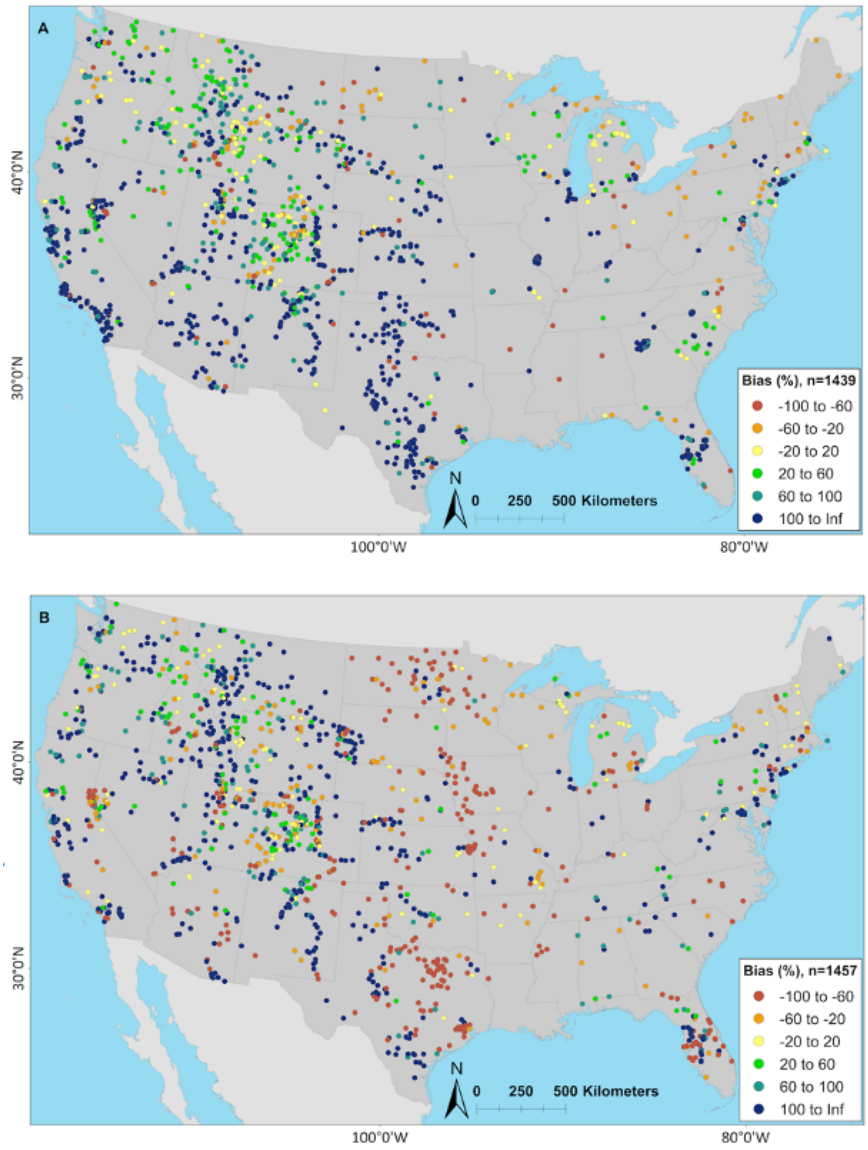


Figure 8: Percent bias (PBIAS) maps for National Water Model v2.1 (NWMv2.1) (A) and National Hydrologic Model v1.0 (NHMv1.0) (B), for sites where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark. Cooler colors are where model application is overestimating volume and warmer colors are where model is underestimating volume.

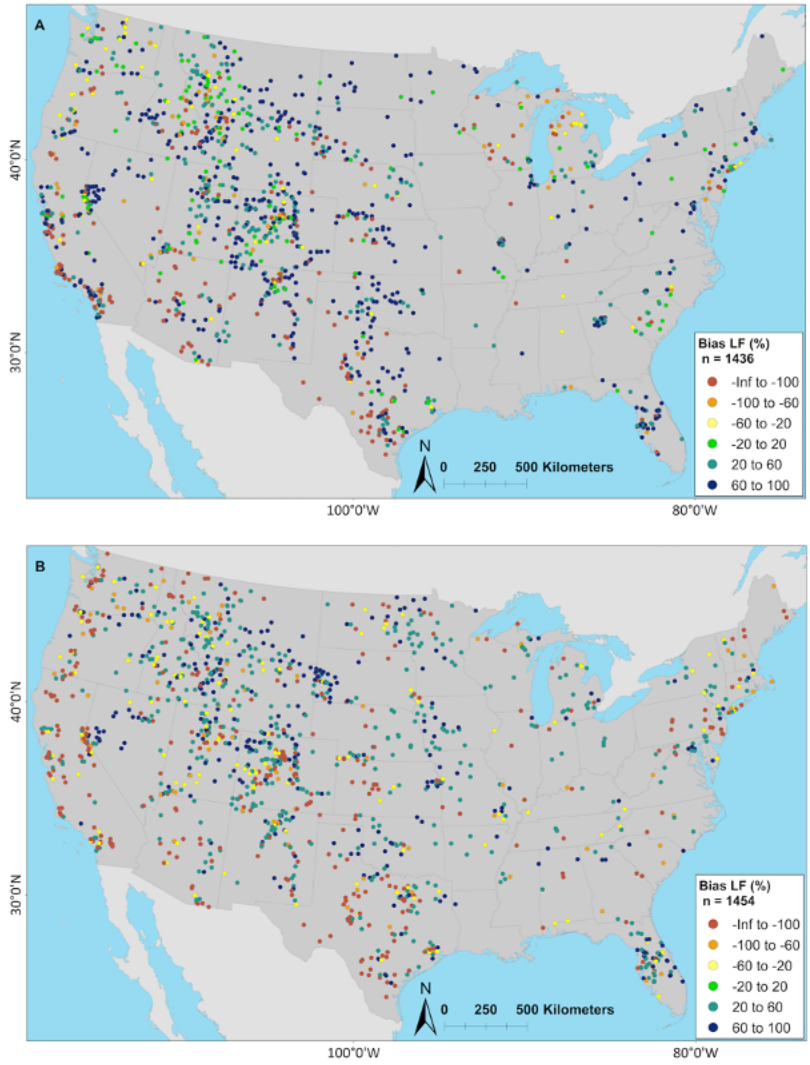
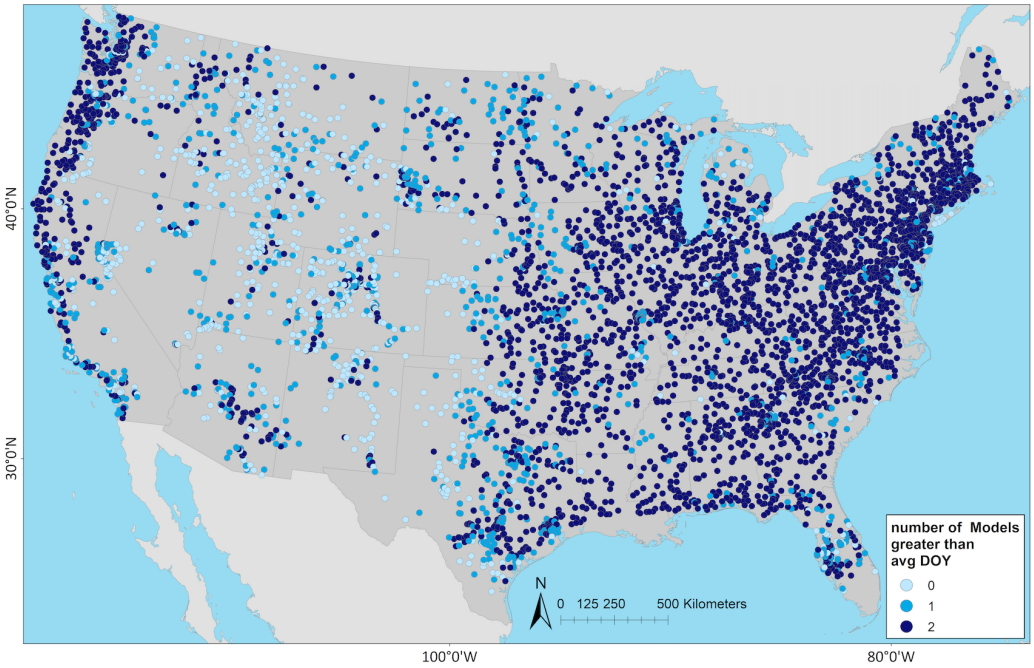


Figure 9: Percent bias low flow (PBIAS_LF, flows below 30% percentile) maps for National Water Model v2.1 (NWMv2.1) (A) and National Hydrologic Model v1.0 (NHMv1.0) (B), for sites where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark. Cooler colors are where model application is overestimating low flows and warmer colors are where model is underestimating low flows.

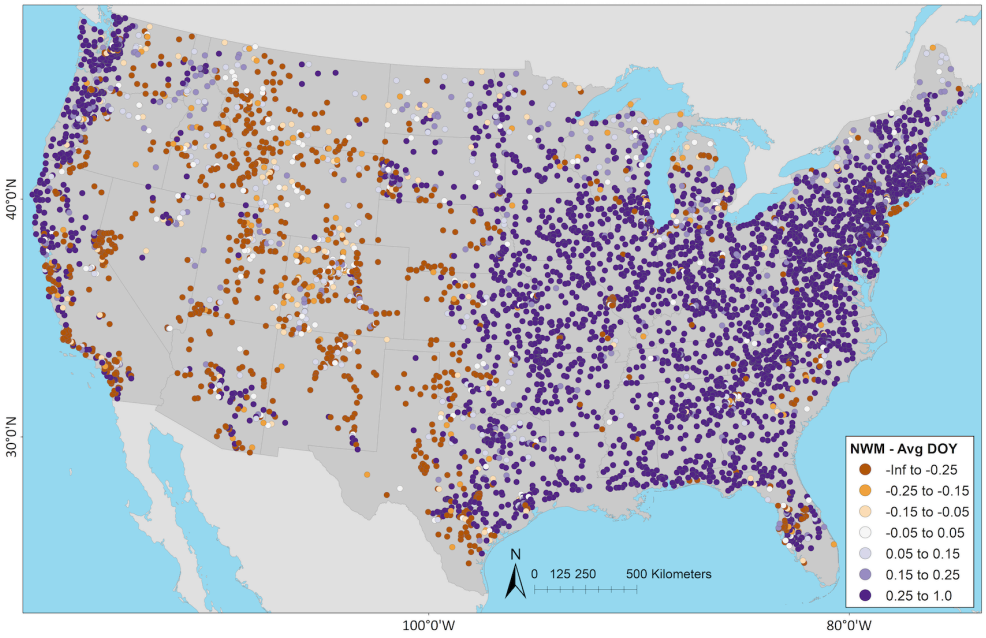
Supplemental Material for:
Benchmarking High-Resolution, Hydrologic Performance of Long-Term Retrospectives in the Contiguous United States

Erin Towler[1], Sydney S. Foks[2], Aubrey L. Dugger[1], Jesse E. Dickinson[3], Hedeff I. Essaid[4], David Gochis[1], Roland J. Viger[2], and Yongxin Zhang[1]

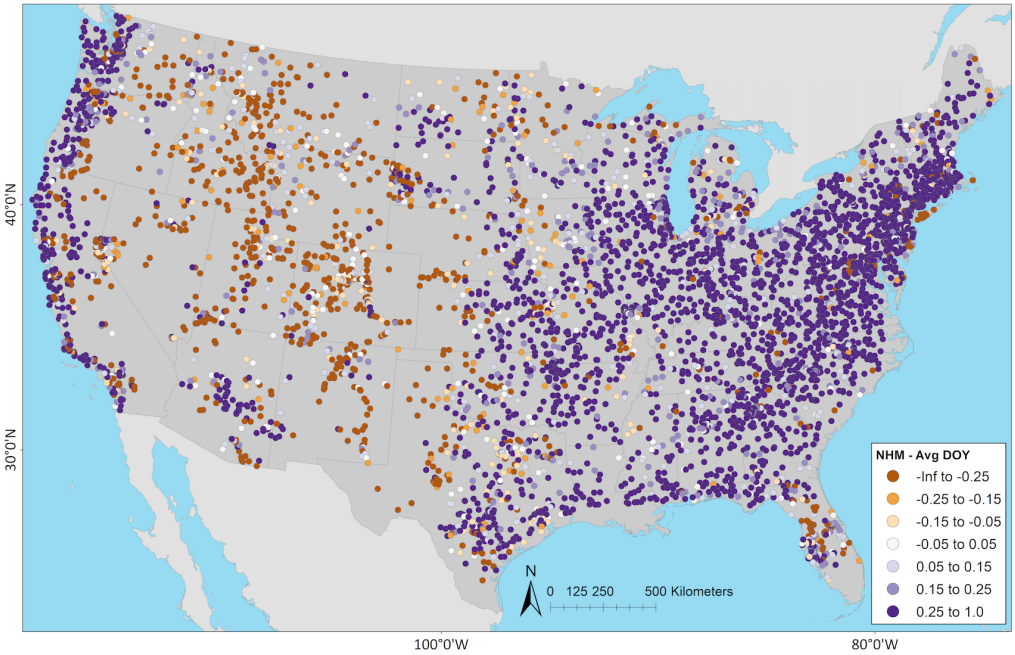
[1]National Center for Atmospheric Research (NCAR), Boulder, CO, USA
[2]U.S. Geological Survey (USGS), Lakewood, CO, USA
[3]U.S. Geological Survey, Arizona Water Science Center, Tucson, AZ, USA
[4]U.S. Geological Survey, Moffett Field, CA, USA



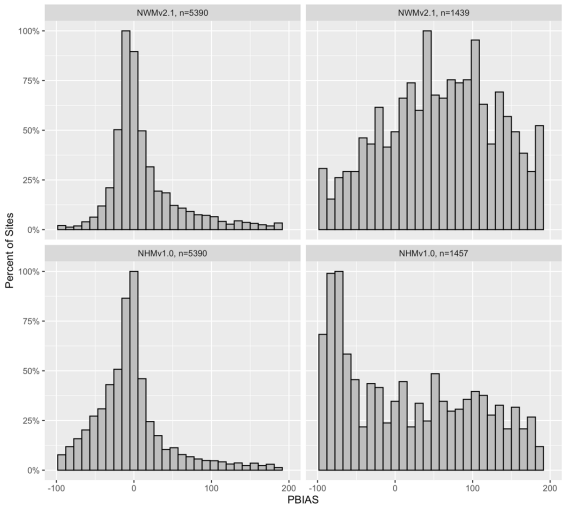
Supplemental Figure 1. For the National Water Model v2.1 (NWMv2.1) and the National Hydrologic Model v1.0 (NHMv1.0), the number of models where the KGE value is greater than the AvgDOY; both models are better (n=3396), one model is better (n = 1083), or neither model is better (n=911).



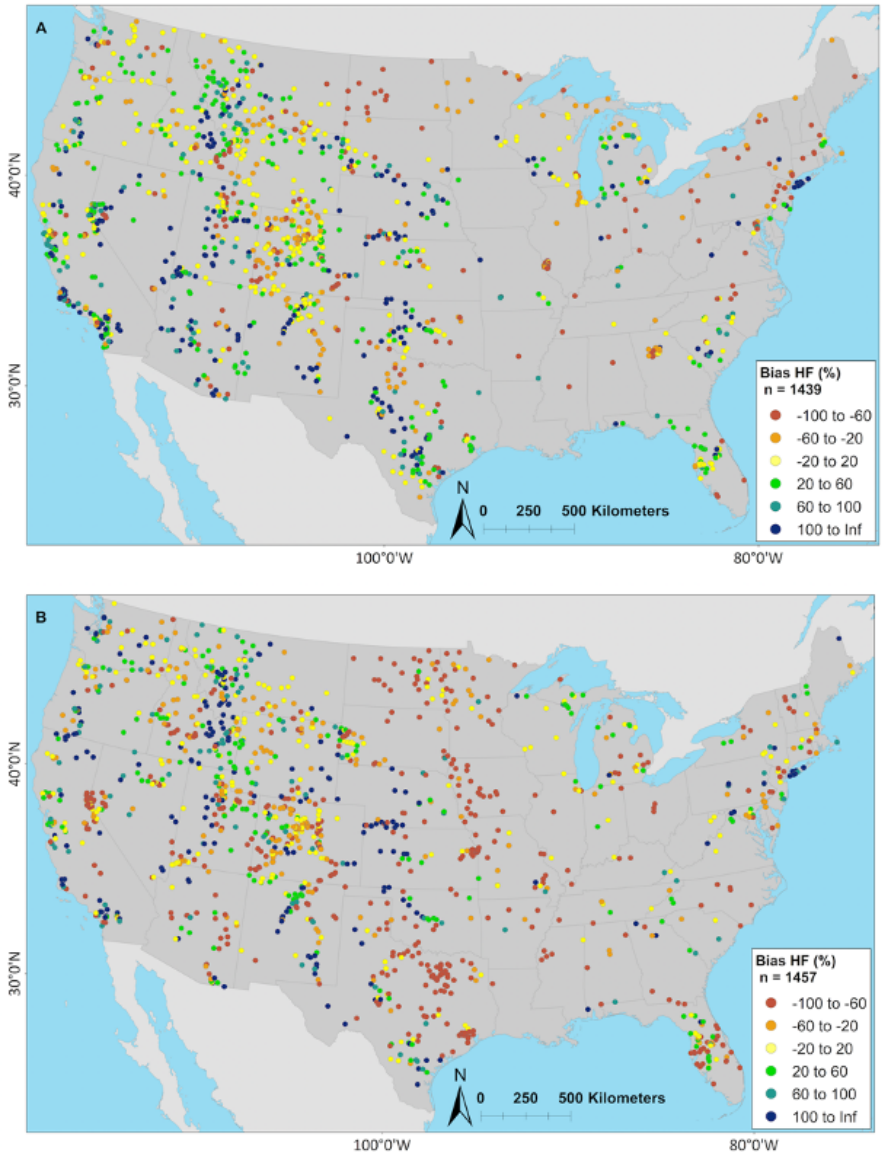
Supplemental Figure 2. Difference between the Kling–Gupta efficiency (KGE) from the National Water Model v2.1 (NWMv2.1) and the seasonal benchmark based on the average day-of-year flows (AvgDOY); negative (orange) indicates where AvgDOY has a higher (better) KGE, positive (purple) indicates that the NWMv2.1 has a higher (better) KGE.



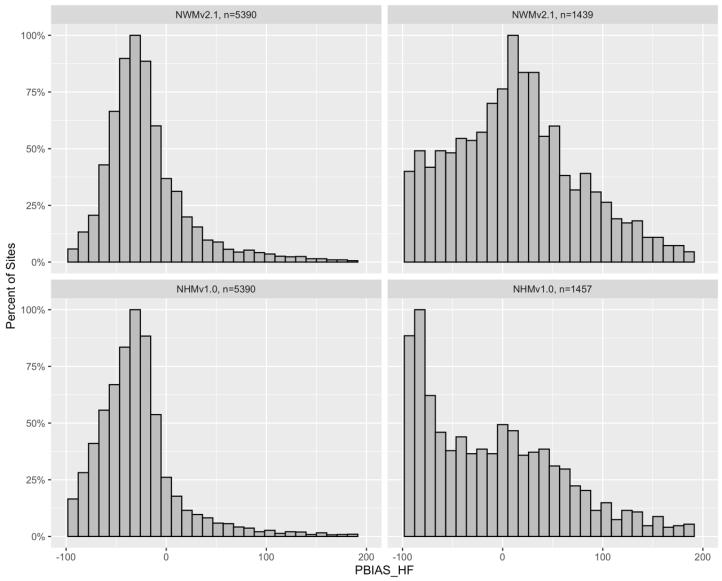
Supplemental Figure 3. Difference between the Kling–Gupta efficiency (KGE) from the National Hydrologic Model v1.0 (NHMv1.0) and the seasonal benchmark based on the average day-of-year flows (AvgDOY); negative (orange) indicates where AvgDOY has a higher (better) KGE, positive (purple) indicates that the NHMv1.0 has a higher (better) KGE.



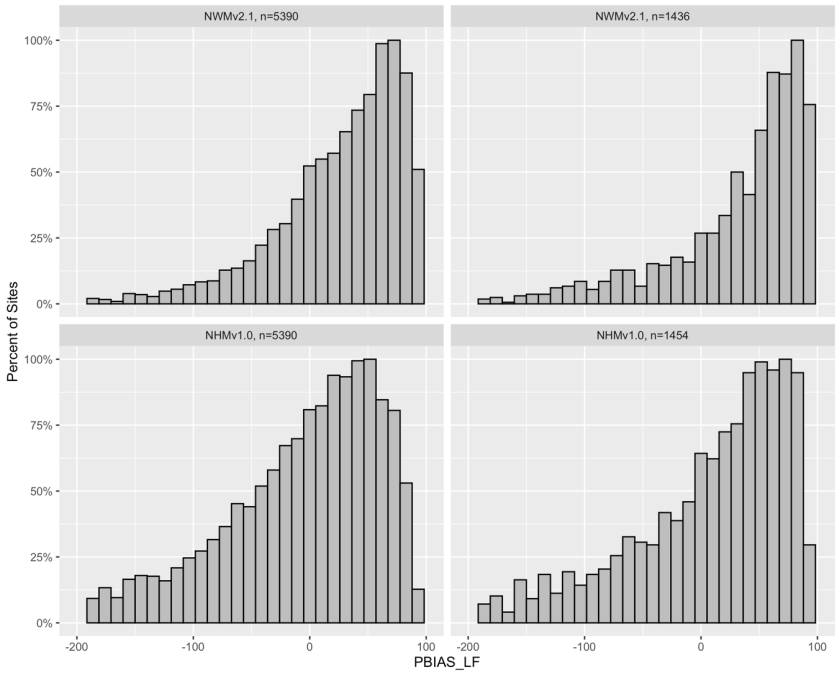
Supplemental Figure 4: Normalized histograms of PBIAS for National Water Model v2.1 (NWMv2.1, top) and National Hydrologic Model v1.0 (NHMv1.0, bottom), for all sites (left) and for sites where the model's KGE score is less than the average day-of-year flow benchmark (right).



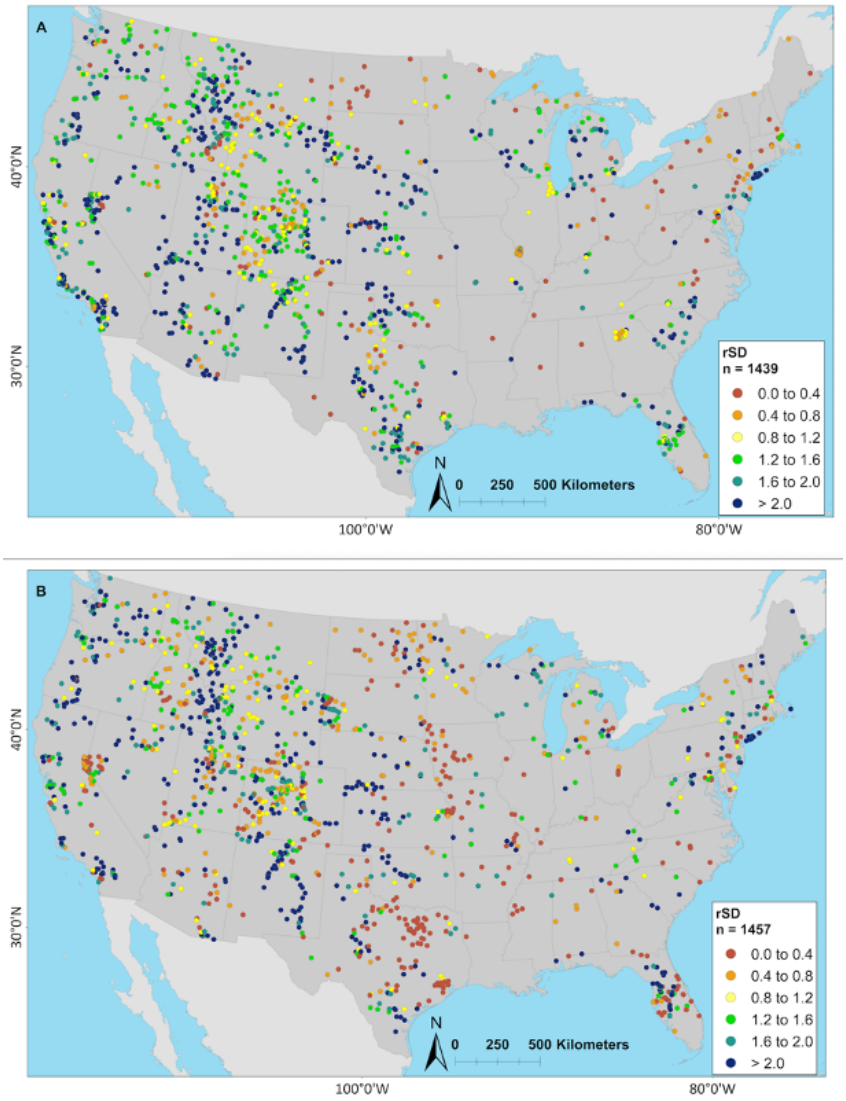
Supplemental Figure 5: Percent bias of high flow (PBIAS_HF; i.e., exceeding top 2%) maps for National Water Model v2.1 (NWMv2.1) (A) and National Hydrologic Model v1.0 (NHMv1.0) (B), for sites where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark. Cooler colors are where model application is overestimating high flow bias and warmer colors are where model is underestimating high flow bias.



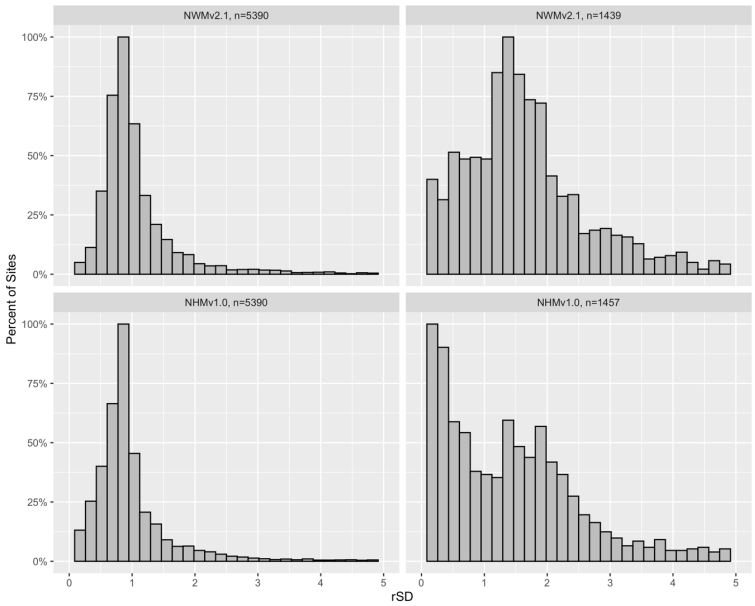
Supplemental Figure 6: Normalized histograms of Percent bias of high flow (PBIAS_HF; i.e., exceeding top 2%) for National Water Model v2.1 (NWMv2.1, top) and National Hydrologic Model v1.0 (NHMv1.0, bottom), for all sites (left) and for sites where the model's KGE score is less than the average day-of-year flow benchmark (right).



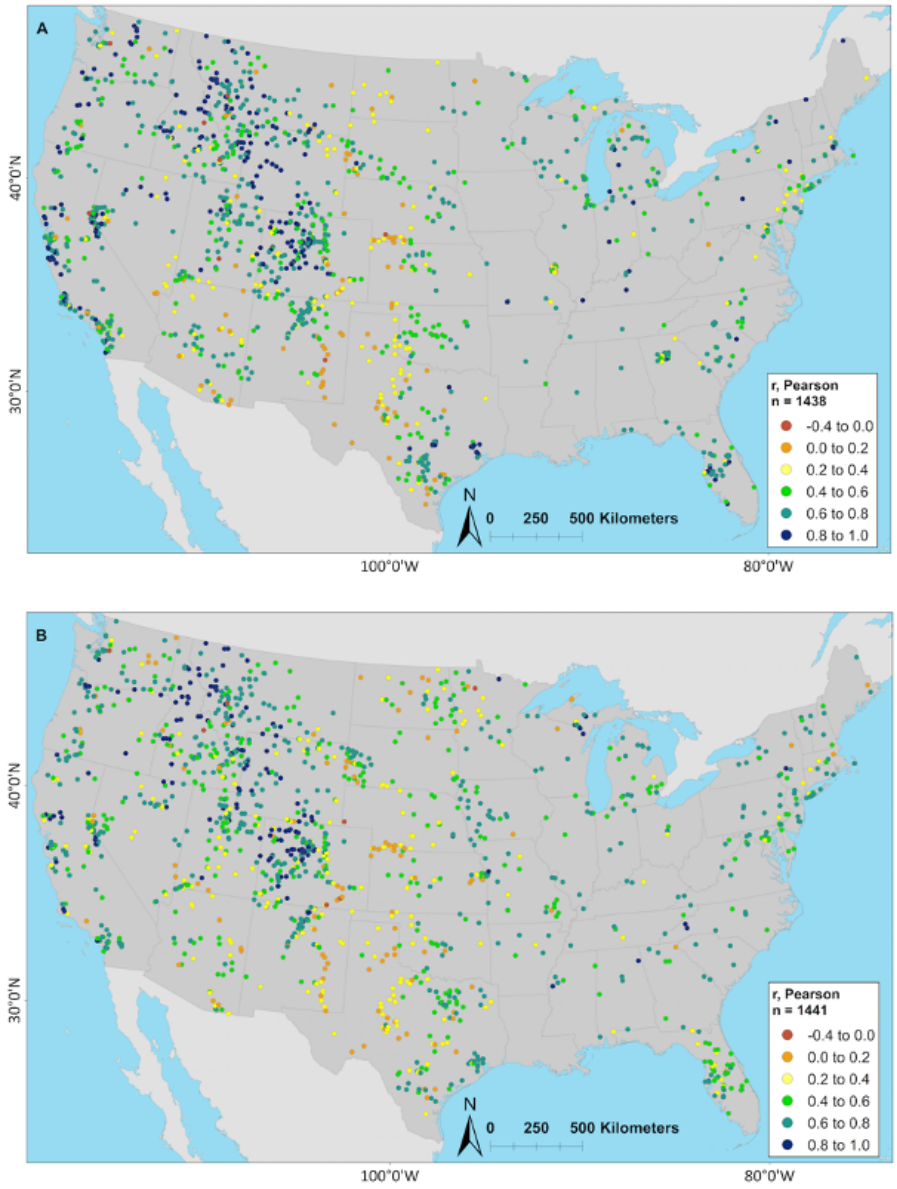
Supplemental Figure 7: Normalized histograms of percent bias of low flow (PBIAS_LF, flows below 30% percentile) for National Water Model v2.1 (NWMv2.1, top) and National Hydrologic Model v1.0 (NHMv1.0, bottom), for all sites (left) and for sites where the model's KGE score is less than the average day-of-year flow benchmark (right).



Supplemental Figure 8: ratio of standard deviation (rSD) maps for National Water Model v2.1 (NWMv2.1) (A) and National Hydrologic Model v1.0 (NHMv1.0) (B), for sites where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark. Cooler colors are where model application is overestimating variability and warmer colors are where model is underestimating variability.



Supplemental Figure 9: Normalized histograms of standard deviation ratio (rSD) for National Water Model v2.1 (NWMv2.1, top) and National Hydrologic Model v1.0 (NHMv1.0, bottom), for all sites (left) and for sites where the model's KGE score is less than the average day-of-year flow benchmark (right).



Supplemental Figure 10: Pearson's correlation coefficient (r) for National Water Model v2.1 (NWMv2.1) (A) and National Hydrologic Model v1.0 (NHMv1.0) (B), for sites where the KGE score is less than the average day-of-year flow (AvgDOY) benchmark.

Equations:

The percent bias in the high flows (PBIAS_HF) is defined as (Yilmaz et al. 2008):

$$PBIAS_{HF} = \frac{\sum_{h=1}^{H}(S_h - O_h)}{\sum_{h=1}^{h} O_h}$$

Where h = 1, 2,… H are the low flow indices for flows with exceedance probabilities lower than 0.02.

The percent bias in the low-flow (PBIAS_LF) is defined as (Yilmaz et al. 2008):

$$PBIAS_{LF} = -1 \cdot \frac{\sum_{l=1}^{L}[log(S_l) - log(S_L)] - \sum_{l=1}^{L}[log\,(O_l - log(O_L)]}{\sum_{l=1}^{L}[log(O_l) - log(O_L)]} \times 100$$

where l = 1, 2,…L is the flow value index in the low-flow segment (0.7–1.0 flow exceedance probabilities) of the flow duration curve and L is the minimum flow index.