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

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

[1] Department of Civil and Environmental Engineering, Virginia Tech, Blacksburg, Virginia, USA – ORCID: 0000-0002-4225-2252

[2] U.S. Geological Survey, Maryland-Delaware-D.C. Water Science Center, Catonsville, Maryland, USA – ORCID: 0000-0002-4935-0736

[3] USDA Forest Service, Pacific Northwest Research Station, La Grande, Oregon, USA – ORCID: 0000-0002-7211-4758

[4] Department of Civil, Environmental and Architectural Engineering, University of Kansas, Lawrence, Kansas, USA – ORCID: 0000-0003-0328-3248

*Admin Husic, husic@vt.edu, Occoquan Watershed Monitoring Laboratory, Virginia Tech, Manassas, VA 20110

**Table S1.** Predictor types, variable names, and BasinATLAS reference names (Linke et al., 2019).

| Variable Name (and type)		BasinATLAS Name
Climate	Precipitation	pre_mm_syr
	Potential ET.	pet_mm_syr
	Actual ET.	aet_mm_syr
	Aridity Index	ari_ix_sav
	Air Temp. Min.	tmp_dc_smn
	Air Temp. Max.	tmp_dc_smx
	Air Temp. Avg.	tmp_dc_syr
	Snow Cover Max.	snw_pc_smx
	Snow Cover Avg.	snw_pc_syr
	Permafrost Extent	prm_pc_sse
	Glacier Extent	gla_pc_sse
Hydrology	Natural Discharge Min.	dis_m3_pmn
	Natural Discharge Max.	dis_m3_pmx
	Natural Discharge Avg.	dis_m3_pyr
	Naturalized Runoff	run_mm_syr
	River Area	ria_ha_usu
	River Volume	riv_tc_usu
	Inundation Extent Min.	inu_pc_smn
	Inundation Extent Max.	inu_pc_smx
	Lake Area Percent	lka_pc_sse
	Lake Volume	lkv_mc_usu
	Reservoir Volume	rev_mc_usu
	Regulation by Dams	dor_pc_pva
Topography	Basin Area	area
	Elevation Avg.	ele_mt_sav
	Elevation Min.	ele_mt_smn
	Elevation Max.	ele_mt_smx
	Stream Gradient	sgr_dk_sav
	Basin Slope	slp_dg_sav
Soils & Geology	Clay Fraction	cly_pc_sav
	Silt Fraction	slt_pc_sav
	Sand Fraction	snd_pc_sav
	Soil Water Content	swc_pc_syr
	Soil Organic Carbon	soc_th_sav
	Groundwater Depth	gwt_cm_sav
	Karst Cover	kar_pc_sse
Natural	Forest Cover	for_pc_sse
	Wetlands	wet_pc_sg2
	Protected Area	pac_pc_sse
Agriculture	Cropland	crp_pc_sse
	Irrigated Area	ire_pc_sse
	Pasture	pst_pc_sse
	Soil Erosion	ero_kh_sav
Urban	Urban Extent	urb_pc_sse
	Road Density	rdd_mk_sav
	Population Count	pop_ct_usu
	Population Density	ppd_pk_sav
	Human Footprint	hft_ix_s09
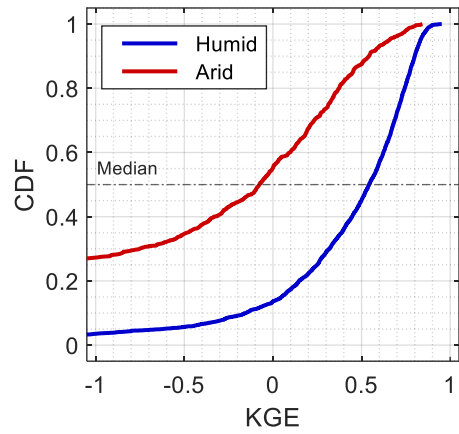
Figure S1. Cumulative distribution function (CDF) of National Hydrologic Model performance for humid (PET/P <1, n = 3,827) and arid (PET/P >1, n = 787) sites as assessed by the Kling–Gupta efficiency (KGE) evaluation metric.
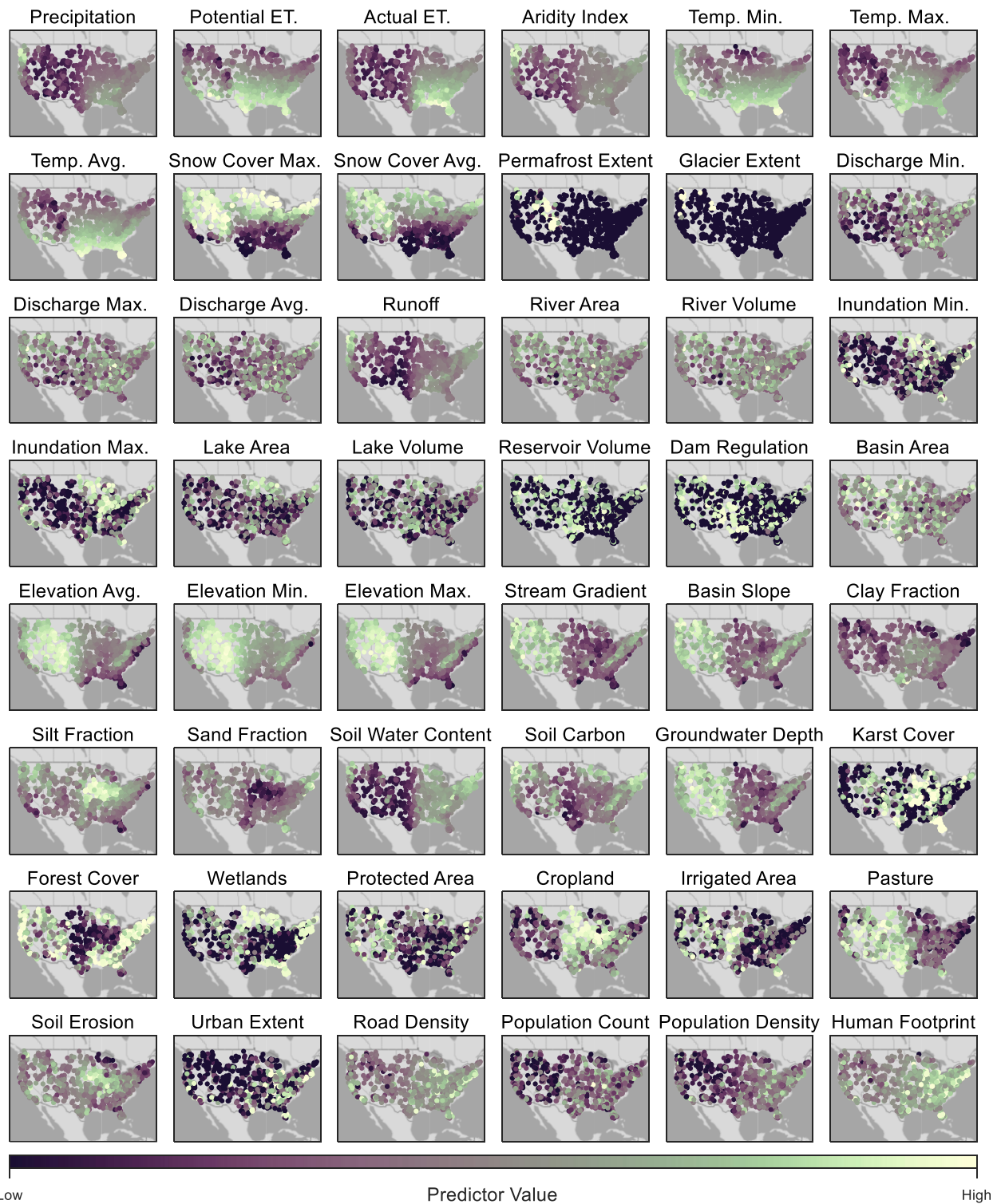
Figure S2. Catchment attributes plotted for the study basins. For a full description of variables see Table S1. The color scale for each predictor is based on a Box-Cox transformation of the data to highlight spatial gradients, not absolute values of features.
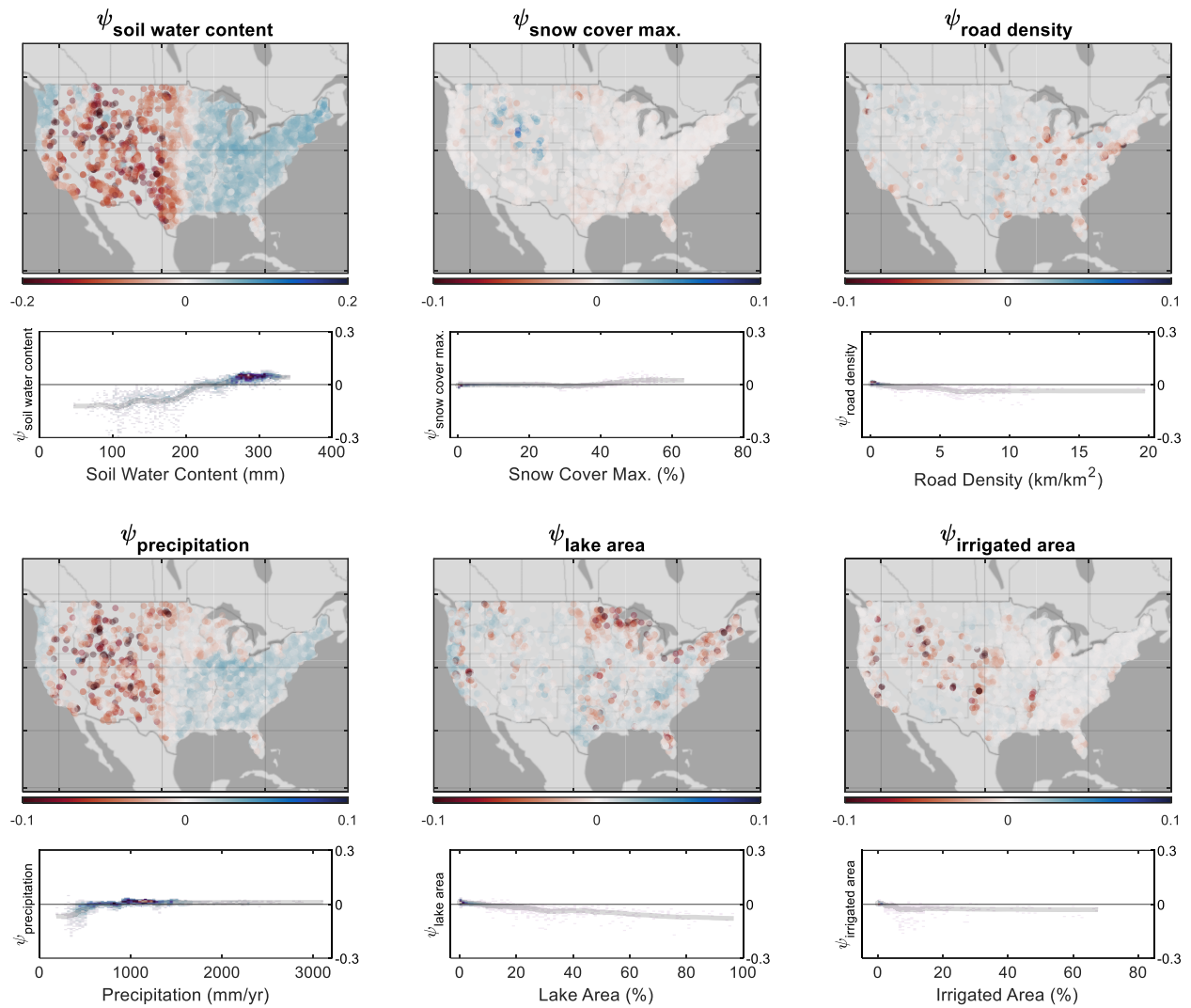
Figure S3. Spatial distribution of Shapley values (ψ) for selected influential features and their impact on Kling–Gupta efficiency (KGE) prediction for the National Hydrologic Model (NHM). The partial dependence plot of each feature is shown. Features value distributions are represented with a heatmap. A moving average of feature values is indicated by a line to show general trends.
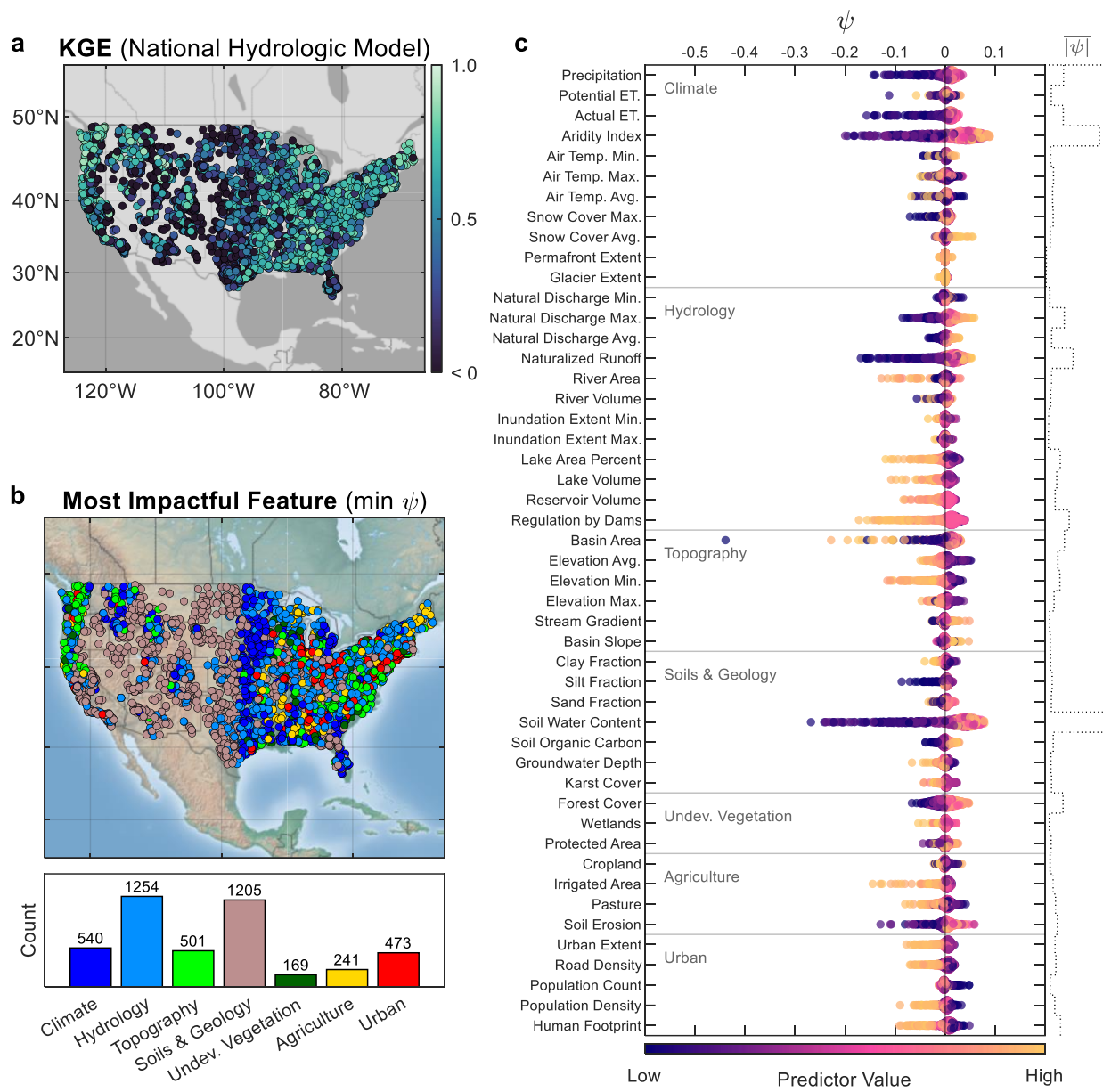
Figure S4. (a) Map of Kling–Gupta efficiency (KGE) for the National Hydrologic Model (NHM). (b) Map and histogram of the most impactful feature causing poor model performance at each site, i.e., the predictor group having the greatest negative Shapley value (ψ) at a site. (c) Swarm chart of Shapley values for KGE prediction showing feature importance for 48 predictors. The staircase plot on the right axis indicates the mean absolute Shapley value $\overline{|\psi|}$) of all observations for a predictor.
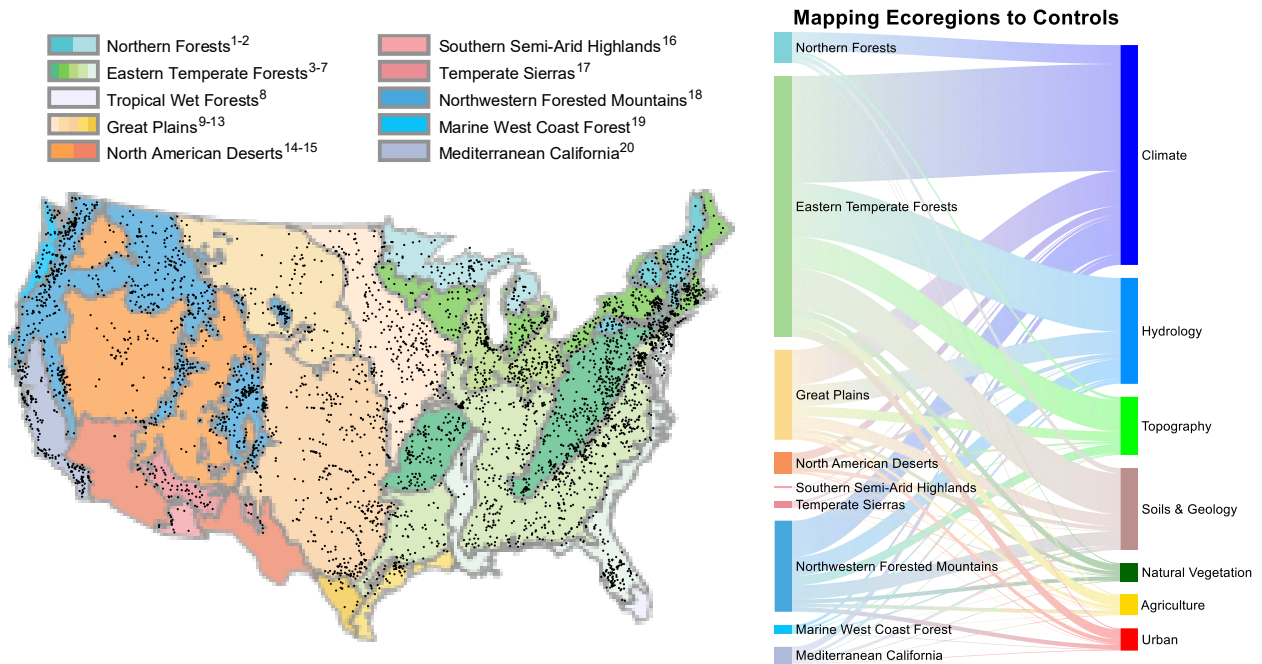
Figure S5. Map of study stream gages (black markers) and the Ecological Regions of North America (as defined in Omernik, 1987). Sankey diagram showing the pairing of ecoregions and impactful feature classes for the National Hydrologic Model (NHM) for the Kling–Gupta efficiency (KGE) evaluation metric. Superscripts in ecoregion classifications are defined in Section 2.3.