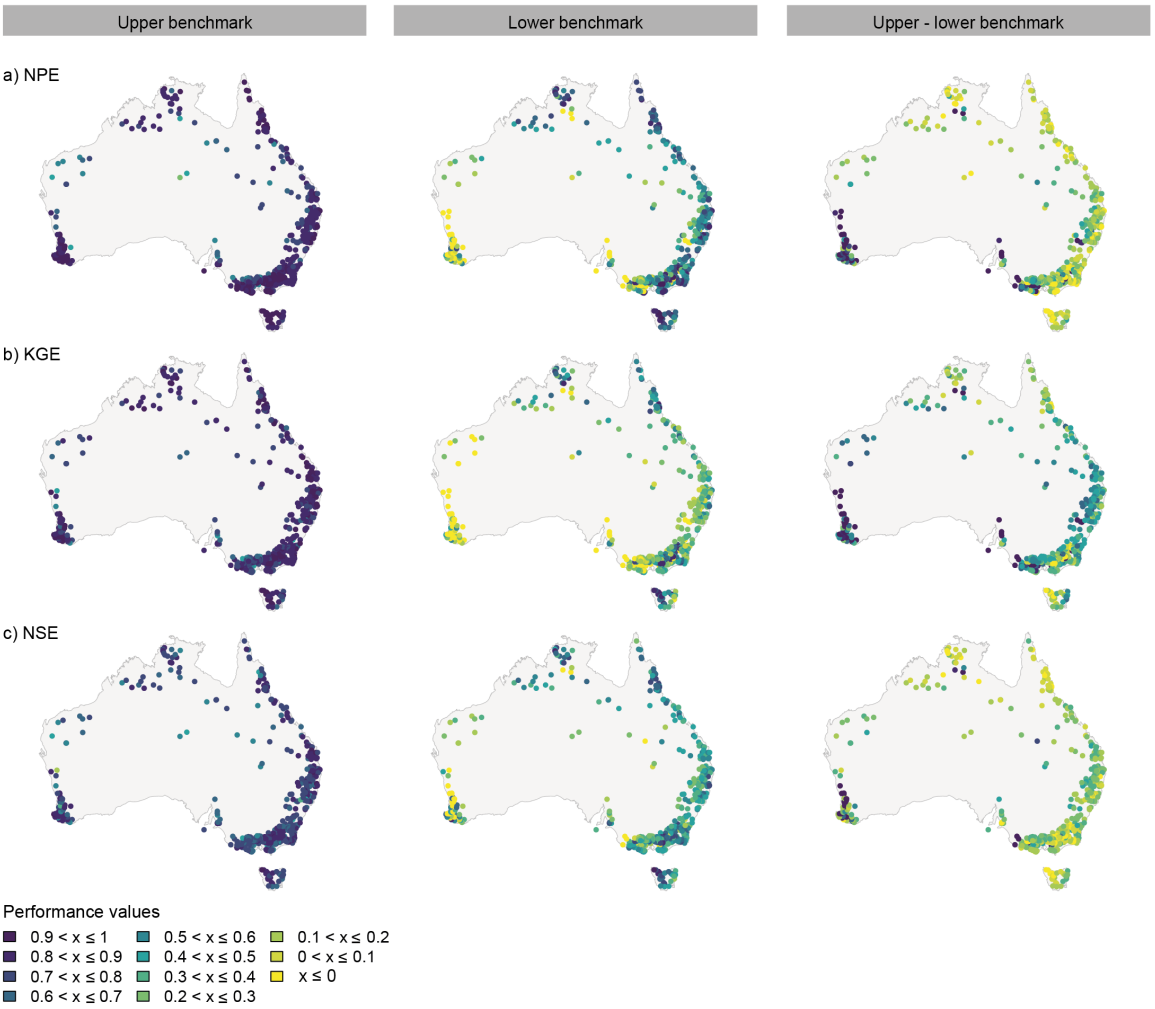

Figure S1: Spatial patterns of model performance for catchments in the Australia (CAMELS-AUS dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

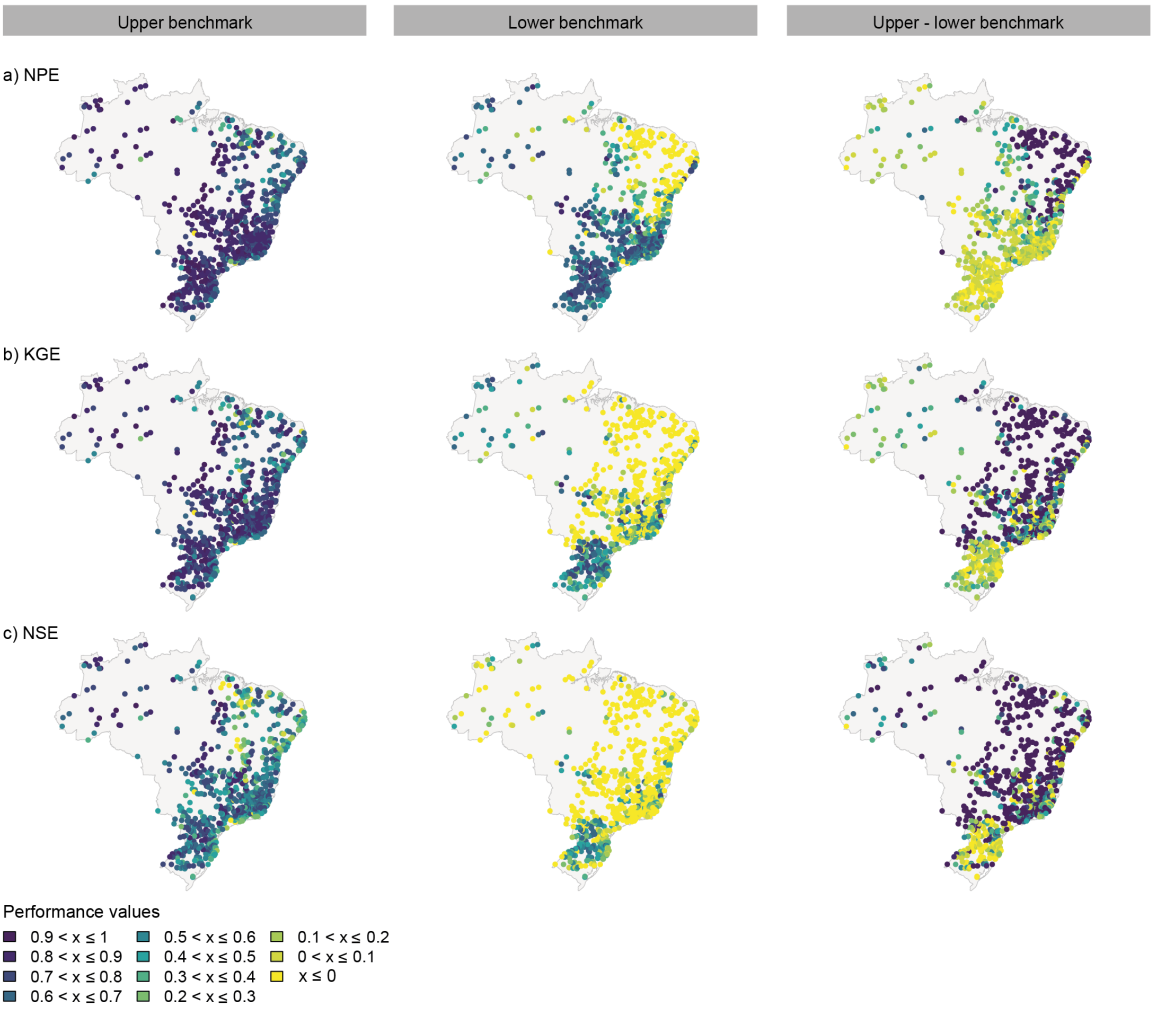

Figure S2: Spatial patterns of model performance for catchments in the Brazil (CAMELS-BR dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

Figure S3: Spatial patterns of model performance for catchments in the Central Europe (LamaH-CE dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

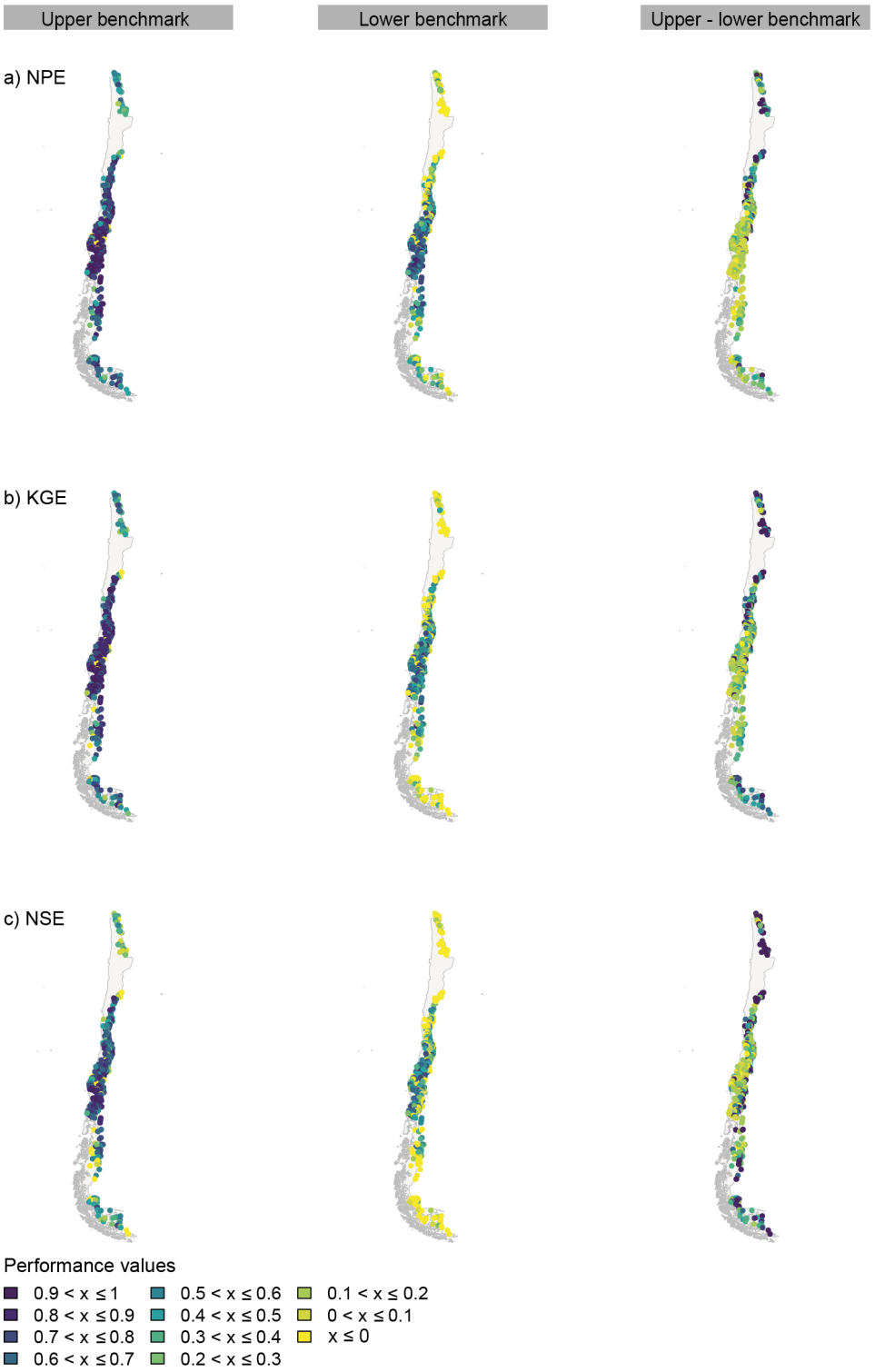

Figure S4: Spatial patterns of model performance for catchments in the Chile (CAMELS-CL dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

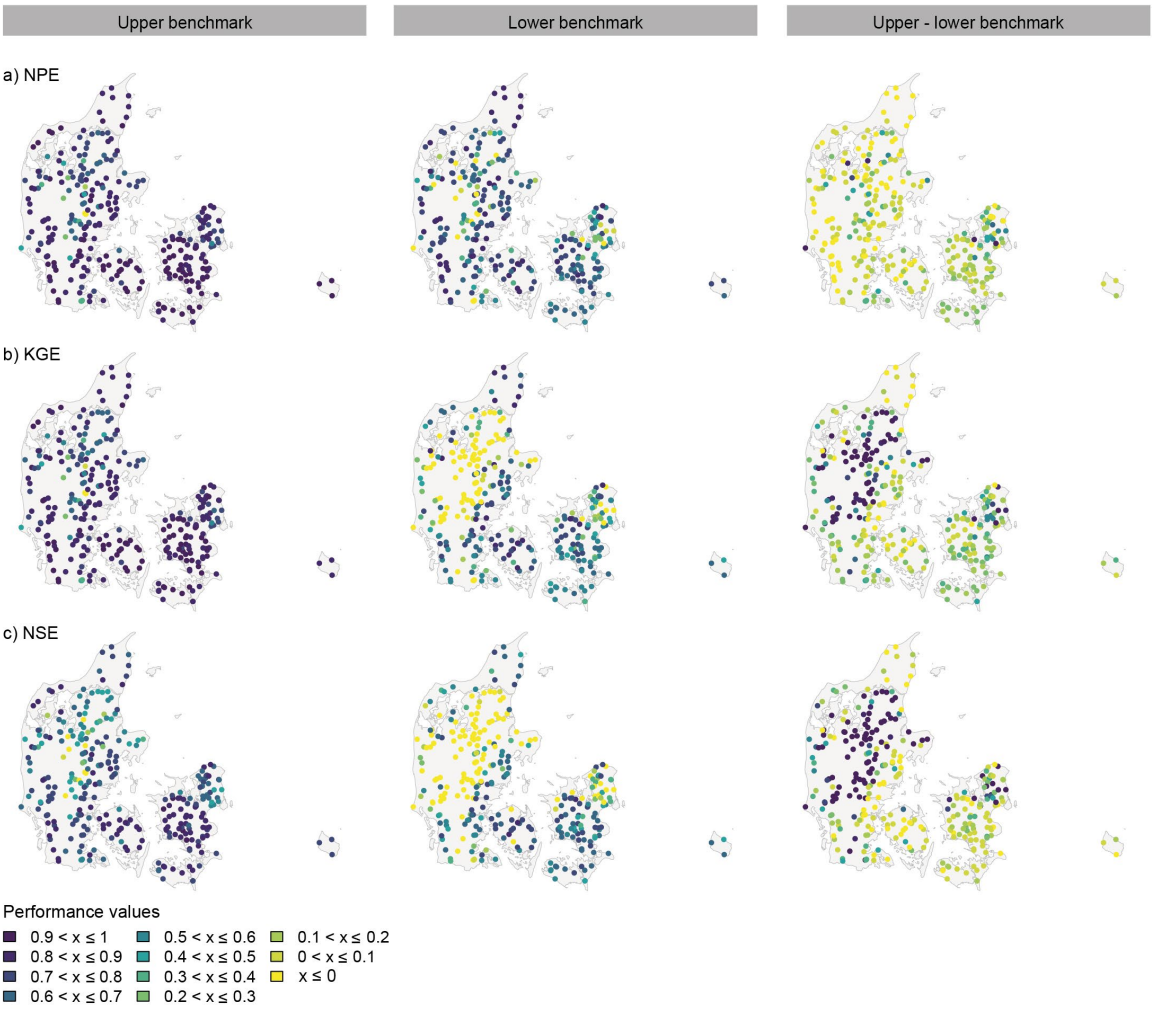

Figure S5: Spatial patterns of model performance for catchments in the Denmark (CAMELS-DK dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

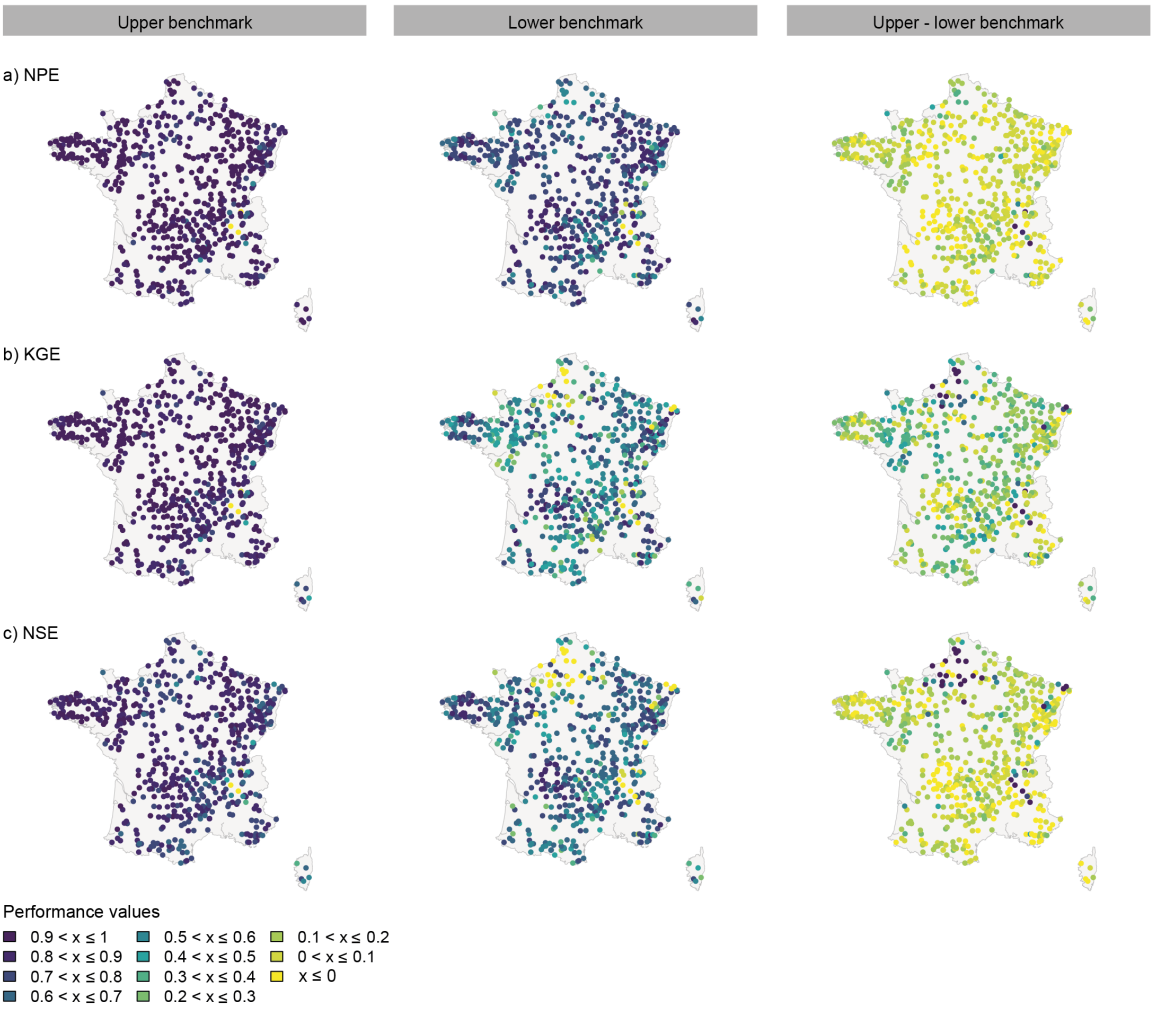

Figure S6: Spatial patterns of model performance for catchments in the France (CAMELS-FR dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

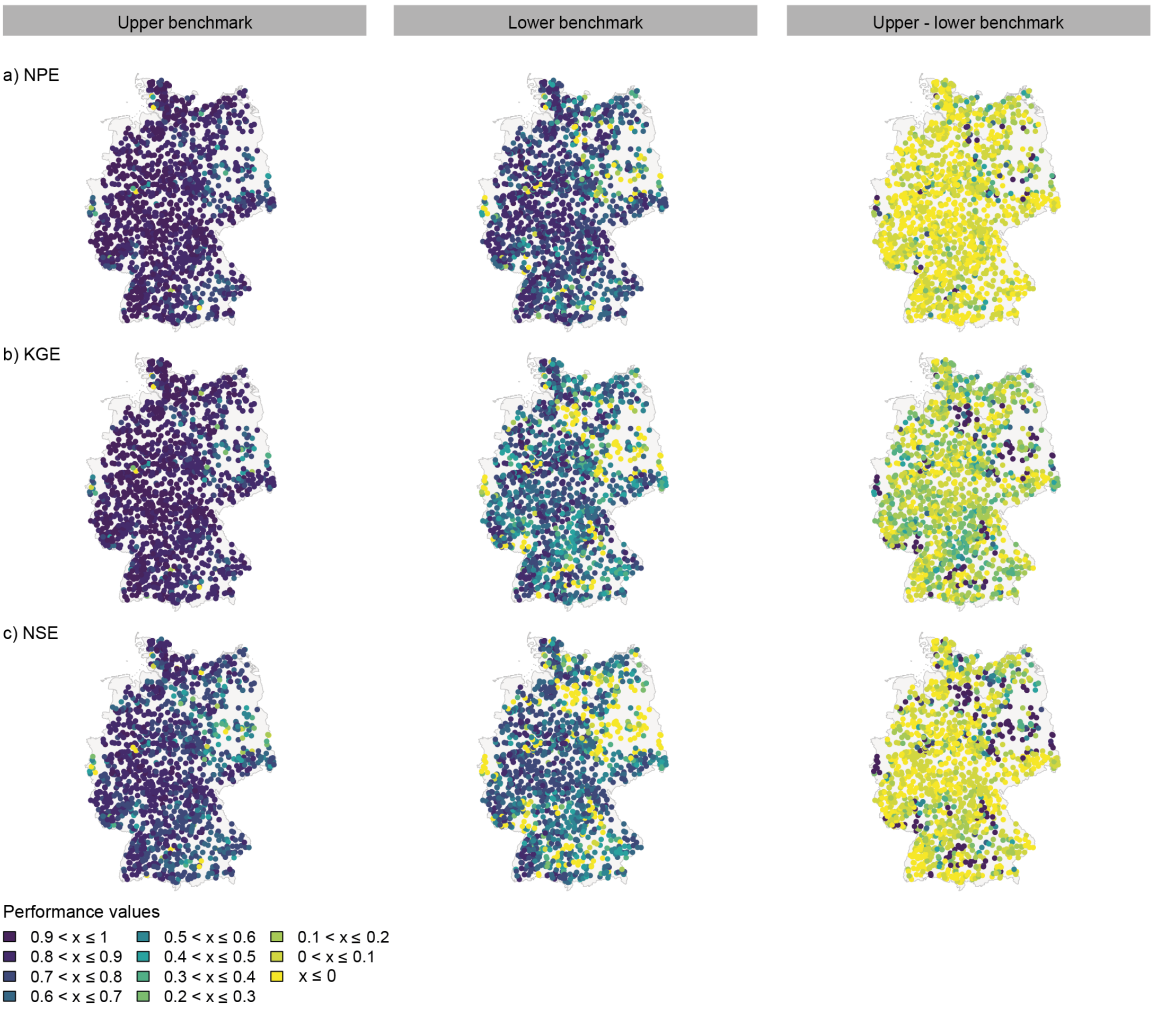

Figure S7: Spatial patterns of model performance for catchments in the Germany (CAMELS-DE dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

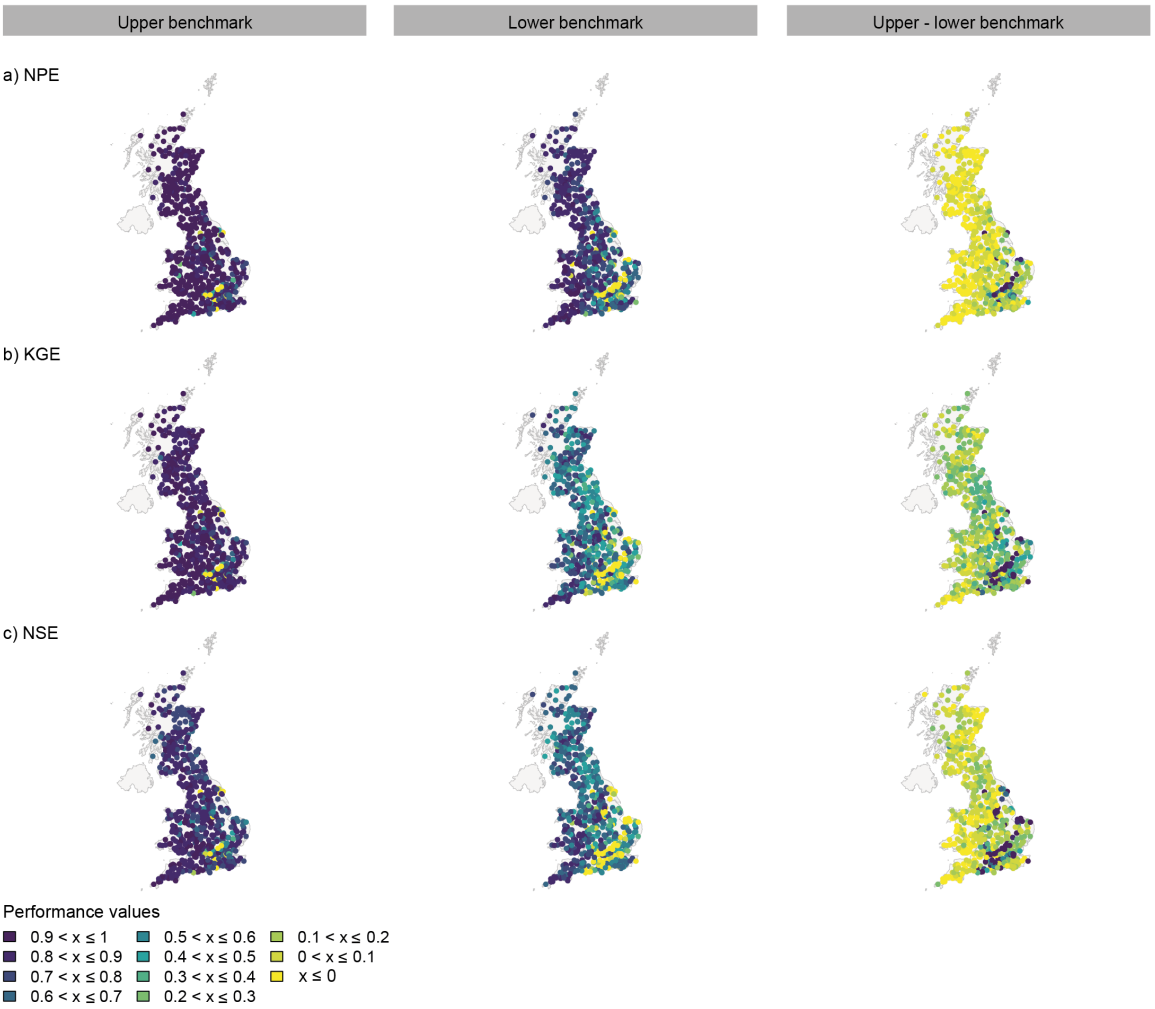

Figure S8: Spatial patterns of model performance for catchments in the Great Britain (CAMELS-GB dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

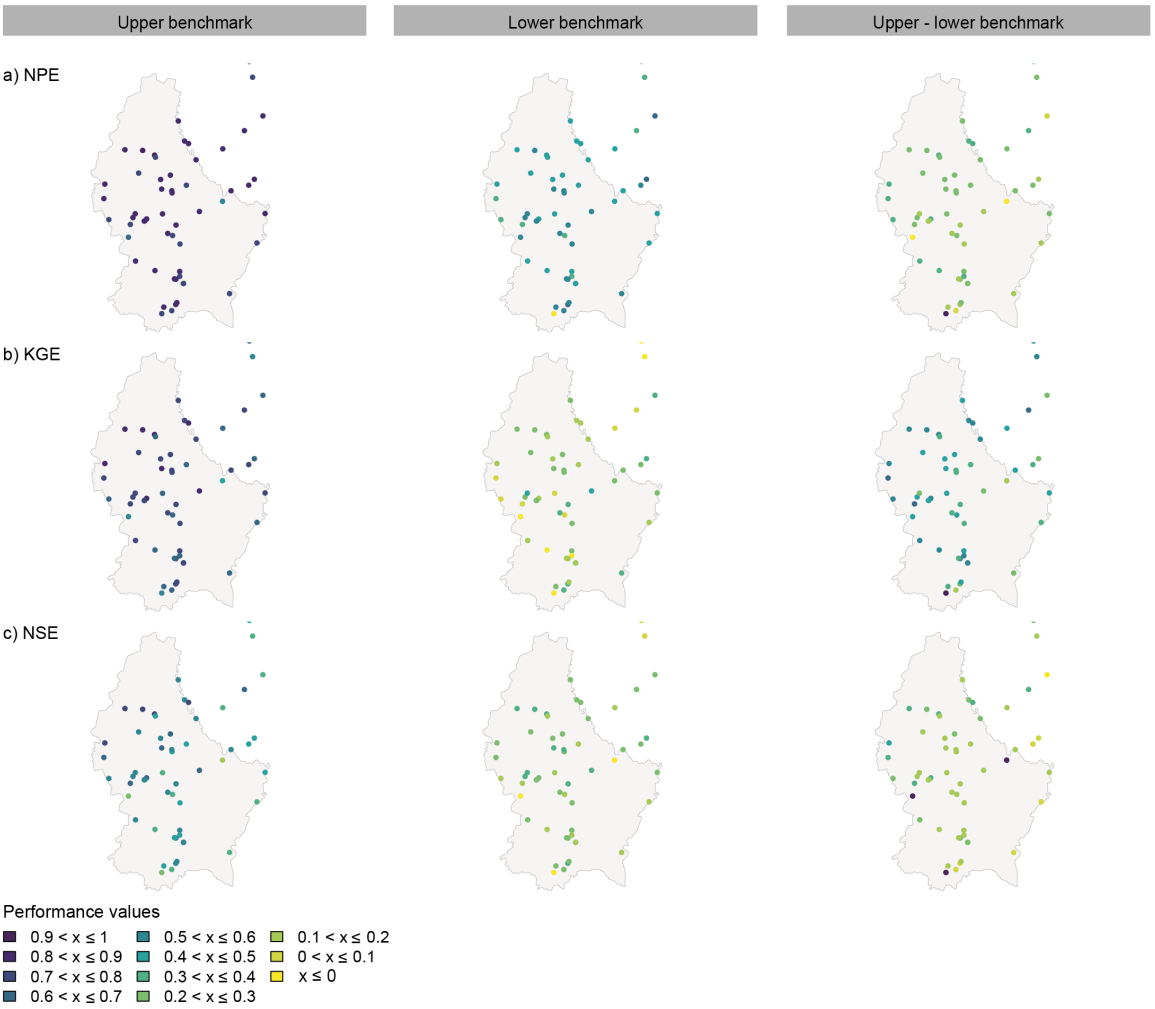

Figure S9: Spatial patterns of model performance for catchments in the Luxembourg (CAMELS-LUX dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

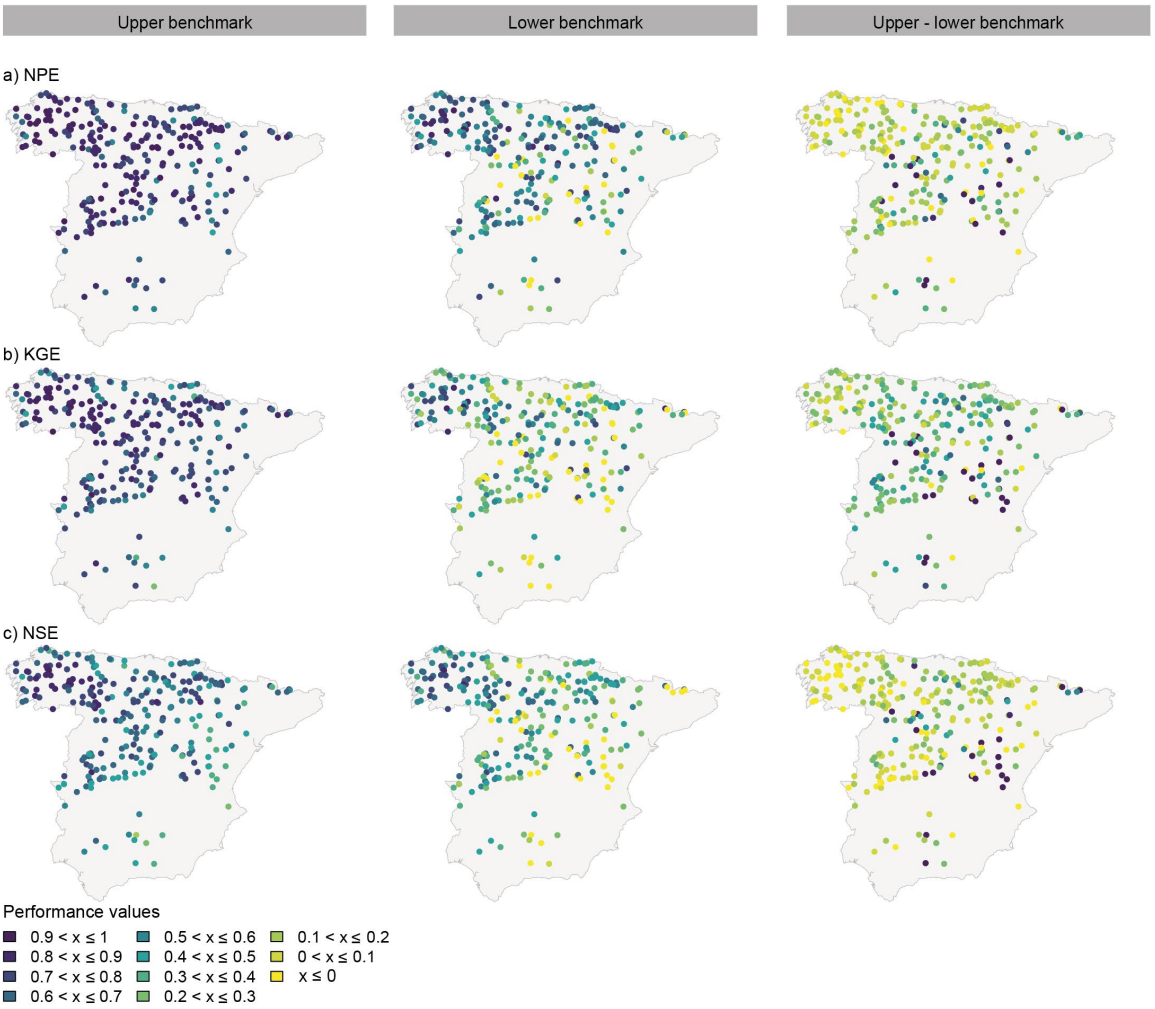

Figure S10: Spatial patterns of model performance for catchments in the Spain (CAMELS-ES dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

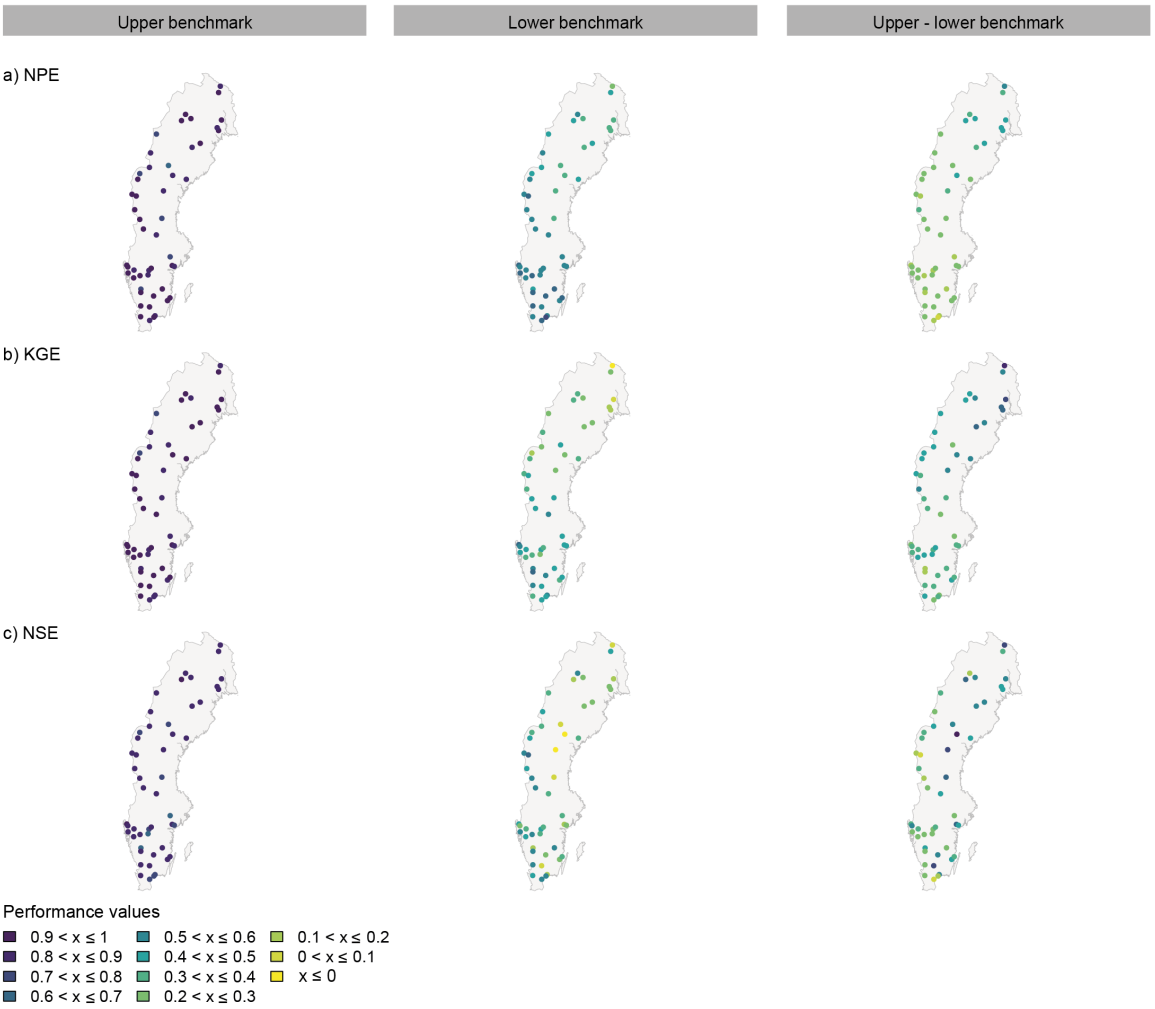

Figure S11: Spatial patterns of model performance for catchments in the Sweden (CAMELS-SE dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

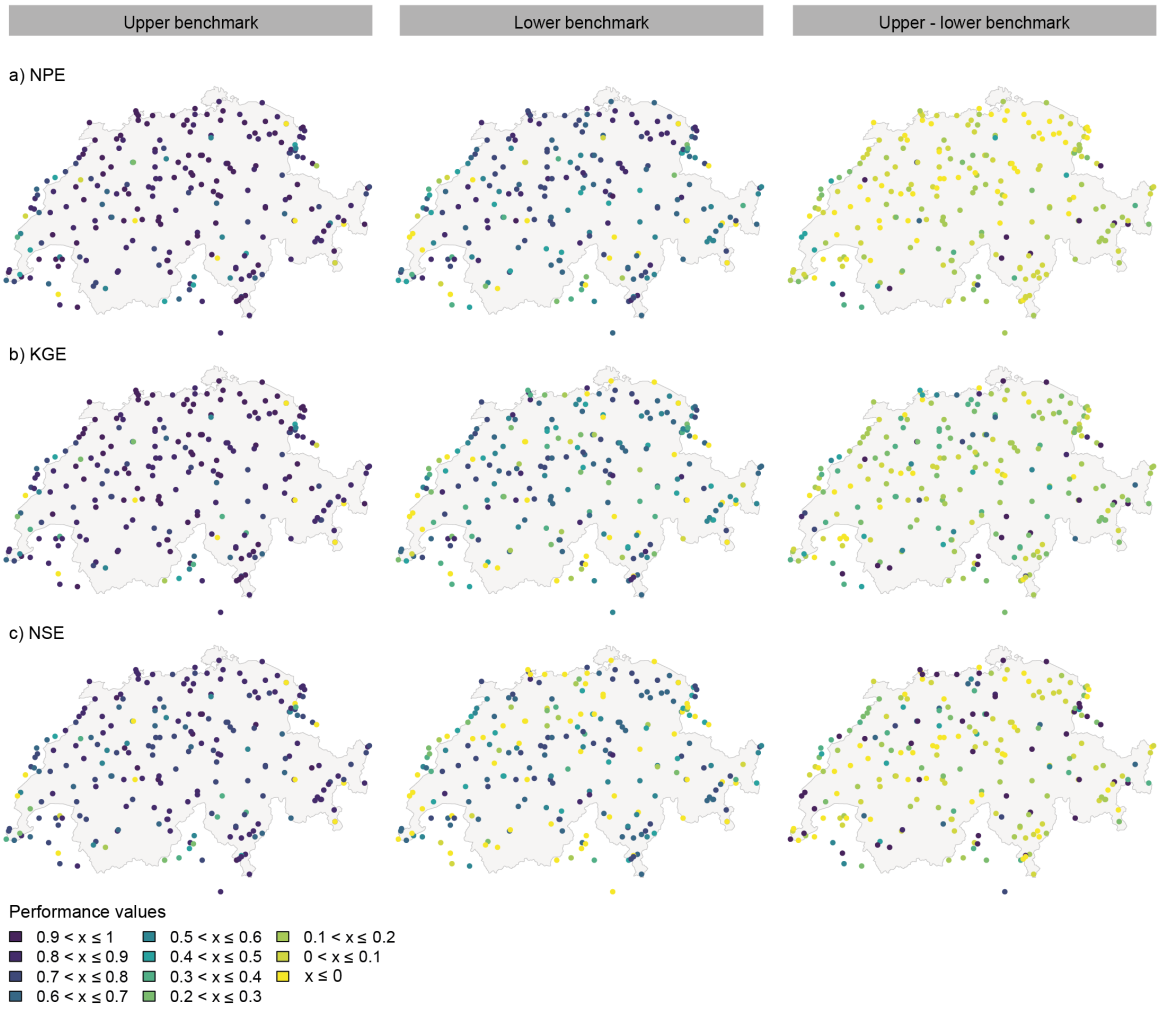

Figure S12: Spatial patterns of model performance for catchments in the Switzerland (CAMELS-CH dataset) for the upper benchmark (first column), lower benchmark (represented by the ensemble mean of 1000 parameterizations, second column), and the difference between upper and lower benchmark (third column). Values are shown for a) NPE, b) KGE, and c) NSE.

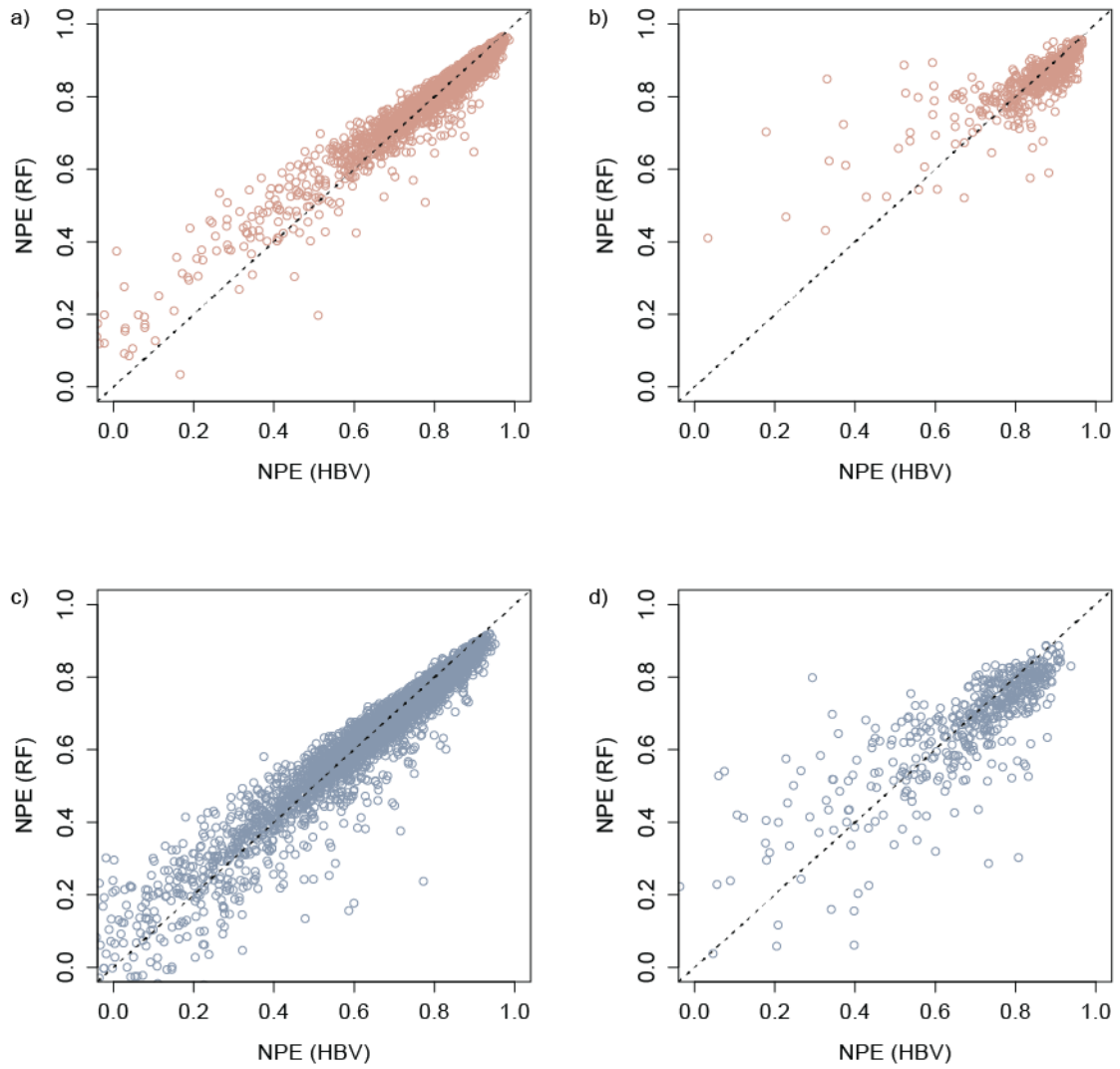

Figure S13: Example of a random forest trained and tested using all CAMELS datasets. The training set was based on 90% of the catchments and the test set was based on the remaining 10% of the catchments. We did a ten-fold validation, i.e., testing and training was repeated 10 times. The example shows one of the ten random forests used to predict RNP (files rf_UB_NPE_1.RDat and rf_LB_NPE_1.RDat). The x-axis shows the "observed" RNP from the HBV model and the y-axis shows the RNP values "simulated" by the random forest. a) RNP for the upper benchmark in the training data set, b) RNP for the upper benchmark in the test data set, c) RNP for the lower benchmark in the training data set, and d) RNP for the lower benchmark in the test data set.

Table S1: Performance of the ten individual random forest models (RF.1 to RF.10) in terms of Pearson correlation coefficient (r) and mean absolute error (MAE) when predicting RNP.

Case	Metric	Sample	RF.1	RF.2	RF.3	RF.4	RF.5	RF.6	RF.7	RF.8	RF.9	RF.10	RF.Mean
UB	*r*	Training	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
UB	*r*	Testing	0.76	0.78	0.82	0.79	0.85	0.82	0.83	0.77	0.96	0.94	0.83
LB	*r*	Training	0.96	0.96	0.97	0.96	0.95	0.97	0.96	0.96	0.96	0.96	0.96
LB	*r*	Testing	0.91	0.71	0.59	0.82	0.94	0.7	0.72	0.84	0.82	0.88	0.79
UB	*MAE*	Training	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
UB	*MAE*	Testing	0.05	0.05	0.05	0.04	0.05	0.05	0.05	0.06	0.04	0.05	0.05
LB	*MAE*	Training	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
LB	*MAE*	Testing	0.13	0.14	0.13	0.13	0.11	0.12	0.13	0.12	0.14	0.12	0.13

Table S2: same as Table S1, but for RKG.

Case	Metric	Sample	RF.1	RF.2	RF.3	RF.4	RF.5	RF.6	RF.7	RF.8	RF.9	RF.10	RF.Mean
UB	*r*	Training	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98	0.98
UB	*r*	Testing	0.67	0.73	0.83	0.82	0.86	0.83	0.89	0.78	0.95	0.92	0.83
LB	*r*	Training	0.97	0.96	0.97	0.96	0.96	0.97	0.97	0.97	0.96	0.96	0.97
LB	*r*	Testing	0.96	0.66	0.65	0.77	0.91	0.69	0.74	0.85	0.83	0.91	0.8
UB	*MAE*	Training	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
UB	*MAE*	Testing	0.06	0.06	0.05	0.05	0.05	0.05	0.05	0.06	0.05	0.06	0.06
LB	*MAE*	Training	0.1	0.1	0.09	0.1	0.1	0.09	0.09	0.1	0.09	0.1	0.1
LB	*MAE*	Testing	0.19	0.22	0.25	0.22	0.2	0.22	0.25	0.2	0.23	0.2	0.22

T*able S3: same as Table S1, but for RNS.*

Case	Metric	Sample	RF.1	RF.2	RF.3	RF.4	RF.5	RF.6	RF.7	RF.8	RF.9	RF.10	RF.Mean
UB	*r*	Training	0.96	0.97	0.96	0.96	0.96	0.96	0.97	0.97	0.97	0.97	0.96
UB	*r*	Testing	0.4	0.39	0.45	0.74	0.59	0.64	0.84	0.72	0.93	0.94	0.66
LB	*r*	Training	0.97	0.96	0.97	0.95	0.95	0.97	0.97	0.97	0.98	0.97	0.97
LB	*r*	Testing	0.98	0.23	0.29	0.37	0.37	0.27	0.33	0.43	0.65	0.85	0.48
UB	*MAE*	Training	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
UB	*MAE*	Testing	0.11	0.12	0.1	0.1	0.11	0.09	0.09	0.14	0.13	0.14	0.11
LB	*MAE*	Training	1.22	1.22	1	1.27	1.3	1.08	1.04	1.23	0.93	1.12	1.14
LB	*MAE*	Testing	2.05	2.57	3.9	2.38	2.8	3.12	3.39	1.84	5.29	2.84	3.02

Table S4: Version number and the column names (in green) or file names (in blue) of the precipitation (P), temperature (T), potential evapotranspiration (PET), and discharge (Q) time series that were used for the various model runs.

Camels	Version	P series	T series	PET series	Q series
CAMELS-AUS	Version 2.02	precipitation_SILO.csv	tmin_SILO.csv and tmax_SILO.csv	et_morton_point_silo.csv	streamflow_mmd.csv
CAMELS-BR	Version 1.2	p_chirps	tmean_era5land	pet_gleam	streamflow_mm
CAMELS-CH	Version 0.7	precipitation(mm/d)	temperature_mean(degC)	pet_sim(mm/d)	discharge_spec(mm/d)
CAMELS-CL	-	4_CAMELScl_precip_cr2met.txt	10_CAMELScl_tmean_cr2met.txt	12_CAMELScl_pet_hargreaves.txt	3_CAMELScl_streamflow_mm.txt
CAMELS-DE	Version 1.0.0	precipitation_mean	temperature_mean	pet_hargreaves	discharge_spec_obs
CAMELS-DK	Version 3.0	precipitation	temperature	pet	Qobs
CAMELS-ES	Version 1.1.0	total_precipitation_sum	temperature_2m_mean	potential_evaporation_sum_FAO_PENMAN_MONTEITH	streamflow
CAMELS-FR	Version 2.1	tsd_prec	tsd_temp	tsd_pet_pe	tsd_q_mm
CAMELS-GB	v1 [1)]	precipitation	temperature	pet	discharge_spec
CAMELS-LUX	Version v1	RR_rad	t2m	PET_PM	Qspec
CAMELS-SE	Version 1	Pobs_mm	Tobs_C	PET_Hamon_mm [2)]	Qobs_mm
CAMELS-US	Version 1.2	prcp(mm/day)	tmin(C) and tmax(C)	[3)]	[4)]
LamaH-CE	Version 1.0	prec	2m_temp_mean	PET_A	qobs

1) No official version number, but on 2025-10-21 CAMELS-GB v2 was published.
2) PET data files received per email.
3) PET values are not provided. PET values were calculated by us using the Priestley-Taylor equation instead.
4) Column names are missing, but streamflow values are given in the 5th column.