Supplementary information for

Constraining climate model projections with observations amplifies future runoff declines

Hanjun Kim[1], Flavio Lehner[1,2,3],

Katherine Dagon[2], David M. Lawrence[2], Sean Swenson[2], Andrew W. Wood[2,4]

[1]Department of Earth and Atmospheric Sciences, Cornell University, Ithaca, NY, USA

[2]Climate and Global Dynamics Laboratory, NSF National Center for Atmospheric Research, Boulder, CO, USA

[3]Polar Bears International, Bozeman, MT, USA

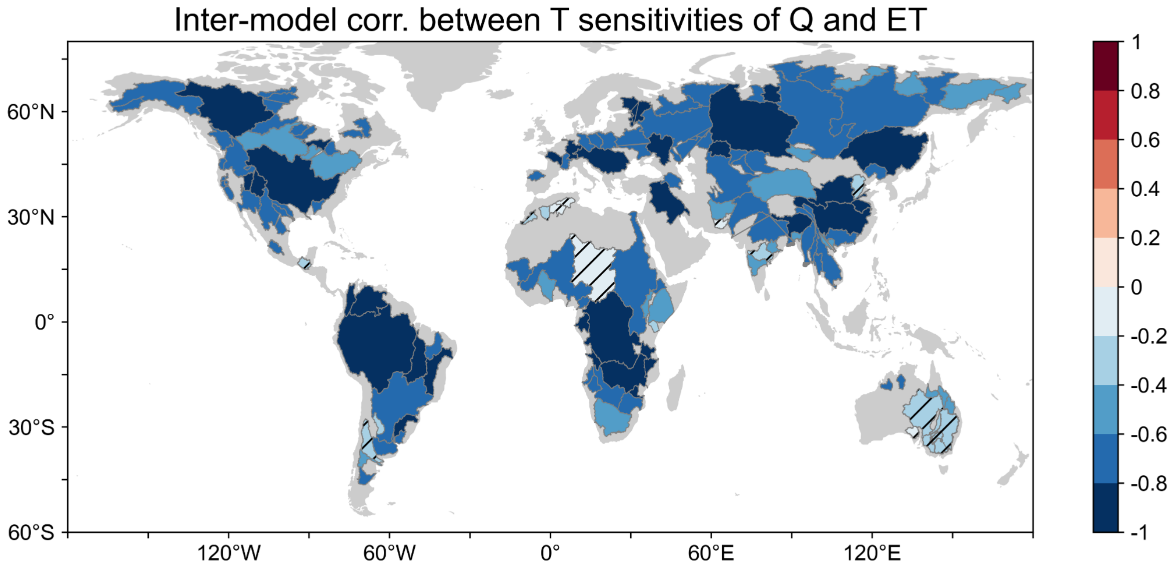
[4]Department of Civil and Environmental Engineering, Colorado School of Mines, Golden, CO, USA

[*]*Corresponding author*: Hanjun Kim (hk764@cornell.edu)

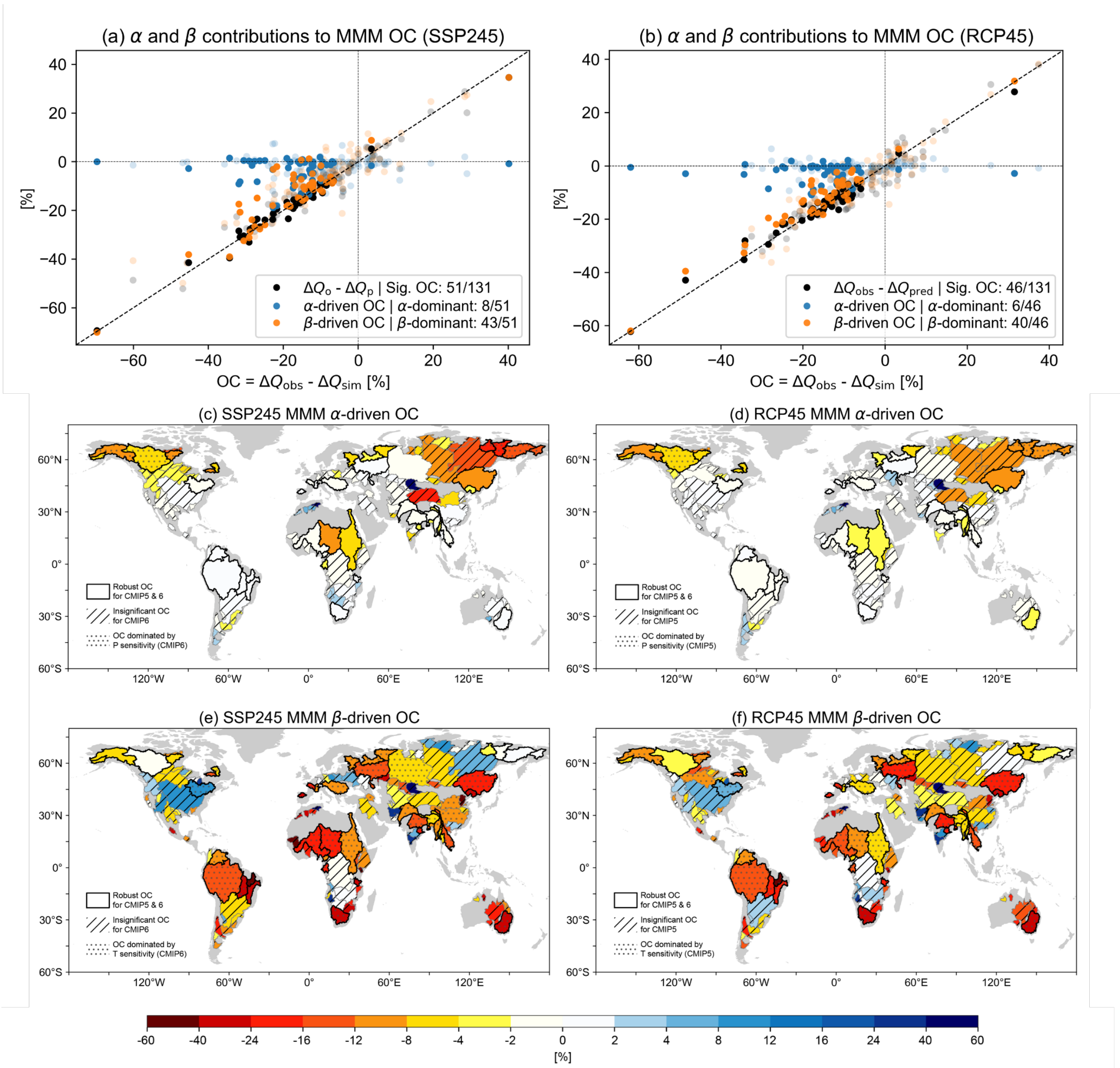
CMIP6 models			
Model names	Institution	Fraction of closed basins (area weighted)	# of Basins with negative values (>10%)
ACCESS-CM2	CSIRO and ARCCSS, Australia	0.819	
ACCESS-ESM1-5	CSIRO and ARCCSS, Australia	0.843	
BCC-CSM2-MR	BCC, China	0.816	11
CAMS-CSM1-0	CAMS, China	0.693	
CAS-ESM2-0	CAS, China	0.283	1
CESM2	NCAR, USA	0.975	2
CESM2-WACCM	NCAR, USA	0.977	2
CMCC-CM2-SR5	CMCC, Italy	0.032	5
CMCC-ESM2	CMCC, Italy	0.044	5
CNRM-CM6-1	CNRM, France	0.613	
CNRM-CM6-1-HR	CNRM, France	0.664	
CNRM-ESM2-1	CNRM, France	0.602	
CanESM5	CCCma, Canada	0.728	
CanESM5-1	CCCma, Canada	0.716	
EC-Earth3	SMHI, Sweden / EC-Earth Consortium	0.762	
EC-Earth3-CC	SMHI, Sweden / EC-Earth Consortium	0.882	
EC-Earth3-Veg	SMHI, Sweden / EC-Earth Consortium	0.873	
EC-Earth3-Veg-LR	SMHI, Sweden / EC-Earth Consortium	0.849	
FIO-ESM-2-0	FIO, China	0.943	2
GFDL-CM4	NOAA GFDL, USA	0.112	
GFDL-ESM4	NOAA GFDL, USA	0.114	
GISS-E2-1-G	NASA/GISS, USA	0.412	
GISS-E2-1-H	NASA/GISS, USA	0.322	
HadGEM3-GC31-LL	MOHC, UK	0.655	
INM-CM4-8	INM, Russia	0.929	9
INM-CM5-0	INM, Russia	0.936	7
IPSL-CM6A-LR	IPSL, France	0.871	
MCM-UA-1-0	Univ. of Arizona, USA	0.899	
MIROC6	AORI, NIES, and JAMSTEC, Japan	0.926	
MPI-ESM1-2-HR	MPI-M, Germany	0.859	
MPI-ESM1-2-LR	MPI-M, Germany	0.816	
MRI-ESM2-0	MRI, Japan	0.816	
NorESM2-LM	NCC, Norway	0.835	6
NorESM2-MM	NCC, Norway	0.980	4
UKESM1-0-LL	MOHC, UK	0.669	

CMIP5 models			
Model names	Institution	Fraction of closed basins (area weighted)	# of Basins with negative values (>10%)
bcc-csm1-1	BCC, CMA, China	0.882	3
bcc-csm1-1-m	BCC, CMA, China	0.952	6
CCSM4	NCAR, USA	0.953	1
CESM1-BGC	NCAR, USA	0.953	1
CMCC-CM	CMCC, Italy	0.851	
CNRM-CM5	CNRM, France	0.851	
CanESM2	CCCma, Canada	0.896	
FGOALS-g2	CAS, china	0.903	4
FIO-ESM	FIO, China	0.322	4
GFDL-CM3	NOAA GFDL, USA	0.738	
GFDL-ESM2G	NOAA GFDL, USA	0.777	
GFDL-ESM2M	NOAA GFDL, USA	0.800	
GISS-E2-H	NASA/GISS, USA	0.054	
GISS-E2-H-CC	NASA/GISS, USA	0.038	
GISS-E2-R	NASA/GISS, USA	0.031	
GISS-E2-R-CC	NASA/GISS, USA	0.043	
inmcm4	INM, Russia	0.882	9
IPSL-CM5A-LR	IPSL, France	0.702	
IPSL-CM5A-MR	IPSL, France	0.714	
MIROC-ESM	AORI, NIES, and JAMSTEC, Japan	0.899	
MIROC-ESM-CHEM	AORI, NIES, and JAMSTEC, Japan	0.899	
MIROC5	AORI, NIES, and JAMSTEC, Japan	0.911	
MPI-ESM-LR	MPI-M, Germany	0.905	
MPI-ESM-MR	MPI-M, Germany	0.906	
MRI-CGCM3	MRI, Japan	0.758	
NorESM1-M	NCC, Norway	0.852	2
NorESM1-ME	NCC, Norway	0.832	2

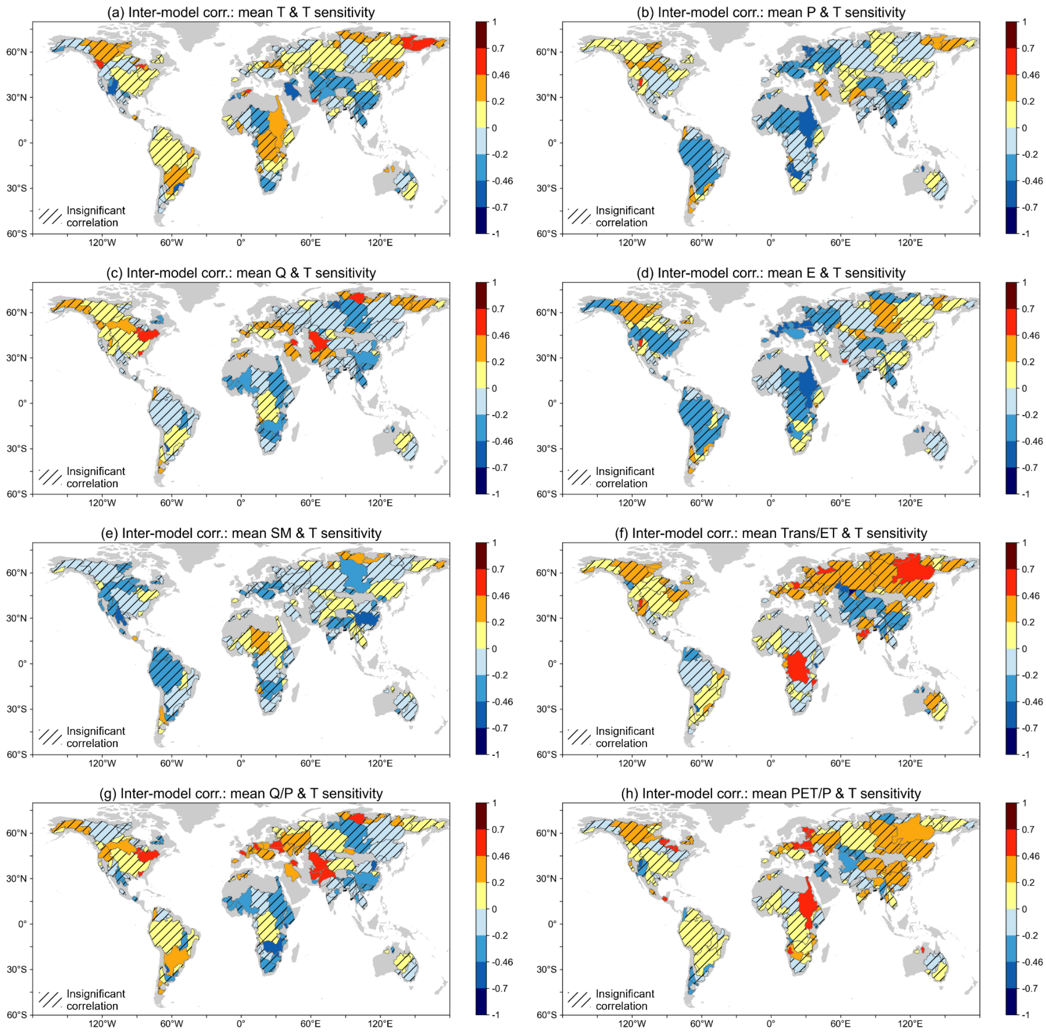
Supplementary Table. 1| List of CMIP5/6 models used in the study. The models providing total runoff, precipitation, surface air temperature, and evapotranspiration across historical and model scenarios (SSP2-4.5, RCP4.5) are listed. Gray shading indicates models which do not satisfy the water budget closure, determined as when the area fraction of basins with closed water balance is less than 0.6 of global river basins (Method). Number of basins with negative runoff values is also described for each model (Method).



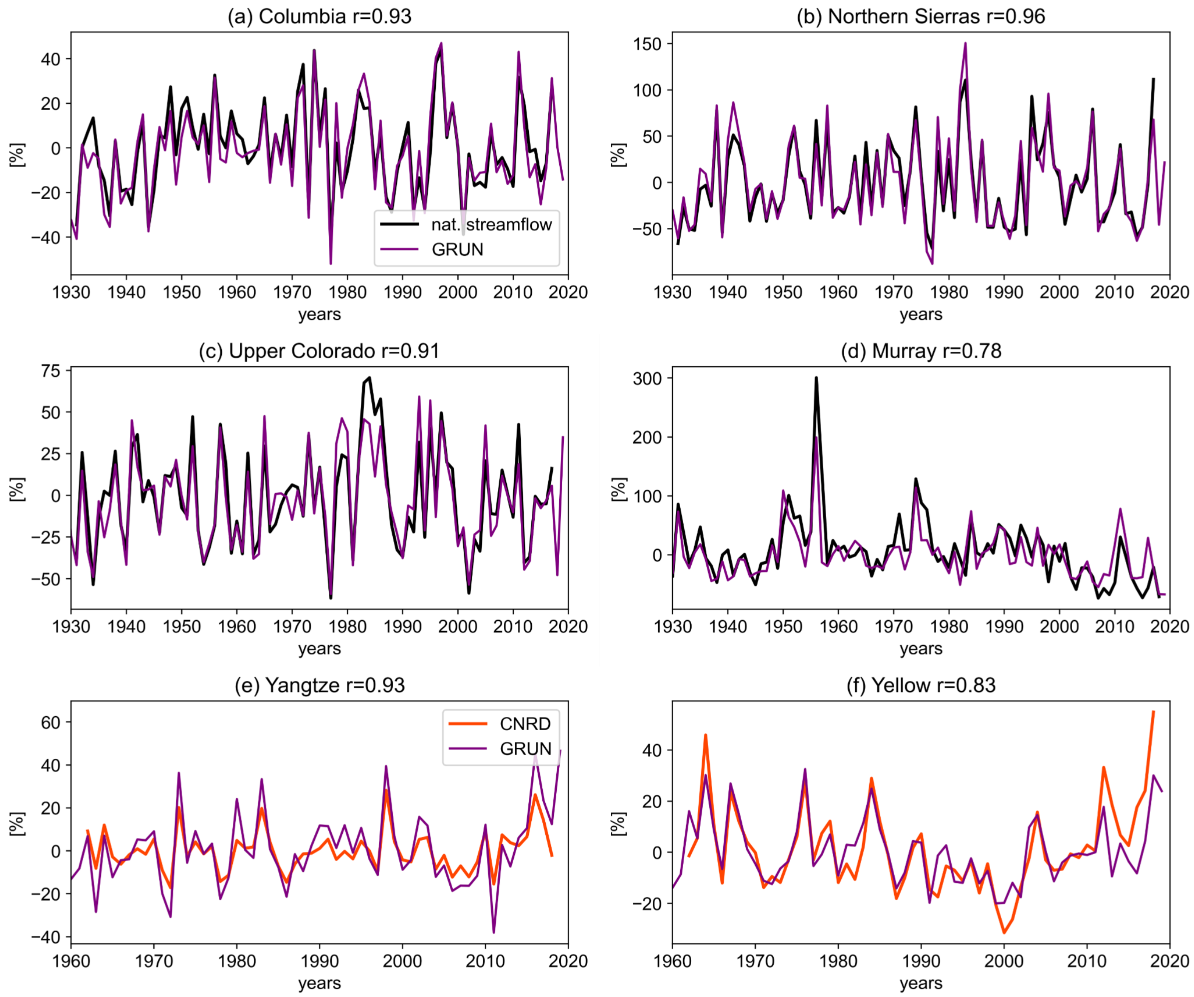
Supplementary Fig. 1| Relationship between the temperature sensitivities of runoff and evapotranspiration. Inter-model correlation coefficients between the historical temperature sensitivities of runoff ($\delta Q/\delta T$) and evapotranspiration ($\delta ET/\delta T$). The sensitivity of evapotranspiration is calculated by substituting evapotranspiration (ET) instead of runoff (Q) in Eq. (3), following the methodology for runoff sensitivity (Methods). Hatched basins indicate where the correlation coefficient is statistically not significant using a *t*-test at 95% confidence level.



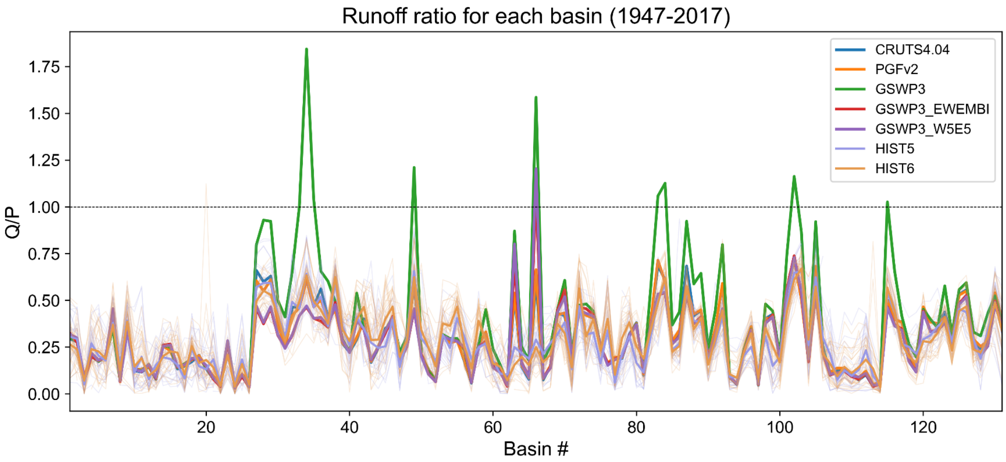
Supplementary Fig. 2| Contribution of temperature and precipitation sensitivity bias to the observational constraint. **a,b,** Observational constraint relative to predictions versus relative to simulated runoff projections for (**a**) CMIP6 and (**b**) CMIP5. Observational constraining effects are further decomposed into components driven by T sensitivity (α) and P sensitivity (β). **c-f**, The multi-model median (MMM) observational constraining effect driven by P sensitivity bias for (**c**) SSP2-4.5 and (**d**) RCP4.5 scenarios, and T sensitivity bias for (**e**) SSP2-4.5 and (**f**) RCP4.5 scenarios.



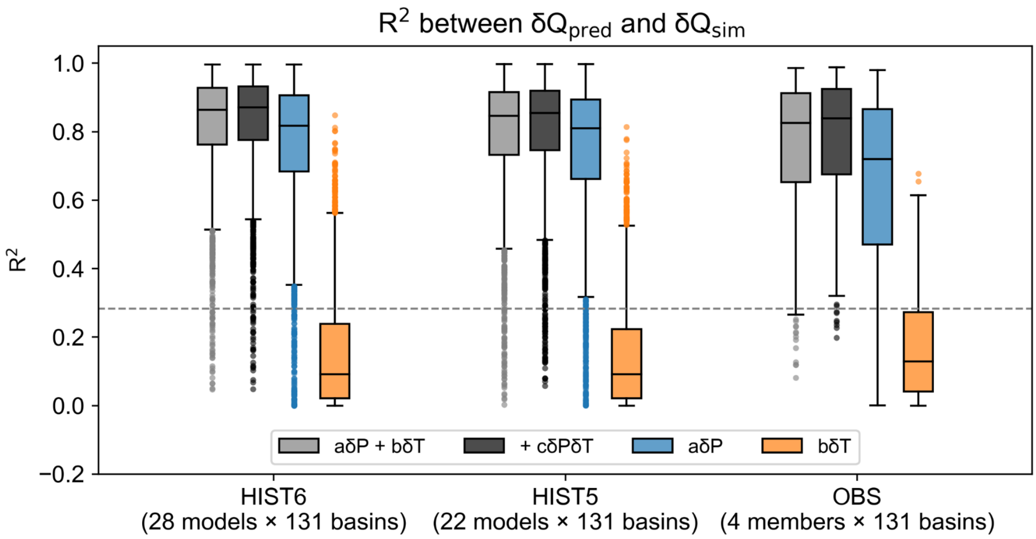
Supplementary Fig. 3| Inter-model correlation of temperature sensitivity to various mean state variables. **a-h**, Inter-model correlation coefficient of historical (1947-2017) temperature sensitivity to historical mean state variables: (**a**) temperature (T), (**b**) precipitation (P), (**c**) runoff (Q), (**d**) evapotranspiration (ET), (**e**) soil moisture in the upper 10 cm of the upper layer (SM), (**f**) transpiration divided by evapotranspiration (Trans/ET), (**g**) runoff ratio (Q/P), and (**h**) potential evapotranspiration divided by precipitation (PET/P). Hatchings indicate regions where the correlation coefficient is not statistically significant at the 95% confidence level, with the criterion for significance being r=±0.37.



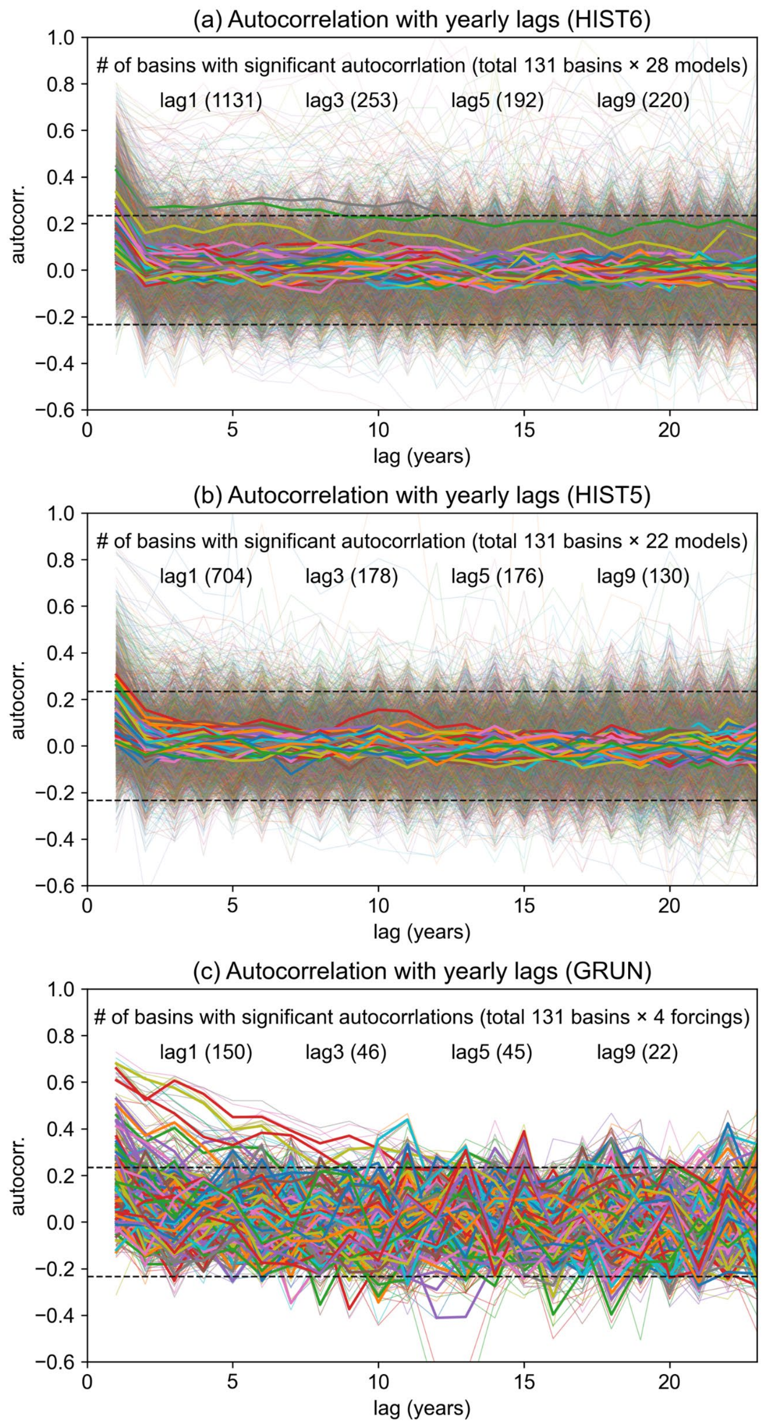
Supplementary Fig. 4| Comparison between Global Runoff Reanalysis (GRUN) and other observational runoff datasets. **a-c**, Water-year and basin averaged runoff for (**a**) Columbia, (**b**) Norther Sierras, (**c**) Upper Colorado River, and (**d**) Murray River basins with naturalized streamflow. The GRUN runoff is averaged for the basin mask corresponding to the streamflow gauge catchment. **e,f,** Water-year and basin averaged runoff for (**e**) Yangtze and (**f**) Yellow river basins, where the China Natural Runoff Dataset[1] based on hydrologic models are compared to the GRUN runoff. Inter-annual correlations between GRUN and other datasets are indicated in the titles. The runoff and streamflow are represented as percent anomalies relative to the long-term mean (1962-2017). Note that the naturalized streamflow for the Chinese river basins is produced by the Variable Infiltration Capacity model, which is specifically tunned for natural or near-natural gauge streamflow[1]. To validate GRUN runoff sensitivity with naturalized streamflow, we only use the dataset which corrects the gauge streamflow directly (a-d).



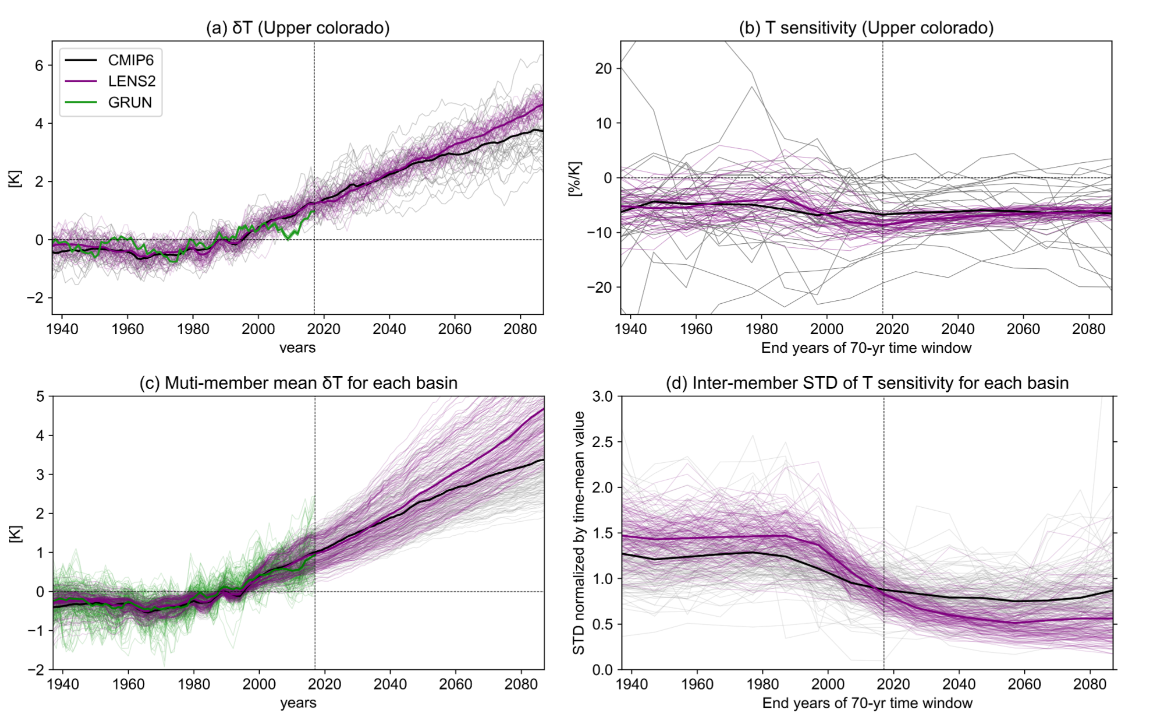
Supplementary Fig. 5| Historical runoff ratio in models and GRUN dataset. Runoff ratio calculated by dividing runoff by precipitation for each basin. The runoff ratio is calculated for each water year and then averaged for historical period (1947-2017). Results are shown for the GRUN dataset and individual CMIP5/CMIP6 models, with thin lines representing the individual CMIP models. Note that the GRUN dataset derived from GSWP3 is excluded from the main paper due to its unrealistically large runoff ratio compared to others.



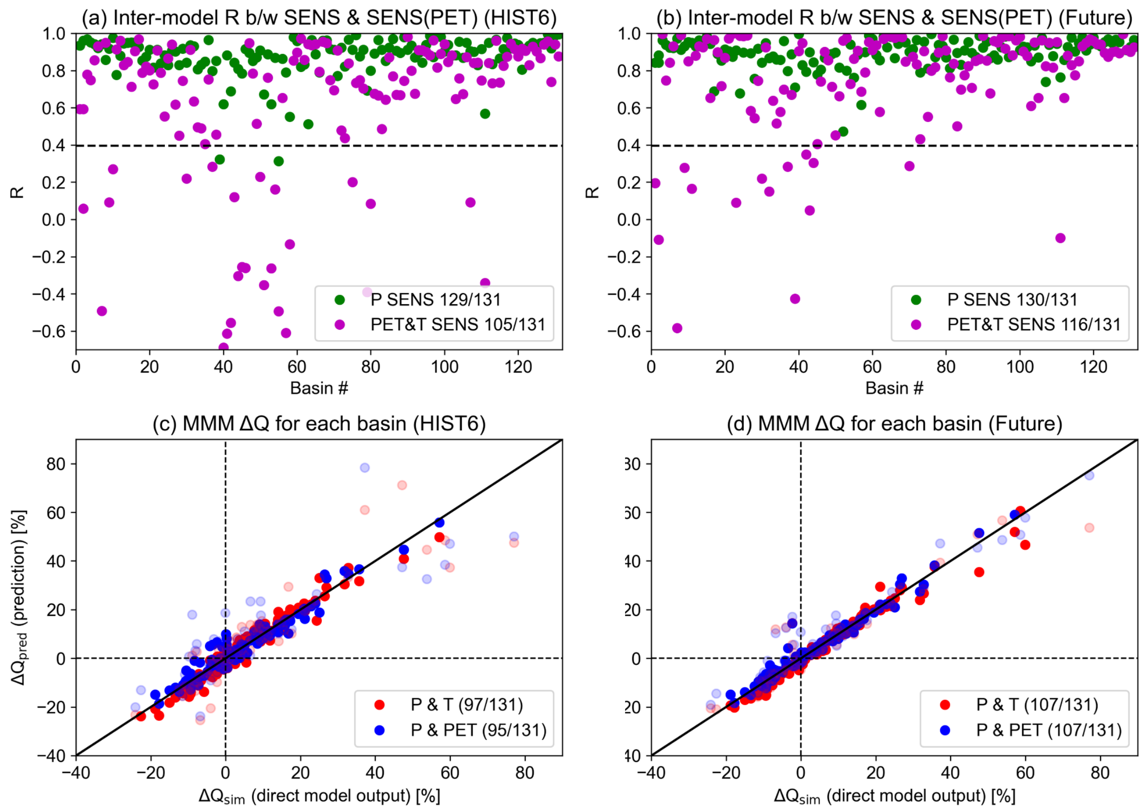
Supplementary Fig. 6| Training accuracy of the regression models using runoff sensitivity. The coefficient of determination (R^2) between runoff variations and predictions by runoff sensitivity, calculated with 5-year moving windows. R^2 values for historical period (1947-2017) of CMIP6, CMIP5, and GRUN dataset are calculated separately for each model and basin and then aggregated into a box and whisker plot. Predictions are decomposed into contributions from precipitation and temperature variations without interaction term (gray), precipitation and temperature variations with interaction term (black), precipitation variation (blue), and temperature variation (orange). The dotted line indicates statistically significant R^2, determined using *t*-test at a 95% confidence interval.



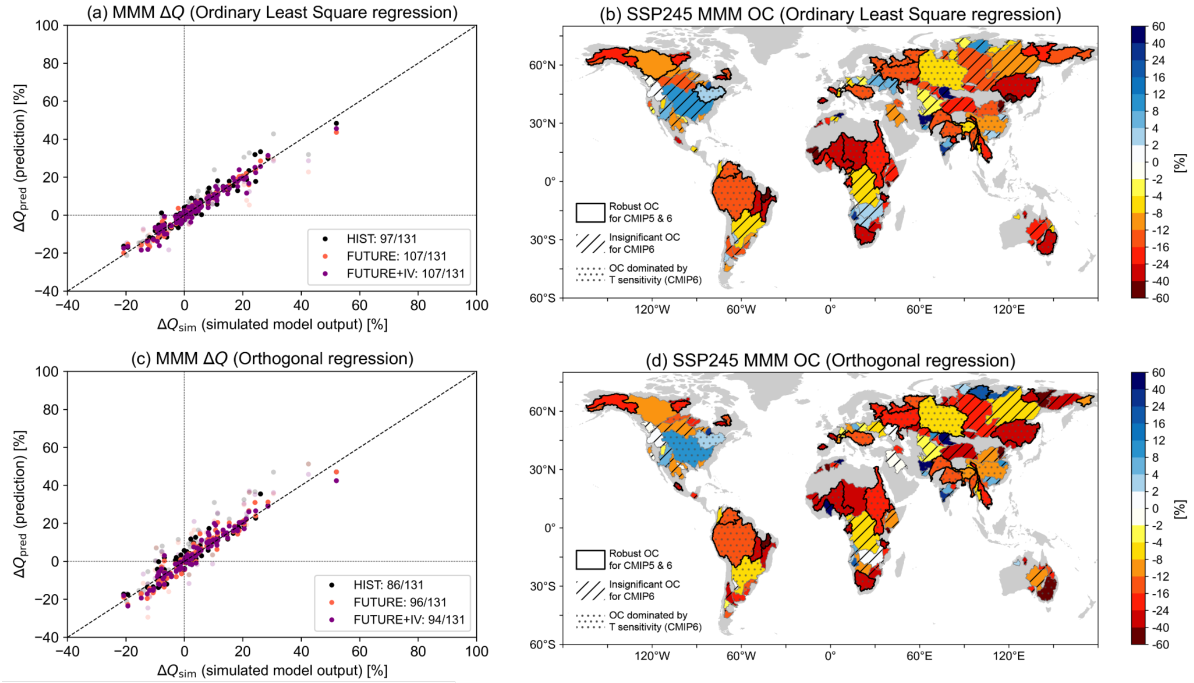
Supplementary Fig. 7| Effect of runoff storage shown by autocorrelation. a-c, Autocorrelation of annual runoff with increasing time lag years for the historical period (1947-2017) in (**a**) CMIP6, (**b**) CMIP5, and (**c**) GRUN dataset. In (a-b), thin lines represent the autocorrelation for each basin and each model, while thick lines denote the multi-model mean. In (c), thin lines correspond to each basin and each atmospheric forcing of GRUN dataset, and thick lines indicate the average across all forcing datasets. Gray dotted lines mark the statistically significant autocorrelation at 95% confidence interval.



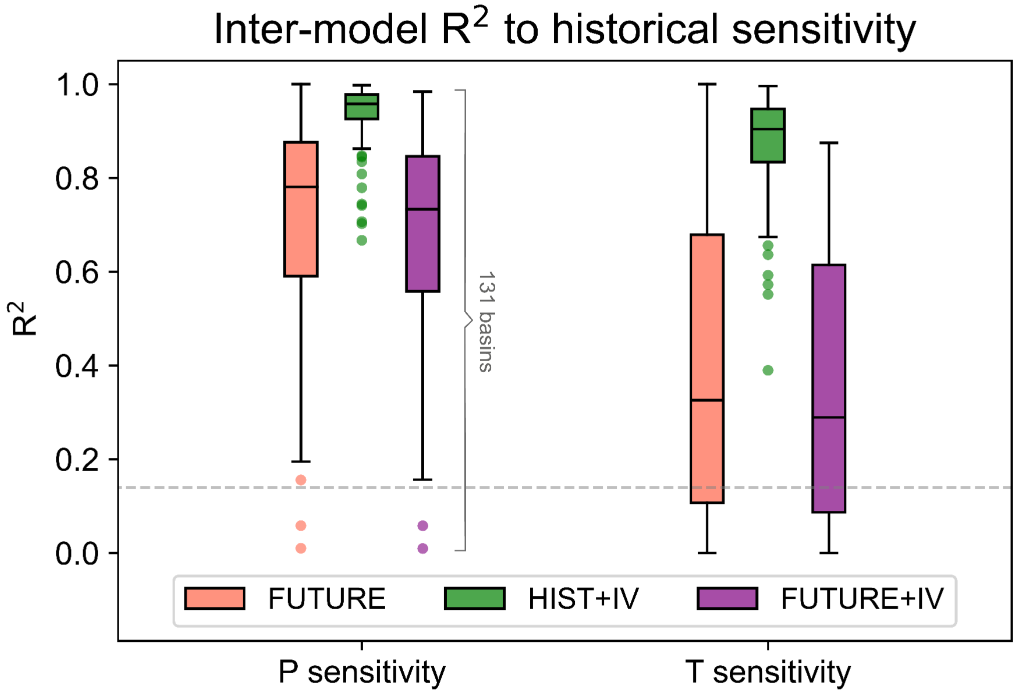
Supplementary Fig. 8| Impact of recent warming on the temperature sensitivity estimation. a, Timeseries of annual temperature anomalies in the Upper Colorado River basin compared to the historical period (1947-2017), averaged using 5-year moving windows. Thick lines indicate the multi-ensemble average, and thin lines indicate each member. **b,** T sensitivity calculated with 70-year time windows advanced from the early historical period (1867-1937) to the future (2017-2087) with 10-year intervals. **c,** Multi-member averaged temperature anomalies for 131 global river basins. **d**, Inter-member standard deviation of T sensitivity for each 70-year time window, calculated for 131 global river basins. The black vertical dashed line indicates the end year of the historical period. Note that the T sensitivity tends to converge as recent warming emerges, indicating that T sensitivity in the historical period reflects runoff sensitivity to climate change.



Supplementary Fig. 9| Temperature as a proxy for potential evapotranspiration when predicting the runoff changes. a,b, Inter-model correlation between runoff sensitivities using temperature (T) and potential evapotranspiration (PET) calculated for (**a**) historical and (**b**) future period. The black dashed line indicates statistically significant correlation coefficients based on a *t*-test at a 95% confidence interval, and the number of basins with significant correlations are indicated in the legend. **c,d,** Scatter plots between multi-model median simulated runoff projections and predictions using (**c**) historical and (**d**) future runoff sensitivities. The number of basins with significant predictions are indicated in the legend following the method applied in Fig. 2b. Note that 25 CMIP6 models are used for this analysis according to data availability related to PET.



Supplementary Fig. 10| Consistent results regardless of the regression methods. a,b, Same to Fig. 2b and Fig. 4b with runoff sensitivity estimated using the ordinary least square regression method. **c,d,** same to (a,b) but with runoff sensitivity estimated from the orthogonal regression method.



Supplementary Fig. 11| Impact of non-stationarity and internal climate variability on the runoff sensitivity. Inter-model R^2 between historical runoff sensitivity and various combinations of the runoff sensitivity for each basin, aggregated across 131 basins to a typical box and whisker plot.

Reference

1. Gou, J. *et al.* CNRD v1.0: A High-Quality Natural Runoff Dataset for Hydrological and Climate Studies in China. *Bulletin of the American Meteorological Society* **102**, E929–E947 (2021).