Supplementary Materials for

Historical and future changes in global flood magnitude – Evidence from a model-observation investigation

Hong Xuan Do[(1)(2)(3)(*)], Fang Zhao[(4)(5)(*)], Seth Westra[(1)], Michael Leonard[(1)], Lukas Gudmundsson[(6)], Jinfeng Chang[(7)], Philippe Ciais[(7)], Dieter Gerten[(5)(8)], Simon N. Gosling[(9)], Hannes Müller Schmied[(10)(11)], Tobias Stacke[(12)], Boulange Julien Eric Stanislas[(13)], Yoshihide Wada[(14)].

(1) School of Civil, Environmental and Mining Engineering, University of Adelaide, Adelaide, Australia.

(2) Faculty of Environment and Natural Resources, Nong Lam University, Ho Chi Minh City, Vietnam.

(3) School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan, United States.

(4) School of Geographical Sciences, East China Normal University, Shanghai, China.

(5) Potsdam Institute for Climate Impact Research, Potsdam, Germany.

(6) Institute for Atmospheric and Climate Science, Department of Environmental Systems Science, ETH Zurich, Zurich, Switzerland.

(7) Laboratoire des Sciences du Climat et de l'Environnement, CEA-CNRS-UVSQ/IPSL, Université Paris Saclay, 91191 Gif sur Yvette, France.

(8) Geography Dept., Humboldt-Universität zu Berlin, Berlin, Germany.

(9) School of Geography, University of Nottingham, Nottingham, United Kingdom.

(10) Institute of Physical Geography, Goethe University Frankfurt, Frankfurt am Main, Germany.

(11) Senckenberg Leibnitz Biodiversity and Climate Research Centre (SBiK-F), Frankfurt am Main, Germany.

(12) Max Planck Institute for Meteorology, Hamburg, Germany.

(13) Center for Global Environmental Research, Japan.

(14) International Institute for Applied Systems Analysis, Laxenburg, Austria.

[(*)] *Corresponding authors:* Hong Xuan Do (hong.do@adelaide.edu.au) *and* Fang Zhao (fangzhao@pik-potsdam.de)

Contents:

1 Simulation information

This section summarises the key simulation settings of each global hydrological model (GHM). Note that more detailed information is available in the protocols of the Inter-Sectoral Impact Model Intercomparison Project (ISIMIP) available at https://www.isimip.org/protocol.

The following two input datasets were used for the GHM simulations, with specific model runs summarised in Table S1:

1. **Climate & CO_2 concentration scenarios (i.e. atmospheric forcing)**
   - GSWP3: observations-based dataset providing the climate forcing data.
   - RCP2.6: future climate and CO_2 concentration from RCP2.6
   - RCP6.0: future climate and CO_2 concentration from RCP6.0
   - HINDCAST: historical modelled climate and CO_2 concentration.
2. **Human influence and land-use scenarios**
   - nosoc: Naturalized runs (no human impact). No irrigation. No population and GDP data prescribed.
   - varsoc: Varying historical land use and other human influences over historical period.
   - 2005soc: Fixed year-2005 land use and other human influences.

Note that GSWP3 was used as the sole observational atmospheric forcing dataset in this investigation. We also used modelled atmospheric forcing datasets introduced by four global climate models (GCM): GFDL-ESM2M, HadGEM2-ES, IPSL-CM5A-LR and MIROC5.

Table S1. Simulation set up of GHMs used in this investigation. 'Climate' represents atmospheric forcing dataset while 'human' represents human influence and land-use scenarios. Note that a more detailed inventory of available model runs is provided in Table S2.

<table>
<tr><th>Model</th><th>GSWP3_VARSOC</th><th>GSWP3_NOSOC</th><th>GCMHIND</th><th>GCMRCP2.6</th><th>GCMRCP6.0</th></tr>
<tr><td>H08</td><td rowspan="4">Climate: GSWP3
Human: varsoc</td><td rowspan="6">Climate: GSWP3
Human: nosoc</td><td>Climate: HINDCAST
Human: 2005soc</td><td rowspan="6">Climate: rcp26
Human: 2005soc</td><td rowspan="6">Climate: RCP6.0
Human: 2005soc</td></tr>
<tr><td>LPJmL</td><td rowspan="5">Climate: HINDCAST
Human: varsoc (except for ORCHIDEE using nosoc)</td></tr>
<tr><td>PCR-GLOBWB</td></tr>
<tr><td>WaterGAP2</td></tr>
<tr><td>MPI-HM</td><td rowspan="2">Simulations not available</td></tr>
<tr><td>ORCHIDEE</td></tr>
</table>

The results of preliminary assessment over 3666 observation locations suggest minor influence of human influence and land-use scenarios on the characteristics of trends in streamflow extremes (see section 4 of this supplementary material), and thus only GSWP3_NOSOC was used in the main text (denoted as GSWP3 in the main text).

2 Simulated streamflow extraction

For very large catchments, where excess rainfall takes a significant amount of time to reach the outlet, the routing scheme plays an important role in model performance related to high flow events (Zhao et al., 2017) and thus routed discharge is the more appropriate measure of simulated streamflow. The same simulation product, however, potentially does not perform well for small catchments, partially due to the coarse resolution of GHMs (Hunger and Döll, 2008). To address this concern, we adopted a common threshold of 9,000km^2 (approximate the size of 1°×1 ° grid cell) to separate the selected catchments into two groups and applied different procedures to extract simulated streamflow.

2.1 Weighted-area average for stations with catchment from 0 to 9000 km^2

2.1.1 Producing weighted-area tables

For stations with catchment area less than or equal to 9000 km^2, the catchment boundary was superimposed to the ISIMIP grid to identify intersecting cells, and a weighted-area table was calculated for each case. Simulated runoff was extracted by averaging the un-routed surface runoff from all intersect cells (considering weight). Runoff was then converted into discharge data.

Figure S1 provides an illustration of the weighted-area table for station US_0002282 (red dot; Merrill catchment of Pascagoula River, Mississippi, US) which has the total number of 15 upstream cells (dark-grey cells). Two components of the weighted-area table were used to label intersect cells: (1) cell number (dark red) and (2) normalised fraction of each cell (weights) that is covered by the catchment boundary (dark blue). The normalisation was performed such that the weights add up to one for each catchment, and these weights are used to extract simulated runoff for this catchment.

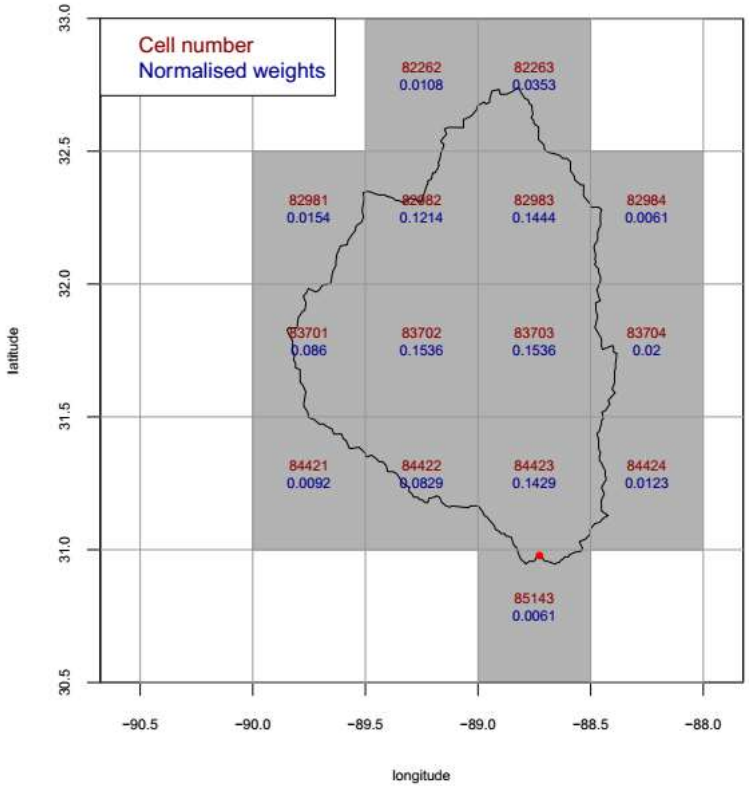
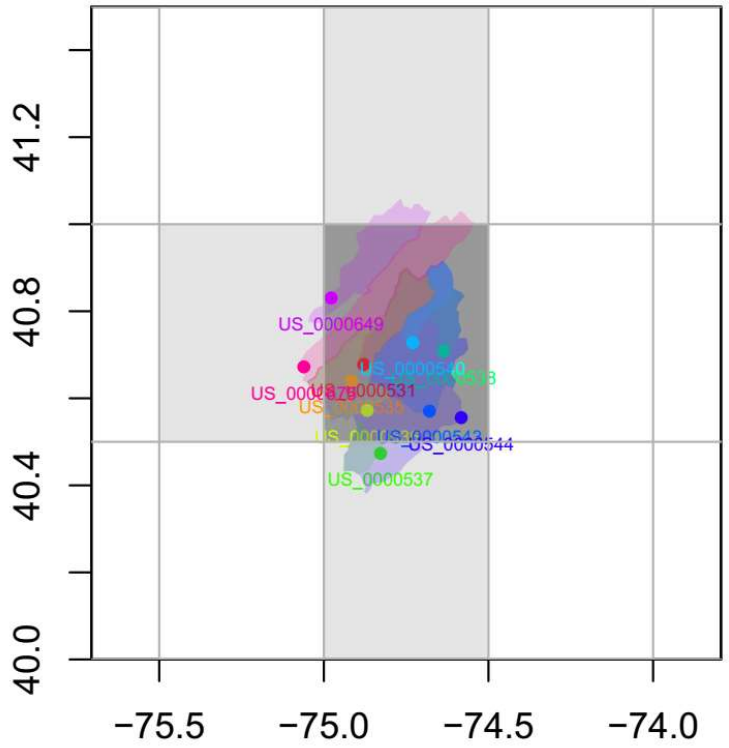


Figure S1. Illustration of the table of weights.

2.1.2 Averaging approach for cases where there were more than one catchment sharing similar weighted-area tables

Among catchments that have area less than 9000km^2, there are many instances where two or more catchments have (almost) identical simulated runoff as they have similar weighted-area tables. All ISIMIP models have a common assumption of uniform parameterisation for runoff generation in the 0.5×0.5 grid area, which in concept should represent an average value of runoff at finer resolution. Note that ORCHIDEE in ISIMIP2b (GCMs driven) was run at 1°×1° resolution, and the outputs were disaggregated evenly 0.5×0.5 resolution. Here we also treat catchments that intersect an identical set of dominant contributing grid-cells (total weights of at least 70%) as samples of an identical simulation domain. As a result, the area-weighted mean discharge of these catchments was calculated and used for model-observation comparison.

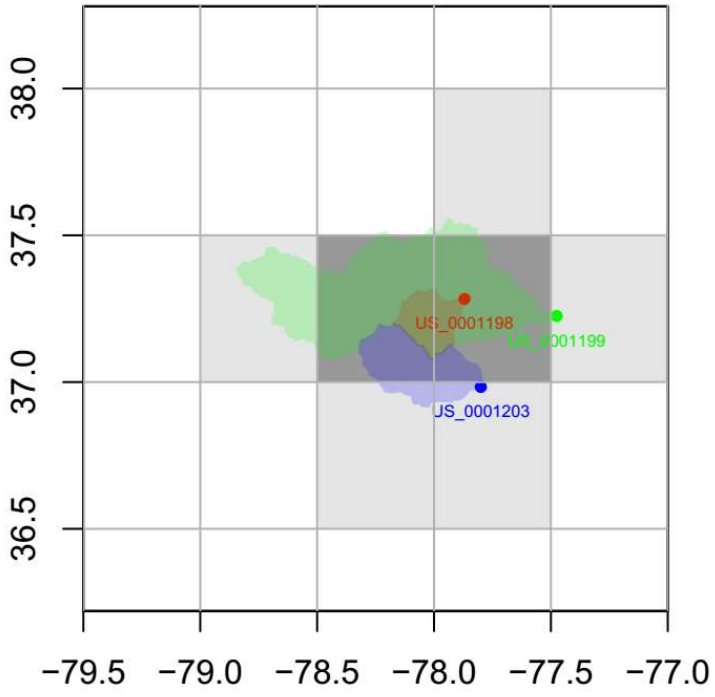
A search was conducted across all weighted-area tables to identify cases that have an identical set of intersecting cells contributing at least 70% to the total weighting. Figure S2 provides an example of these cases. In the top panel, boundaries of ten catchments were superimposed on top of the ISIMIP gridline (0.5×0.5 degree), demonstrating that they share a common cell (number 70051) which contributes at least 70% to the total weight (showed in the bottom panel).



	US_0000531	US_0000535	US_0000536	US_0000537	US_0000538	US_0000540	US_0000543	US_0000544	US_0000649	US_0000675
70051	1	0.75	0.947	1	1	1	1	0.86	0.818	0.933
71491	0	0.25	0.053	0	0	0	0	0.018	0	0.067
70770	0	0	0	0	0	0	0	0.123	0	0
70771	0	0	0	0	0	0	0	0	0.182	0

Figure S2. Example of instances where there is a significant overlap in contributing cells. Top panel: locations of 10 catchments that share a common contributing grid-cell (cell number 70051 (in dark-grey colour) contributes at least 70% to the total weight of each catchment) although specific catchments have different contributing cells. Bottom panel: weighted-area table of these 10 catchments.

Figure S3 illustrates another case where three different catchments share two common cells (no. 76524 and 76525). These cells contribute 100%, 79.1%, and 76.4% to the weighted-area tables of catchment US_0001198, US_0001199, and US_0001203 respectively. In both examples, the identified catchments were considered samples of the same modeling domain.



	US_0001198	US_0001199	US_0001203
76525	0.562	0.455	0.529
76524	0.438	0.336	0.235
75806	0	0.014	0
75803	0	0.189	0
75085	0	0.007	0
75805	0	0	0.147
75804	0	0	0.088

Figure S3. Similar to Figure S2, but here we have two contributing cells. The total weight of these common cells (number 76524 and 76525, highlighted in dark-grey colour) is higher than 0.7 in all cases and thus these three catchments were considered samples of the same modelling domain.

For each set of *n* catchments with similar weighted-area tables, a single average discharge $\bar{Q}$(m^3/s) was calculated to represent these individual time series in the model-observation comparison following below procedures:

For observed discharge:

1. Convert discharge Q (units: m^3/s) to runoff rate R (units: m/day) using catchment area A (units: m^2) for each catchment *i*.
$$R_i = Q_i \times 24 \times 3600/A_i \quad \text{(m/day)}$$
Average catchment size was also recorded:
$$\bar{A} = \frac{1}{n}\sum_{i=1}^{n} A_i \quad \text{(m}^2\text{)}$$
2. Average runoff rate across all catchments (considering area-weights)
$$\bar{R} = \frac{\sum_{i=1}^{n} R_i A_i}{\sum_{i=1}^{n} A_i} \quad \text{(m/day)}$$
3. Back-calculate average discharge (m^3/s):
$$\bar{Q} = \frac{\bar{R}\bar{A}}{24\times3600} \quad \text{(m}^3\text{/s)}$$

For simulated discharge:

1. Extract runoff rate using weighted-area tables as described in Section 2.1 for all catchments.
2. Follow Step 2 and Step 3 of the observation procedure.

2.3 Discharge output identification for catchment with area greater than 9000 km^2

For catchments with area greater than 9000km^2, the 'discharge output' approach was adopted to find GHM cells corresponding to the catchments following Zhao et al. (2017). For a specific catchment, the grid cell corresponding to the catchment outlet was identified by matching catchment area available in a 0.5° drainage direction map (DDM30 dataset, freely available at http://www.uni-frankfurt.de/45218101/DDM30) and the reported area. The identified grid cell was then used to extract simulated discharge available in the ISIMIP data repository. Stations were removed if the procedure could not identify any DDM30 grid cell surrounding the reported geographical location with a drainage area discrepancy less than 30% (see supplementary of Zhao et al. (2017) for detail).

3 Supplementary Figures

3.1 Capacity of GHMs to reproduce observed trends at continental scale

As stream gauges are not evenly distributed across the world, Figure S4 provides a zoomed-in map for four regions with relatively high number of stations (North America, Europe, South America, and Oceania). The most notable feature is a significantly lower strength of trends exhibited through GSWP3/GCMHIND ensemble average compared to GSIM observed trends. This pattern is likely the result of averaging technique (smoothed out variability of ensemble members) as the feature is more pronounced in GCMHIND (21 simulations) compared to GSWP3 (6 simulations). Visual inspection of these results suggests that the overall spatial pattern of observed trends seems to be preserved in GSWP3 while GCMHIND simulations tend to incorrectly simulate some spatial pattern of trends (e.g. over Oceania).

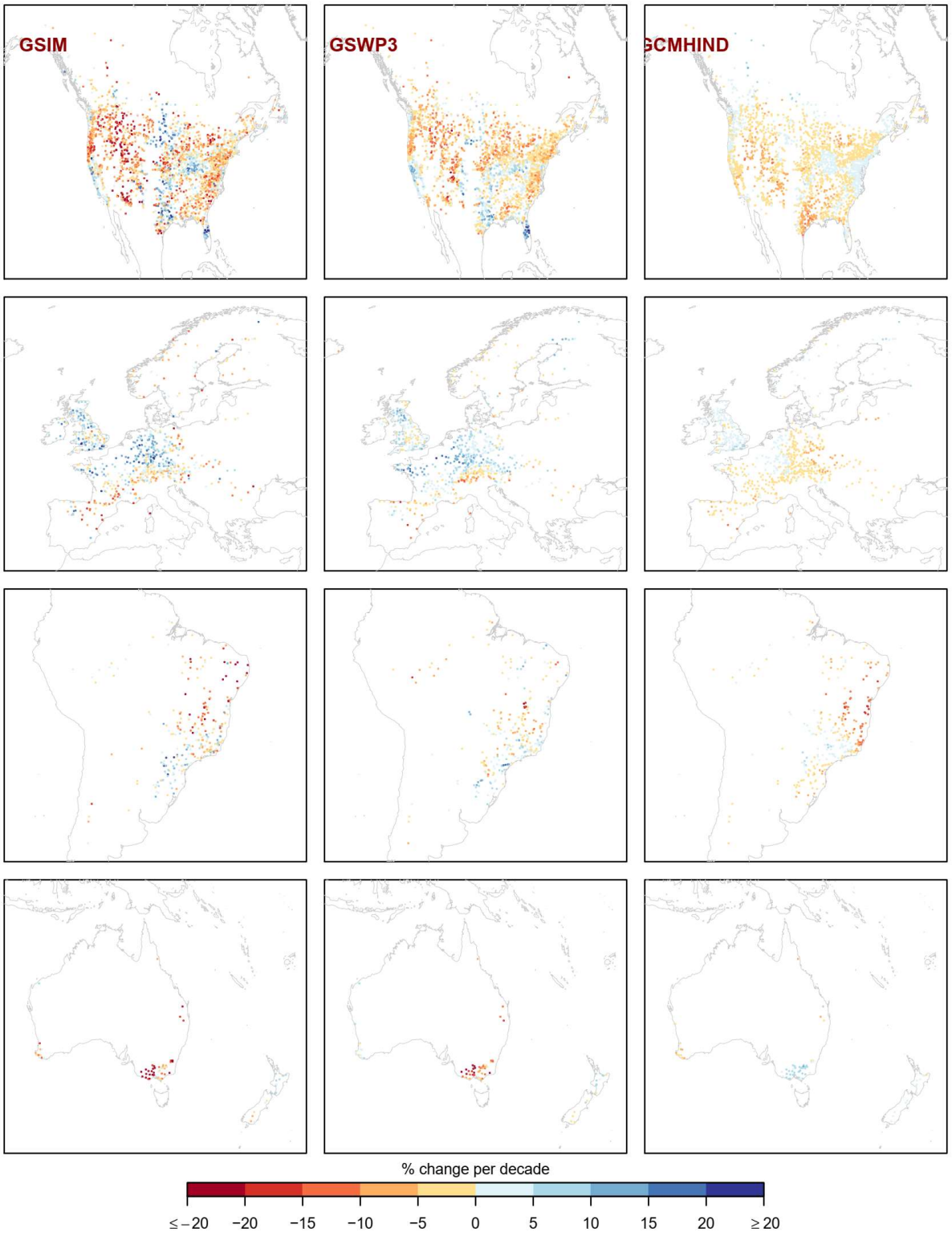


Figure S4. Normalised Theil-Sen slope for historical trends in flood magnitude (MAX7 index) over South America, Europe, South America and Oceania (left panels: GSIM; middle panels: GSWP3; right panels: GCMHIND). Multi-model average is shown for simulated trends. Trend is expressed in % change per decade.

Figure S5 illustrates the mean and standard deviation of simulated trends across all locations (% change per decade) for each individual ensemble member (multi-model average was showed in the manuscript). The mean and standard deviation of all trends (referred to as trend mean and trend standard deviation here-after) obtained from GSIM archive were also showed as dark blue line. GSWP3 simulations generally produced a higher trend mean and a lower trend standard deviation across all continents compared to the observed trends. The discrepancy varies substantially across different regions. For instance, Oceania exhibited a discrepancy up to 7% per decade for the trend mean and 8% per decade for the trend standard deviation.

This feature indicates a substantial inconsistency between simulated trends and observed trends. Among the six GHMs, ORCHIDEE, PCR-GLOBWB and WaterGAP tend to have a higher trend mean with the exception of Africa. This pattern potentially indicates the influence of either (i) parameterisation, (ii) model capacity in reproducing observed trend characteristics, or (iiI) a bias of the GSWP3 forcing trends.

Figure S5 also shows relatively lower capacity of GCMHIND simulation in terms of reproducing observed trend mean and trend standard deviation in streamflow maxima. There is no clear ranking pattern in terms of the modelled atmospheric forcing being used, suggesting that uncertainty in GCM model was inherited differently across GHMs, likely due to the variation of parameterisation strategies.

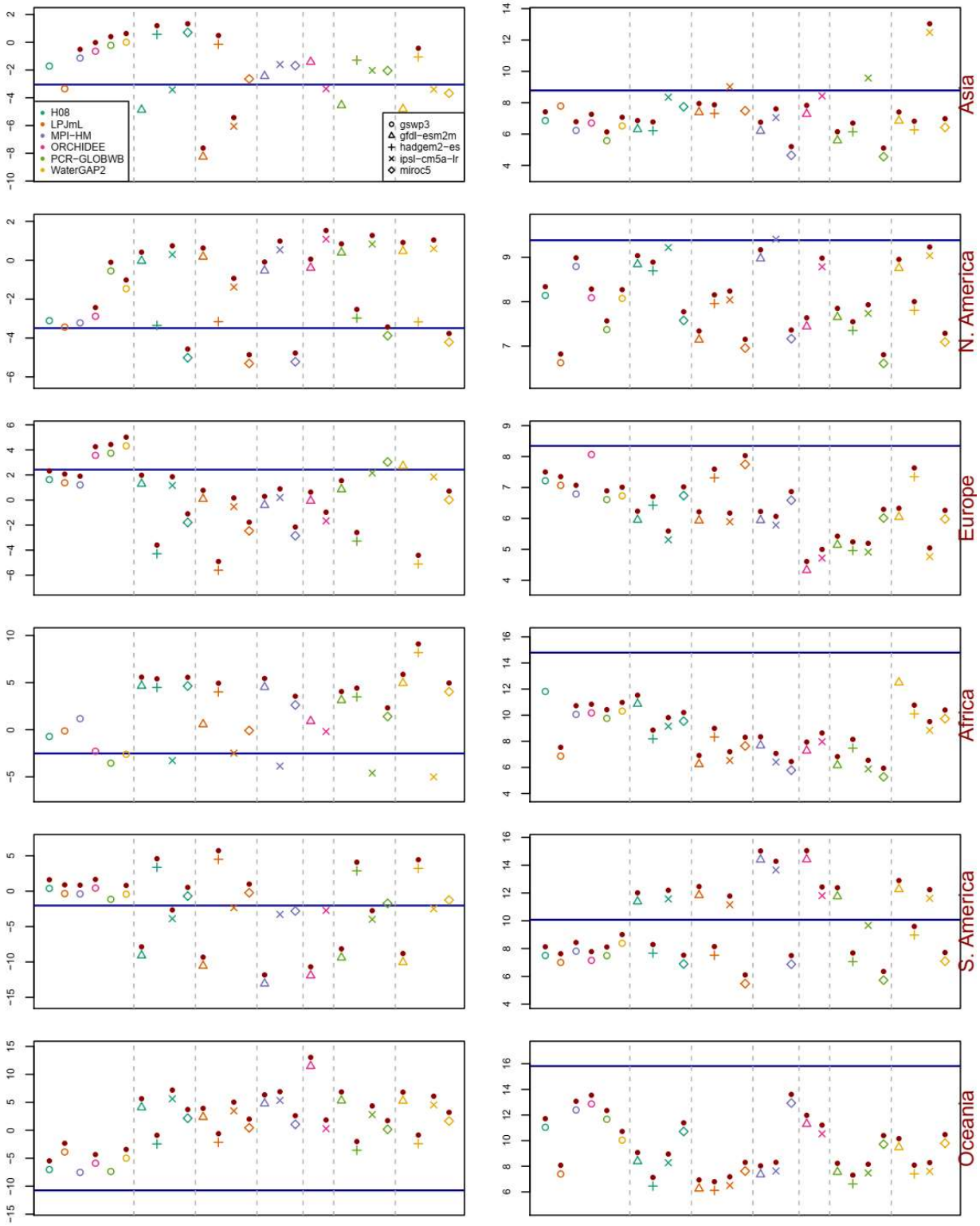


Figure S5. Mean (left panels) and standard deviation (right panels) of trends (% change per decade) exhibited from GSIM (horizontal blue line) observed trends and GSWP3/GCMHIND (hollow dots) simulated trends at the continental scale. The x-axis indicates different models. Note that y-axis range varies across panels. A null-hypothesis test was conducted to assess whether the mean/standard deviation of simulated trends is statistically different to that obtained from observed GSIM trends (horizontal blue line). Dark-red filled dots indicate simulations rejecting the null-hypothesis (i.e. which is that simulated trend mean/trend standard deviation is not statistically different to that obtained from GSIM).

3.2 Spatial uncertainty across simulated trends forced with different modelled atmospheric forcing

The assessment in section 3.3 of the main text suggests the combined GCM-GHM uncertainty has led to the presence of high uncertainty in terms of regions with significant projected trends in streamflow extremes. That is, a region could be projected by an overall increasing trend by one member and a decreasing trend by another member. This feature is illustrated in Figure S6, which shows a notable mismatch in the spatial structure of projected trends in MAX7 index between two ensemble members. Under the RCP2.6 greenhouse gas emission scenario, H08 forced with GFDL-ESM2M (top panels) projects an increasing trend for the majority of Australia and Siberia, while ORCHIDEE forced with IPSL-CM5A-LR (bottom panels) projects an overall decreasing trend for the same regions. This spatial uncertainty could come from either the climate trends introduced by GCMs (differentiate across GCMs), different RCPs, and model characteristics.

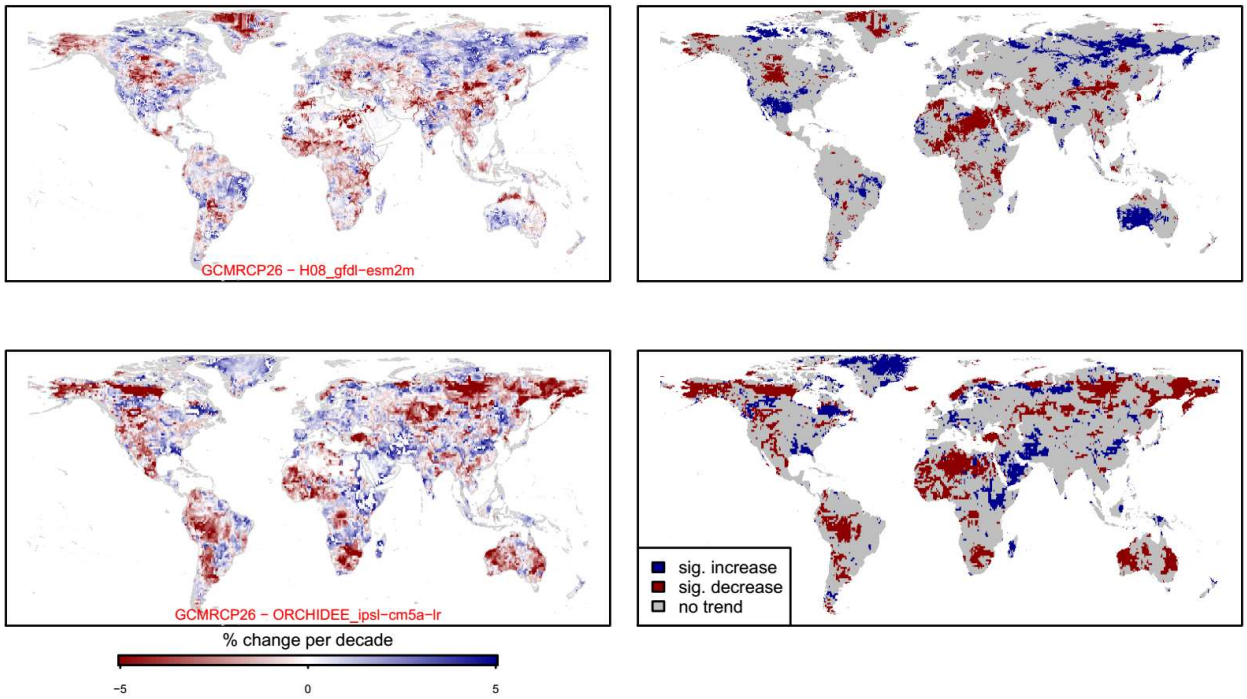


Figure S6. The magnitude (left panels) and significance (right panels) of trends in simulated MAX7 time series across all grid cells under RCP26 greenhouse gas emission scenario (2006-2099). Top panels: H08 forced with gfdl-esm2m climate data; bottom panels: ORCHIDEE forced with ipsl-cm5a-lr climate data. These two models had the lowest value of pattern similarity (correlation of -0.17).

4 Supplementary Tables

Considering a large number of simulations available (73 in total), the main text mostly used multi-model min/max/average to illustrate the results for cases where there is more than one simulation available for an identical GHM/spatial-domain. Table S2 provides a list of all 73 available models reported in this section together, with their simulation settings. Note that:

(i) GSWP3_VARSOC simulations (listed in Table S2 as H08_GSWVAR, LPJ_GSWVAR, PCR_GSWVAR, and WAT_GSWVAR) were not reported in the main text as (1) there were only four simulations available (comparing to six simulations of GSWP3_NOSOC) and (2) the results obtained from GSWP3_NOSOC and GSWP3_VARSOC are similar (Table S3).

(ii) In the main text, OBSHIS_NOSOC simulations were denoted as GSWP3.

Table S2. Available ISIMIP streamflow simulations and associated setting.

Seq	Streamflow simulations	GHM	Climate	Human	Period
1.	H08_GSWVAR	H08	Observation (GSWPv3)	varsoc	1971-2005
2.	H08_GSWNO		Observation (GSWPv3)	nosoc	
3.	H08_HIN_G		HINDCAST (GFDL-ESM2M)	2005soc	
4.	H08_HIN_H		HINDCAST (HadGEM2-ES)		
5.	H08_HIN_I		HINDCAST (IPSL-CM5A-LR)		
6.	H08_HIN_M		HINDCAST (MIROC5)		
7.	H08_RCP2.6_G		RCP2.6 (GFDL-ESM2M)		2006-2099
8.	H08_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
9.	H08_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
10.	H08_RCP2.6_M		RCP2.6 (MIROC5)		
11.	H08_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
12.	H08_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
13.	H08_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
14.	H08_RCP6.0_M		RCP6.0 (MIROC5)		
15.	LPJ_GSWVAR	LPJmL	Observation (GSWPv3)	varsoc	1971-2005
16.	LPJ_GSWNO		Observation (GSWPv3)	nosoc	
17.	LPJ_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
18.	LPJ_HIN_H		HINDCAST (HadGEM2-ES)		
19.	LPJ_HIN_I		HINDCAST (IPSL-CM5A-LR)		
20.	LPJ_HIN_M		HINDCAST (MIROC5)		
21.	LPJ_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
22.	LPJ_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
23.	LPJ_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
24.	LPJ_RCP2.6_M		RCP2.6 (MIROC5)		
25.	LPJ_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
26.	LPJ_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
27.	LPJ _RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
28.	LPJ_RCP6.0_M		RCP6.0 (MIROC5)		
29.	MPI_GSWNO	MPI-HM	Observation (GSWPv3)	nosoc	1971-2005
30.	MPI_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
31.	MPI_HIN_I		HINDCAST (IPSL-CM5A-LR)		
32.	MPI_HIN_M		HINDCAST (MIROC5)		
33.	MPI_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
34.	MPI_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
35.	MPI_RCP2.6_M		RCP2.6 (MIROC5)		
36.	MPI_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
37.	MPI_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
38.	MPI_RCP6.0_M		RCP6.0 (MIROC5)		
39.	ORC_GSWNO	ORCHIDEE	Observation (GSWPv3)	nosoc	1971-2005
40.	ORC_HIN_G		HINDCAST (GFDL-ESM2M)		
41.	ORC_HIN_I		HINDCAST (IPSL-CM5A-LR)		
42.	ORC_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	nosoc (land use changes was considered)	2006-2099
43.	ORC_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
44.	ORC_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
45.	ORC_RCP6.0_G		RCP6.0 (IPSL-CM5A-LR)		
46.	PCR_GSWVAR		Observation (GSWPv3)	varsoc	

47.	PCR_GSWNO	PCR-GLOBWB	Observation (GSWPv3)	nosoc	1971-2005
48.	PCR_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
49.	PCR_HIN_H		HINDCAST (HadGEM2-ES)		
50.	PCR_HIN_I		HINDCAST (IPSL-CM5A-LR)		
51.	PCR_HIN_M		HINDCAST (MIROC5)		
52.	PCR_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
53.	PCR_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
54.	PCR_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
55.	PCR_RCP2.6_M		RCP2.6 (MIROC5)		
56.	PCR_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
57.	PCR_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
58.	PCR_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
59.	PCR_RCP6.0_M		RCP6.0 (MIROC5)		
60.	WAT_GSWVAR	WaterGAP2	Observation (GSWPv3)	varsoc	1971-2005
61.	WAT_GSWNO		Observation (GSWPv3)	nosoc	
62.	WAT_HIN_G		HINDCAST (GFDL-ESM2M)	varsoc	
63.	WAT_HIN_H		HINDCAST (HadGEM2-ES)		
64.	WAT_HIN_I		HINDCAST (IPSL-CM5A-LR)		
65.	WAT_HIN_M		HINDCAST (MIROC5)		
66.	WAT_RCP2.6_G		RCP2.6 (GFDL-ESM2M)	2005soc	2006-2099
67.	WAT_RCP2.6_H		RCP2.6 (HadGEM2-ES)		
68.	WAT_RCP2.6_I		RCP2.6 (IPSL-CM5A-LR)		
69.	WAT_RCP2.6_M		RCP2.6 (MIROC5)		
70.	WAT_RCP6.0_G		RCP6.0 (GFDL-ESM2M)		
71.	WAT_RCP6.0_H		RCP6.0 (HadGEM2-ES)		
72.	WAT_RCP6.0_I		RCP6.0 (IPSL-CM5A-LR)		
73.	WAT_RCP6.0_M		RCP6.0 (MIROC5)		

Most results of the main text only showed the multi-model average for GCMHIND simulations of each GHM (up to four simulations per GHM) (e.g. Table 3 of the main text, which presents the characteristics of trends in the MAX7 index over 1971-2005 period across 3666 locations globally). The following tables, therefore, provide the results of each experiment at the global scale for individual models to complement the key findings, in which:

- Table S3 (adapted from Table 2 in the main text) describe the hypothesis tests.
- Table S4 and S5 report trend mean/standard deviation, percentage of locations exhibiting significant trends and the correlation of simulated trends against observed trends (historical period from 1971 to 2005). The results of hypothesis test (described in Table S3) are also highlighted in Table S4 and Table S5.
- Tables S6 and S7 report the value of simulated trend mean/trend standard deviation and the percentage of cells exhibiting significant trends for future period (2006-2099). Note that the statistical test described in Table S3 was not adopted for these results.

As noted in the main text, trends in peak discharge exhibited from 'naturalised runs' (GSWP3_NOSOC) are similar to those obtained from 'human impact runs' (GSWP3_VARSOC). This is specifically illustrated through Table S4, in which the trends characteristic are quite similar between two settings. For instance, PCR_GSWVAR suggests a global trend mean (standard deviation) of 0.0 (7.7) % change per decade, with a spatial correlation against observed trends of 0.5. These results are very similar to that reported for PCR_GSWNO.

Table S3. Summary of the hypothesis tests conducted to address the first two objectives. The significance of these tests was reported in Table S4 and S5.

Objective	Null-Hypotheses	Streamflow dataset	Statistical tests
Objective 1: Capacity of GHMs to reproduce observed trends in flood hazards	Hypothesis 1: Trend means obtained from two streamflow datasets over observation locations were not statistically different from each other.	(i) Observed discharge across 3,666 observation locations (ii) Simulated discharge across 3,666 observation locations (extraction processes outlined in Section 2)	Two-sample *t*-test at the 10% two-sided significance level
	Hypothesis 2: Trend standard deviations obtained from two streamflow datasets over observation locations were not statistically different from each other.		Two-variance *F*-test at the 10% two-sided significance level
	Hypothesis 3: Percentage of significant trends obtained from all observation locations of a specific streamflow dataset was not produced by random chance.		Field significance test similar to that presented in Do et al. (2017) was adopted. A moving-block-bootstrap (block-length $L = 2$) was used to derive a null-hypothesis distribution of the change that occurred due to random chance. The null hypothesis is rejected at 5% one-sided significance level when the true percentage falls on the right-hand side of the 95th percentile of the resampled distributions.
	Hypothesis 4: The correlation between trends obtained from two streamflow datasets was not significantly higher than '0' (i.e. zero pattern similarity).		'Zero pattern similarity' was compared to the probability distribution function (PDF) of pairwise correlation between simulated and observed trends, drawn from a bootstrap procedure similar to that proposed by Kiktev et al. (2003). The null hypothesis is rejected at 5% one-sided significance level when zero correlation falls on the left-hand side of the 5th percentile of the resampled distributions.
	Hypothesis 5: The correlation between GCMHIND simulated trends and observed trends was not significantly lower than the correlation between GSWP3 simulated trends and observed trends		The actual pairwise correlation between GCMHIND simulated trends and observed trends (denoted by $r_{GCMHIND}$) was compared to the bootstrapped PDF of correlation exhibited from GSWP3 simulated trends (denoted by r^*_{GSWP3}). If $r_{GCMHIND}$ falls on the left-hand side of the 5th percentile r^*_{GSWP3}, there is evidence to reject the null-hypothesis at the 5% one-sided significance level.
Objective 2: The representativeness of	Hypothesis 6: Trend mean obtained from observation locations was not statistically different to that obtained from all grid cells.	(i) Simulated discharge across 3,666 observation locations (extraction	Two-sample *t*-test at the 10% two-sided significance level

observation locations in the GHM simulations	Hypothesis 7: Trend standard deviation obtained from observation locations was not statistically different to that obtained from all grid cells.	processes outlined in Section 2) (ii) Routed discharge across all landmass grid cells (59,033 cells)	Two-variance *F*-test at the 10% two-sided significance level
	Hypothesis 8: Percentage of significant trends obtained from all grid cells of a specific streamflow dataset was not produced by random chance.		Field significance test similar to that presented in Hypothesis 3 but trends obtained from all grid cells were the subject of the assessment.

Table S4. Characteristics of trends in the MAX7 index (introduced by GHMs) over the 1971-2005 period averaged across the 3666 locations. Trend mean and trend standard deviation have units of %-change per decade. Gauge-based significant trends were identified using a Mann-Kendall test (10% two-sided significance level). The global significance of this result is then calculated using field significance test (5% one-sided significant level; highlighted in boldface text). Trend mean, trend standard deviation and trend spatial structure were compared against that exhibited by GSIM (see Hypothesis 1 to hypothesis 5 of Table S3 for description of hypothesis tests; significant values were represented in boldface text).

Streamflow simulations	**Trend mean**	**Trend standard deviation**	**Percentages of significant**		**Correlation against observed trends**
			Increasing trend	**Decreasing trend**	
H08_GSWVAR	-2.0	**8.3**	4.8	6.7	**0.4**
LPJ_GSWVAR	-2.6	**7.5**	4.6	9.2	**0.4**
PCR_GSWVAR	**0.0**	**7.7**	9.4	6.1	**0.5**
WAT_GSWVAR	**-0.7**	**8.5**	8.4	5.8	**0.5**
H08_GSWNO	**-1.9**	**8.3**	4.8	6.7	**0.4**
LPJ_GSWNO	-2.2	**7.1**	4.5	7.3	**0.4**
ORC_GSWNO	**-1.4**	**8.6**	7	8.2	**0.4**
MPI_GSWNO	-2.1	**8.7**	5.6	7.5	**0.5**
PCR_GSWNO	**0.1**	**7.7**	9.6	6.1	**0.5**
WAT_GSWNO	**-0.3**	**8.2**	8.5	4.2	**0.5**
H08_HIN_G	**-0.4**	**8.9**	6.1	7.8	**0.1**
H08_HIN_H	**-2.8**	**8.4**	2.2	10.8	-0.1
H08_HIN_I	**0.1**	**8.9**	7.7	4.4	0.0
H08_HIN_M	**-3.6**	**7.8**	3.4	12.0	**0.1**
LPJ_HIN_G	**-0.8**	**8.0**	6.3	8.3	**0.1**
LPJ_HIN_H	**-2.9**	**8.1**	2.8	**14.6**	0.0
LPJ_HIN_I	**-1.3**	**8.0**	4.1	10.1	**0.1**
LPJ_HIN_M	**-4.1**	**7.3**	3.5	**17.3**	**0.2**
ORC_HIN_G	**-0.9**	**8.6**	5.2	7.6	0.0
ORC_HIN_I	**0.1**	**8.6**	8.6	6.4	**0.1**
MPI_HIN_G	**-1.3**	**9.5**	5.9	7.9	**0.1**
MPI_HIN_I	**0.2**	**9.2**	8.8	5.6	0.0
MPI_HIN_M	**-4.2**	**7.3**	2.3	**16.3**	**0.1**
PCR_HIN_G	**-0.2**	**8.0**	8.3	9.0	**0.1**
PCR_HIN_H	-2.5	**7.1**	2.7	11.0	0.0
PCR_HIN_I	**0.6**	**7.6**	12.2	4.1	0.0
PCR_HIN_M	-2.1	**7.0**	6.9	13.5	**0.1**
WAT_HIN_G	**0.2**	**9.2**	8.2	5.6	**0.1**
WAT_HIN_H	**-2.9**	**8.1**	2.7	10.9	-0.1
WAT_HIN_I	**0.5**	**8.8**	6.2	4.2	-0.1
WAT_HIN_M	**-2.9**	**7.3**	4.3	11.4	**0.1**

Table S5. Trend mean, trend standard deviation and percentage of significant trends averaged across all simulation grid cells. Trend mean and trend standard deviation have units of %-change per decade. Cell-based significance was identified using the Mann-Kendall test at the 10% significance level. The global significance of this result is then calculated using field significance test at 5% one-sided level (highlighted in boldface text). Trend mean and trend standard deviation across all land mass were compared against that obtained across 3666 observation locations (reported in Table S4) and significant values are highlighted in boldface text (see Hypothesis 6 to hypothesis 8 of Table S3 for description of hypothesis tests).

Streamflow simulations	**Trend mean**	**Trend standard deviation**	**Percentages of significant**	
			Increasing trend	**Decreasing trend**
H08_GSWVAR	**-0.5**	**10.1**	8.4	**10.7**
LPJ_GSWVAR	**-1.6**	**10.4**	7.2	**14.0**
PCR_GSWVAR	**-1.1**	**11.0**	10.4	**15.0**
WAT_GSWVAR	**-0.3**	**11.4**	10.8	**11.0**
H08_GSWNO	**-0.3**	**9.9**	8.3	9.6
LPJ_GSWNO	**-0.9**	**9.9**	7.4	**11.5**
ORC_GSWNO	**-0.9**	**9.6**	6.1	**7.8**
MPI_GSWNO	**-0.7**	**10.2**	6.4	**7.5**
PCR_GSWNO	**-1.0**	**10.9**	10.7	**14.7**
WAT_GSWNO	**0.0**	**11.1**	**10.9**	10.1
H08_HIN_G	**1.5**	**10.8**	**15.4**	10.4
H08_HIN_H	**0.0**	**8.5**	7.4	9
H08_HIN_I	**-0.7**	**9.3**	7	**10.7**
H08_HIN_M	**0.4**	**8.9**	8.7	8
LPJ_HIN_G	**-0.3**	**9.3**	8.9	9.1
LPJ_HIN_H	**-1.1**	**8.7**	5.1	**9.9**
LPJ_HIN_I	-1.1	**8.7**	6.1	9.2
LPJ_HIN_M	**-0.8**	**9.1**	7.7	9.4
ORC_HIN_G	**0.6**	9.5	**8.4**	6.3
ORC_HIN_I	**-0.9**	**8.2**	3.9	**6.8**
MPI_HIN_G	**-0.1**	7.3	4.5	5
MPI_HIN_I	**-0.2**	**10.3**	**10.9**	**11.2**
MPI_HIN_M	**-1.4**	**9.3**	5.5	**11.1**
PCR_HIN_G	**1.3**	**11.3**	**14.9**	11.1
PCR_HIN_H	**-0.4**	**8.7**	8.1	10.5
PCR_HIN_I	**-1.3**	**10.7**	7.7	**12.2**
PCR_HIN_M	**0.4**	**9**	11.7	9.9
WAT_HIN_G	**1.5**	**10.9**	**15.3**	7.2
WAT_HIN_H	**0.0**	**9.1**	6.3	7.3
WAT_HIN_I	**0.0**	**9.4**	6.9	7.5
WAT_HIN_M	**0.4**	**9.7**	**10.8**	7.2

Table S6. Characteristics of projected trends (GCMRCP2.6) across 18 members at the global scale. Mean and standard deviation have unit of %-change per decade. Note that no statistical test was conducted.

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant	
			Increasing trend	Decreasing trend
H08_RCP2.6_G	0.0	2.1	10.9	9.6
H08_RCP2.6_H	0.4	2.7	18.0	11.0
H08_RCP2.6_I	0.0	2.3	11.5	14.2
H08_RCP2.6_M	0.0	2.8	16.2	11.6
LPJ_RCP2.6_G	-0.1	1.8	7.5	7.4
LPJ_RCP2.6_H	0.0	2.1	10.7	10.6
LPJ_RCP2.6_I	-0.1	2.1	9.1	10.6
LPJ_RCP2.6_M	0.0	2.2	12.6	9.0
ORC_RCP2.6_G	-0.3	2.3	9.0	13.9
ORC_RCP2.6_I	-0.6	2.9	9.2	21.2
PCR_RCP2.6_G	0.1	2.1	11.0	9.0
PCR_RCP2.6_H	0.3	2.3	16.6	11.2
PCR_RCP2.6_I	0.0	2.8	15.5	13.9
PCR_RCP2.6_M	0.1	2.5	17.4	12.4
WAT_RCP2.6_G	0.0	2.1	9.6	7.1
WAT_RCP2.6_H	0.4	2.2	14.1	7.5
WAT_RCP2.6_I	0.2	2.3	12.3	10.0
WAT_RCP2.6_M	0.2	2.4	16.1	7.3

Table S7. Characteristics of projected trend (GCMRCP6.0) across 18 members at the global scale. Trend mean and trend standard deviation have unit of %-change per decade. Note that no statistical test was conducted.

Streamflow simulations	Trend mean	Trend standard deviation	Percentages of significant	
			Increasing trend	Decreasing trend
H08_RCP6.0_G	0.3	3.0	19.7	17.1
H08_RCP6.0_H	0.7	4.0	27.2	18
H08_RCP6.0_I	-0.4	3.4	15.3	27.1
H08_RCP6.0_M	0.4	3.3	26.2	14.9
LPJ_RCP6.0_G	-0.1	2.6	17.5	15.7
LPJ_RCP6.0_H	-0.2	3.4	22.3	21.9
LPJ_RCP6.0_I	-0.6	3.1	14.0	24.8
LPJ_RCP6.0_M	0.1	3.0	22.6	16.2
ORC_RCP6.0_G	-0.3	3.0	16.4	21.1
ORC_RCP6.0_I	-1.3	4.1	12.3	35.0
PCR_RCP6.0_G	-0.1	3.0	18.9	18.7
PCR_RCP6.0_H	0.1	3.8	26.0	22.2
PCR_RCP6.0_I	-0.5	3.6	18.3	25.6
PCR_RCP6.0_M	0.5	3.0	27.7	14.4
WAT_RCP6.0_G	0.4	2.6	23.5	9.8
WAT_RCP6.0_H	0.7	3.2	29.6	10.7
WAT_RCP6.0_I	0.0	3.2	20.4	16.9
WAT_RCP6.0_M	0.8	3.1	30.1	9.6

Reference

Do, H. X., Westra, S., and Leonard, M.: A global-scale investigation of trends in annual maximum streamflow, Journal of Hydrology, 10.1016/j.jhydrol.2017.06.015, 2017.

Hunger, M., and Döll, P.: Value of river discharge data for global-scale hydrological modeling, Hydrology and Earth System Sciences Discussions, 12, 841-861, 2008.

Zhao, F., Veldkamp, T. I., Frieler, K., Schewe, J., Ostberg, S., Willner, S., Schauberger, B., Gosling, S. N., Schmied, H. M., and Portmann, F. T.: The critical role of the routing scheme in simulating peak river discharge in global hydrological models, Environmental Research Letters, 12, 075003, 2017.