Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

The importance of compounding drivers for large river floods

Shijie Jiang (shijie.jiang@ufz.de)
Helmholtz Centre for Environmental Research https://orcid.org/0000-0002-2808-9559
Larisa Tarasova
Helmholtz Centre for Environmental Research - UFZ https://orcid.org/0000-0002-7675-0751
Guo Yu
Desert Research Institute
Jakob Zscheischler
Helmholtz Centre for Environmental Research https://orcid.org/0000-0001-6045-1629

Physical Sciences - Article

Keywords:

Posted Date: March 27th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2733042/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: There is **NO** Competing Interest.

The importance of compounding drivers for large river floods

Shijie Jiang[1], Larisa Tarasova[2], Guo Yu[3], Jakob Zscheischler[1,4]

[1]Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research – UFZ, Leipzig, Germany

[2]Department Catchment Hydrology, Helmholtz Centre for Environmental Research – UFZ, Halle (Saale), Germany

[3]Division of Hydrologic Sciences, Desert Research Institute, Las Vegas, USA

[4]Technische Universität Dresden, Dresden, Germany

Abstract

Estimating river flood risks under global warming is challenging, largely because of the compounding nature of various drivers[1,2]. Yet to date, the interplay of multiple drivers and how they affect river floods are not well understood. Here we use explainable machine learning to disentangle the interactions between flood drivers and identify the compounding drivers of river floods in thousands of catchments around the world. We find that the majority of river floods worldwide over the past 40 years were attributable to compounding drivers, which often amplified river flood magnitude. Furthermore, the role of compounding drivers becomes more important with increasing flood magnitude in nearly all of the studied catchments, with the strength of this relationship generally depending on the catchment physio-climatic conditions. Based on these findings, we demonstrate that traditional statistical methods using flood frequency analysis underestimate the magnitude of extreme floods because compounding drivers are not properly taken into account[3]. Overall, our results highlight the need for careful incorporation of compounding drivers in flood risk assessment to improve estimates of extreme floods, in particular in the face of climate change[4,5].

Main

River floods are among the most common natural disasters and their risk is projected to increase further in the future due to climate and socioeconomic changes, although substantial uncertainties remain[1,6]. A key to improving flood prediction is to enhance the understanding of the mechanisms that lead to floods, especially those associated with extreme floods[7]. Generally, river floods can be generated by a variety of atmospheric processes (e.g., circulation patterns causing heavy precipitation and temperature increases causing snowmelt or glacial melt) that are modified by various catchment conditions and characteristics[2]. Intricate interactions of all these processes determine the timing, duration, extent, temporal clustering, and severity of river floods[8-11], which makes estimating future flood risks particularly challenging because flood drivers may exhibit varying trends in a changing climate[4,12].

Conventionally, river floods have been extensively studied using a process-oriented flood typology, which essentially strives to assign one dominant process to each event, even when multiple drivers may be present[7,13-18]. In many cases, however, the compounding of flood generation processes is given little targeted consideration despite its ubiquity and importance in contributing to extreme floods. Floods can be more severe if multiple processes and conditions that favor high river discharges (e.g., heavy and prolonged precipitation, melting of accumulated snowpack, and high antecedent soil moisture) occur concurrently[3]. The amplified impacts are often due to multiple drivers combining nonlinearly and complex interactions between different physical processes across various spatial and temporal scales[19]. Even for cases in which none of the drivers are extreme from a statistical perspective, their joint contribution can trigger extreme and unprecedented floods that pose an enormous challenge to flood risk management[20]. Therefore, a thorough understanding of compounding river flood drivers under historical conditions is critical to improving current flood risk strategies and developing new strategies in the future[5]. In principle, one can derive physically based hypotheses of different compounding drivers and how they affect flood magnitude, but data-based evidence to support such hypotheses is still lacking at a large scale[3]. Most of the previous studies of compound flood events focused on coastal floods and the interaction between storm surges and

runoff[21,22], whereas only a few studies explicitly examined compounding drivers of inland river flooding[18,23].

In this study, we developed a novel approach based on explainable machine learning (ML) to disentangle the multiple and compounding drivers of river floods. The approach allows for a unified assessment to quantify the contributions of meteorological drivers (rainfall and temperature), catchment preconditions (snow depth and soil moisture), and their interactions on river floods (defined as annual maximum discharges) in a wide variety of catchments. We applied the approach to 3,527 catchments worldwide and quantified the predictive contribution of interaction effects in the predictors for 124,642 annual discharge maxima from 1981 to 2020 (Extended Data Figs. 1-2, see **Methods**). We then identified which floods were associated with multiple drivers based on the interaction effects and examined the spatial distribution of multi-driver floods. We further investigated flood complexity in catchments with different physio-climatic conditions (Extended Data Fig. 3), which we defined as the slope in the number of relevant interaction effects between drivers along increasing flood magnitude. The flood complexity of a catchment, arising from the compounding effects between drivers, characterizes the heterogeneity in the physical processes that generate floods of different magnitudes and is related to the reliability of extreme flood hazard estimates.

Identification of compounding effects

We identified flood drivers using explainable ML that combines Light Gradient Boosting Machine (LightGBM)[24] for the prediction of runoff events and Shapley Additive exPlanation (SHAP) interaction values[25] to explain the results by revealing the contributions of input variables (Fig. 1a). We trained the ML models for each of the studied catchments to learn the nonlinear relationship between discharge peaks (regardless their magnitude) and their corresponding recent rainfall, recent temperature, antecedent soil moisture, and antecedent snowpack (see **Methods**). SHAP interaction values were used to provide predictive contributions of the pairwise interaction between predictors (Fig. 1b). Training was performed by replicating 500 times; each sample was placed in the training set

400 times and the test set 100 times, with all SHAP interaction values being calculated based on samples in the test sets only.

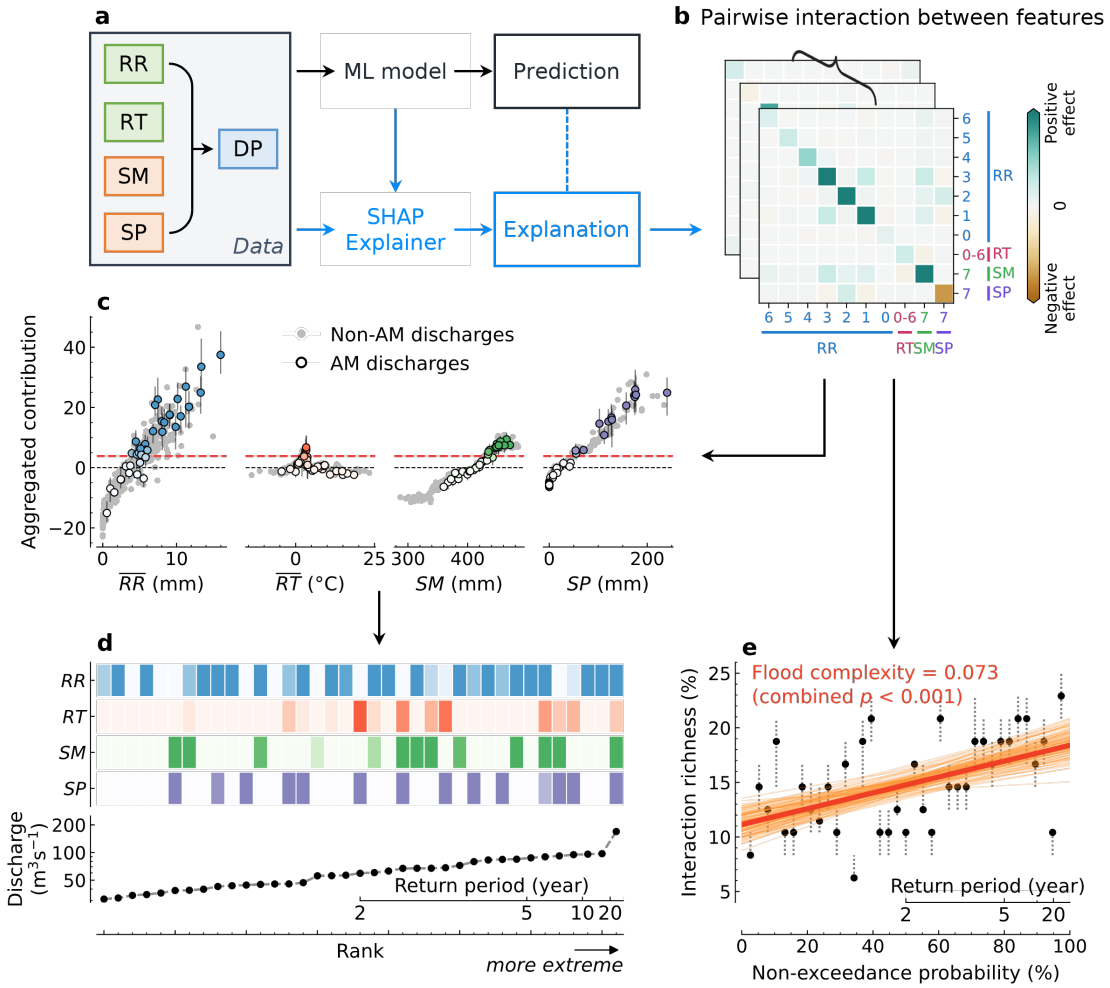


Fig. 1 | Procedure for identifying compounding effects in river flood drivers. A catchment in Slovenia with an outlet at 49.07°N, 18.91°E is used as an example. **a**, Conceptual diagram showing relationships between data, ML model, and explanation. Meteorological drivers (green boxes, including recent rainfall (RR) and recent temperature (RT)) and catchment preconditions (orange boxes, including soil moisture (SM) and snowpack (SP)) are used as input variables. The model output refers to all identifiable discharge peaks (DP) regardless of their magnitude. **b**, Illustration of SHAP interaction values for model outputs. The tick labels indicate the different types of variables, with the number representing the number of days before a discharge peak for the corresponding feature. **c**, The relationship between the aggregated contributions and the event-averaged magnitude of the different types of variables for all samples in the catchment. The points indicate the median of

the respective aggregated contributions over 100 replicates. The error bars indicate the 10th and 90th percentiles of the contribution values for the annual maximum (AM) discharge samples. The red dashed lines indicate the threshold (here, the 80th percentile of the aggregated contributions across all samples) used to identify the main drivers, above which the driver is considered a considerable contributor. Note the small variation in the thresholds (overlapped in the figure) across the 100 replicates. The color saturation of the AM samples indicates the number of exceedances of the corresponding threshold, with higher saturation indicating more exceedances over the 100 replicates. **d**, Summary of the number of exceedances of each threshold for the AM samples, with the x-axis representing the rank of the AM magnitude. Saturation has the same meaning as in panel **c**. **e**, Illustration of the flood complexity. The points indicate the median of the interaction richness over the 100 replicates, with the error bars indicating the 25th and 75th percentiles. The interaction richness (exemplified in Extended Data Fig. 4) is based on the number of interaction values exceeding a threshold (here, the 80th percentile of the positive interaction values across all samples). The dashed lines indicate the fitted slope in each of the replicates, and the flood complexity is the median of these slopes (the solid orange line). The combined *p*-value is estimated using Fisher's method from individual *p*-values that indicate the significance of whether the corresponding slope is positive.

The SHAP interaction values are additive, which allows us to aggregate them into meaningful effects, such as the aggregated contributions of different kinds of variables[25]. Using a catchment in Slovenia as an example, the aggregated SHAP values reveal predictive contributions of the four types of variables in the model (Fig. 1c). A positive contribution indicates that the feature pushed the model output beyond what is expected (e.g., the average of the runoff peaks in the training set), whereas a negative contribution does the opposite. The contributions of recent rainfall, soil moisture, and snowpack all scale with their magnitudes, whereas recent temperature only has a positive contribution when the value is around 5 °C, implying its role in flooding by inducing snowmelt. The robustness of the contribution pattern is confirmed by the error bars, which indicate the variation across the 100 replicates (Fig. 1c). Based on the aggregated contributions, we define a threshold (here, the 80th

percentile of all the aggregated values) to determine the main variables that considerably contribute to discharge peaks. The approach is applied to annual maximum discharges (Fig. 1d), enabling the identification of main drivers and multi-driver floods. Whether a driver is a single or a multi-driver is based on the 100 replicates using a binomial test. Moreover, we define the fraction of main interaction effects (by thresholding the SHAP interaction values) to all possible interaction effects as the interaction richness for individual flood events (Extended Data Fig. 4). We used the simple linear regression slope between the interaction richness and the non-exceedance probability of flood events to measure the flood complexity of a catchment (Fig. 1e), which serves as an indicator of flood heterogeneity resulting from compounding effects between drivers. The flood generation in the Slovenian catchment example (Fig. 1) is likely to be more heterogeneous, and therefore probably less extrapolatable compared to the British catchment shown in Extended Data Fig. 5.

Distribution and impact of compound floods

For all identified flood events, 61.1%, 21.8%, 51.5%, and 20.3% were associated with recent rainfall, recent temperature, soil moisture, and snowpack, respectively. The main variables influencing the annual maximum discharge events vary considerably among catchments (Fig. 2a-d), underlining the spatial heterogeneity of the main processes (e.g., precipitation, soil moisture saturation, snowmelt, and evapotranspiration) that generate floods. The overall patterns of the main driving variables are largely consistent with previously identified global and regional flood types in the literature[13,15,18], although different perspectives were considered and our focus is on the compounding of drivers.

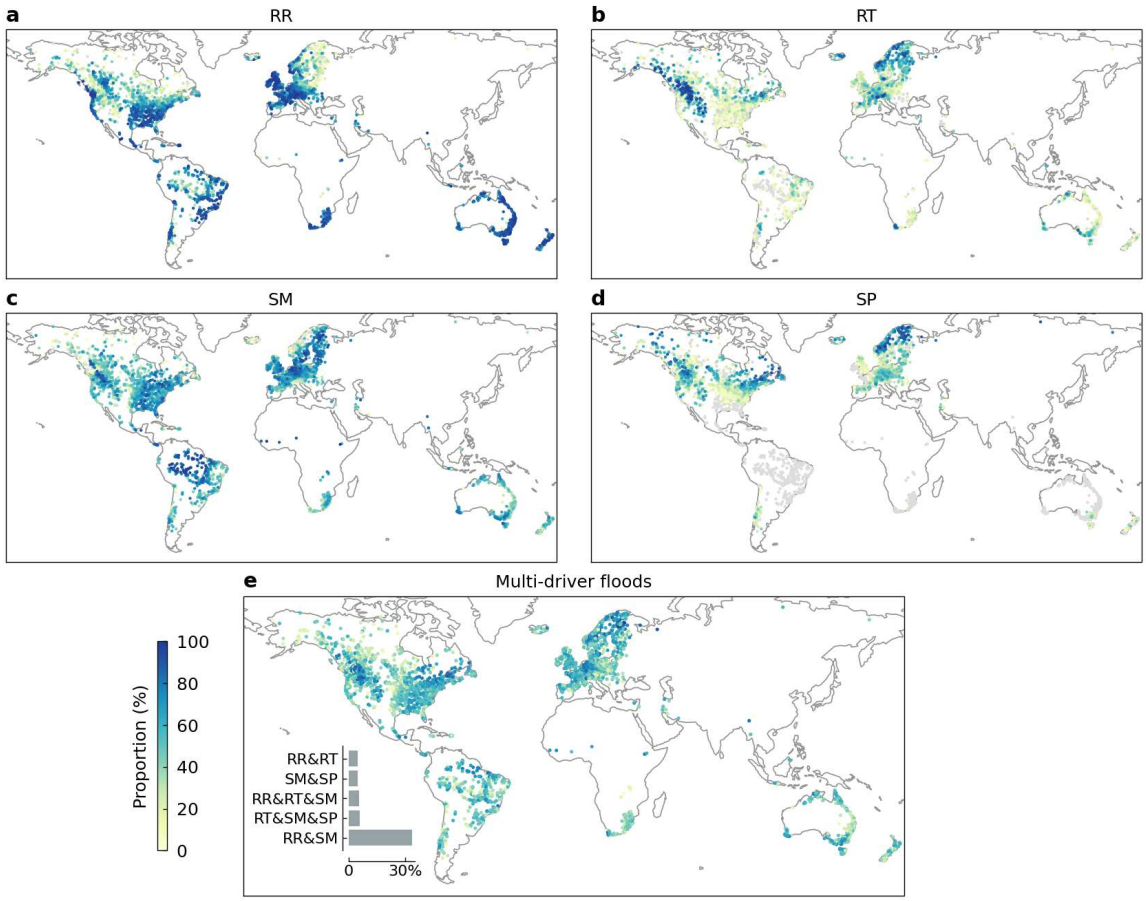
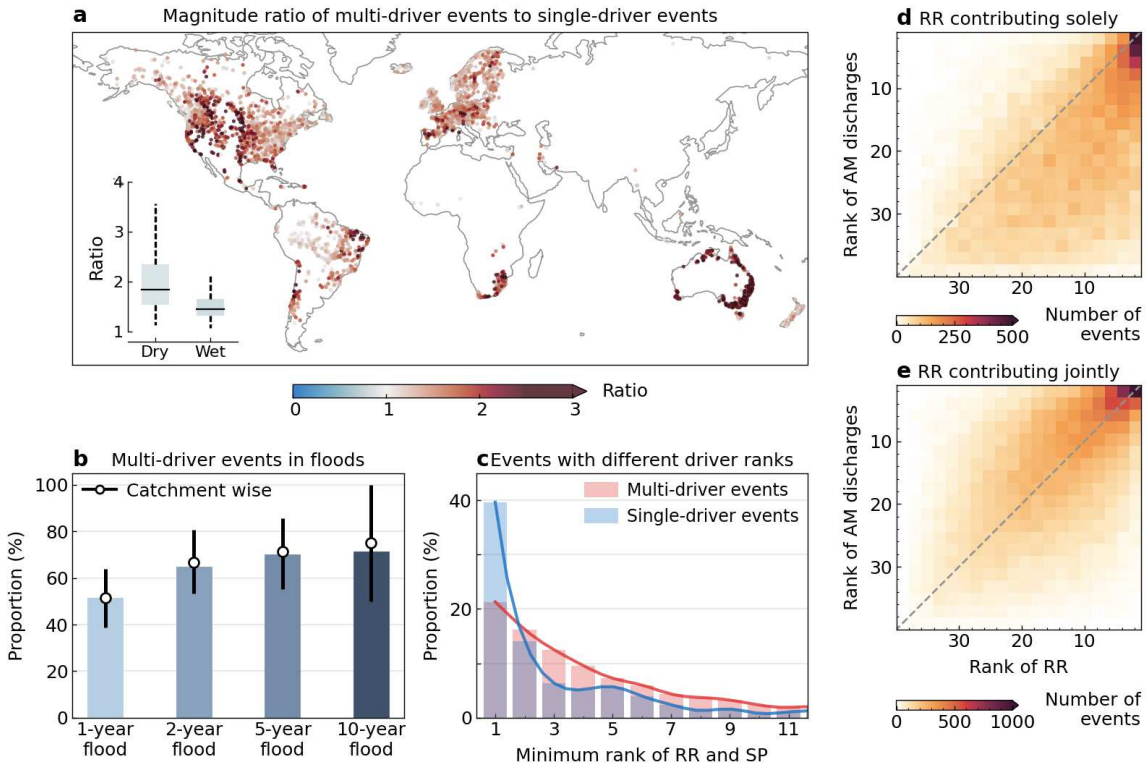


Fig. 2 | Main drivers of annual maximum flood events in the 3,527 catchments. a-d, Proportions of recent rainfall (**a**), recent temperature (**b**), soil moisture (**c**), and snowpack (**d**) as main drivers of AM floods in individual catchments. The gray points indicate catchments for which the respective driver is not considered a main driver in any of the considered floods. **e**, Proportion of flood events for which at least two drivers are considered main drivers. The inset represents the proportion of specific combinations of drivers in all multi-driver flood events (only combinations within the top five proportions are shown). The 80th percentile of the aggregated contributions of all samples in each replicate was used as the cutoff for whether the driver contribution was considered (see Fig. 1c and **Methods**).

We further examined the distribution of multi-driver floods that are associated with at least two of the main drivers identified above (Fig. 2e). Of the 124,642 flood events, 51.6% were attributable to at least two drivers. Almost all the studied catchments have experienced multi-driver floods to a greater

or lesser extent. In 55.1% of the catchments, more than 50% of the floods were multi-driver floods. In particular, the joint contribution of RR and SM accounts for 33.5% of the multi-driver floods, which highlights the preconditioning role of antecedent soil moisture in flood development[15,26]. The various combinations of RR, RT, SM, and SP suggest complex interplays between soil moisture, rainfall, evapotranspiration, and snowmelt in flood generation, emphasizing the importance of understanding river flood risks from a multivariate perspective[5].

The mean magnitude of multi-driver floods is significantly higher in 56.1% of the 3,527 catchments than in single-driver floods (one-sided t-test, α=0.05). Among these catchments, the mean magnitude of multi-driver floods is at least 20% higher than that of non-multi-driver floods in 63.9% of catchments, and at least 50% higher in 28.5% of catchments (Fig. 3a). The catchments with a higher magnitude ratio of multi-driver floods to single-driver floods are generally characterized by a high degree of aridity (Extended Data Fig. 3), with a strong negative correlation between this ratio and the climate moisture index of the catchments (Spearman's rank correlation coefficient = -0.55, p<0.001). Previous studies have consistently found that drier catchments have heavier flood tails and our results suggest that compounding drivers that tend to increase nonlinear interactions between processes may be a possible mechanism[3,27].



166

167 **Fig. 3 | Impact of compounding drivers on the severity of river flood events. a**, The magnitude
168 ratio of multi-driver floods to single-driver floods in individual catchments. The ratio was calculated
169 as the mean magnitude of multi-driver floods divided by the mean magnitude of single-driver floods.
170 The inset box plot compares the magnitude ratio between dry catchments (climate moisture index < 0)
171 and wet catchments (climate moisture index ≥ 0) (Extended Data Fig. 3). **b**, Proportion of multi-driver
172 floods associated with different return periods. The height of the bars indicates the event-wise
173 proportion, whereas the white points indicate the median of the catchment-wise proportion with the
174 error bars indicating the 25th and 75th percentiles. **c**, Comparison between multi-driver and single-
175 driver floods in terms of the minimum rank of recent rainfall (RR) and antecedent snowpack (SP)
176 prior to individual events. Lower ranks correspond to higher magnitudes. **d**-**e**, Event-wise ranks of RR
177 against event-wise ranks of annual maximum (AM) discharges when RR was the sole contributor (**d**)
178 and when RR contributed with other drivers (**e**). The color indicates the two-dimensional histogram of
179 RR-AM discharge rank pairs across all catchments, with dark indicating more counts. The number of
180 bins for each dimension was set to 20.

181

Moreover, the likelihood of floods being associated with multiple drivers generally increases as floods become more extreme (Fig. 3b). The proportion of multi-driver events in all annual maximum floods (124,642 events in total) increased from 51.6% to 64.9%, 70.1%, and 71.3% for floods with a catchment-specific return period of at least 2 years (63,199 events), 5 years (24,289 events), and 10 years (11,338 events), respectively. The catchment-wise proportions largely follow this pattern (Fig. 3b), whereas the variance of the proportions becomes larger for 10-year floods. In some catchments, multi-driver floods may be less dominant for extreme floods than for moderate floods (e.g., 5-year floods) over the past 40 years, particularly when a single driver is extreme enough to dominate the generation of extreme floods[12].

We further compared the extremeness of the drivers between the largest floods globally from 1981 to 2020 with and without multiple associated drivers (Fig. 3c). For the largest floods without multiple drivers, the event-wise rainfall or snowpack tends to have top rankings in its magnitude. In contrast, the extremeness of rainfall or snowpack that trigger multi-driver floods presents a wider distribution, suggesting a considerable chance that non-extreme drivers cause extreme outcomes. A comparison of the relevance of the event-wise rank of the recent rainfall and flood magnitude in each catchment further revealed differences between floods induced by recent rainfall only and floods jointly triggered by recent rainfall and additional drivers. When recent rainfall contributes to flooding events jointly with other drivers, the mean rank of recent rainfall is significantly higher than that of the annual maximum flood (Fig. 3e, one-sided paired t-test, $p<0.001$). In contrast, when recent rainfall is a single flood driver, even extreme recent rainfall does not always cause large floods (high density below the diagonal in Fig. 3d), which is likely due to buffering (i.e., negative contributions) by other drivers (e.g., drier soils). Considering only catchments with a longer observation period (e.g., at least 35 annual maximum discharges) does not change the above conclusion. These combined results demonstrate again the role of compounding drivers in amplifying the flood magnitude of a river and illustrate that extreme outcomes can result from non-extreme drivers, which underscores the importance of considering the compounding nature of flood drivers in risk management.

Since the identification of contributing drivers is highly dependent on the selected threshold, we conducted a number of sensitivity tests. Generally, lower thresholds lead to a higher number of contributing drivers, resulting in more flood events being classified as multi-driver floods, and vice versa for higher thresholds. However, different thresholds lead to similar spatial patterns of multi-driver flood proportions (Extended Data Fig. 6), showing the robustness of our conclusions against different threshold choices. In particular, the proportions of multi-driver floods using different thresholds are strongly correlated. Similarly, the conclusions about the relevance of compounding drivers and flood extremeness drawn in Fig. 3 hold with different thresholds (Extended Data Fig. 7). In addition, a stricter criterion for the predictive performance of ML algorithms does not affect the above conclusion (Extended Data Fig. 8).

Varying degrees of flood complexity

We estimated flood complexity for all 3,527 catchments from 1981 to 2020, which were defined as the regression slope between the interaction richness and non-exceedance probability of flood events (Fig. 4a). Of the catchments studied, 96.1% have a significantly positive slope (combined p-value<0.01 by Fisher's method from the 100 replicates). Catchments with a low flood complexity (often coinciding with non-significant flood complexity overall) were mainly distributed in the northern regions, the Alpine region, and the Amazon Basin. For high-latitude and high-altitude regions, flood generations tend to be uniformly dominated by snowmelt, and therefore the interaction richness is likely to be homogeneous across flood magnitudes. In contrast, floods in the Amazon Basin are typically triggered by saturated soil moisture that has accumulated during the rainy season[28], resulting in low variability in the richness of driver interactions during the few days preceding different magnitudes of flood events.

Regions with average flood complexity of catchments significantly higher than the global average (one-sided t-test, α=0.001) mainly include eastern Brazil, the Andes, eastern Australia, the Rocky Mountains extending to the west coast, and the western and central European plains (Fig. 4a-b). Catchments in these regions typically have multiple flooding mechanisms. For example, catchments

234 in the European plains may experience flooding caused by recent rainfall alone, or by both recent
235 rainfall and snowmelt/antecedent soil moisture[14,18]. The various combinations of factors and processes
236 involved in the generation of the catchment response produce a wide range of hydrologic behaviors
237 with varying degrees of interactions and nonlinearity[29]. The estimated flood complexity is not
238 dependent on the choice of the threshold that determines the main interaction effects, which
239 demonstrates the robustness of the above results (Extended Data Fig. 9a-d).

240 Comparing flood complexity against the catchment-averaged climate moisture index and snow cover
241 extents indicates that drier catchments and wet catchments with moderate snow cover tend to display
242 a high flood complexity (Figs. 4c-d). Generally, rainfall-runoff processes in humid catchments are
243 assumed to be more linear due to the reduced variability of hydrological conditions, whereas
244 catchments in arid environments may experience more disruptions to within-basin connectivity[30-32]. In
245 wet catchments with moderate snow cover, the potential interactions between both rainfall and
246 snowmelt processes make flood generation mechanisms more heterogeneous and complex than in
247 catchments dominated by rainfall or snowmelt alone. This relationship holds under a more stringent
248 criterion for the predictive performance of ML algorithms (Extended Data Fig. 9e-f). In addition to
249 the global spatial variability caused by different climatic conditions, catchments within the same
250 IPCC reference region[33] also exhibit high variance in flood complexity (Fig. 4b). This variance is
251 likely due to local differences in catchment characteristics such as physiography, vegetation, and
252 soils. Spearman's rank correlation analysis between flood complexity and representative local
253 characteristics across various climate reference regions suggests that catchments with higher flood
254 complexity tend to exhibit larger size, flatter terrain, reduced forest cover, and lower sand content.
255 These associations may be related to the effects of scale and soil storage capacity[30,34,35]. However,
256 measurement uncertainty and confounding factors may complicate the correlation analysis for
257 different flood processes, which warrants further investigation.

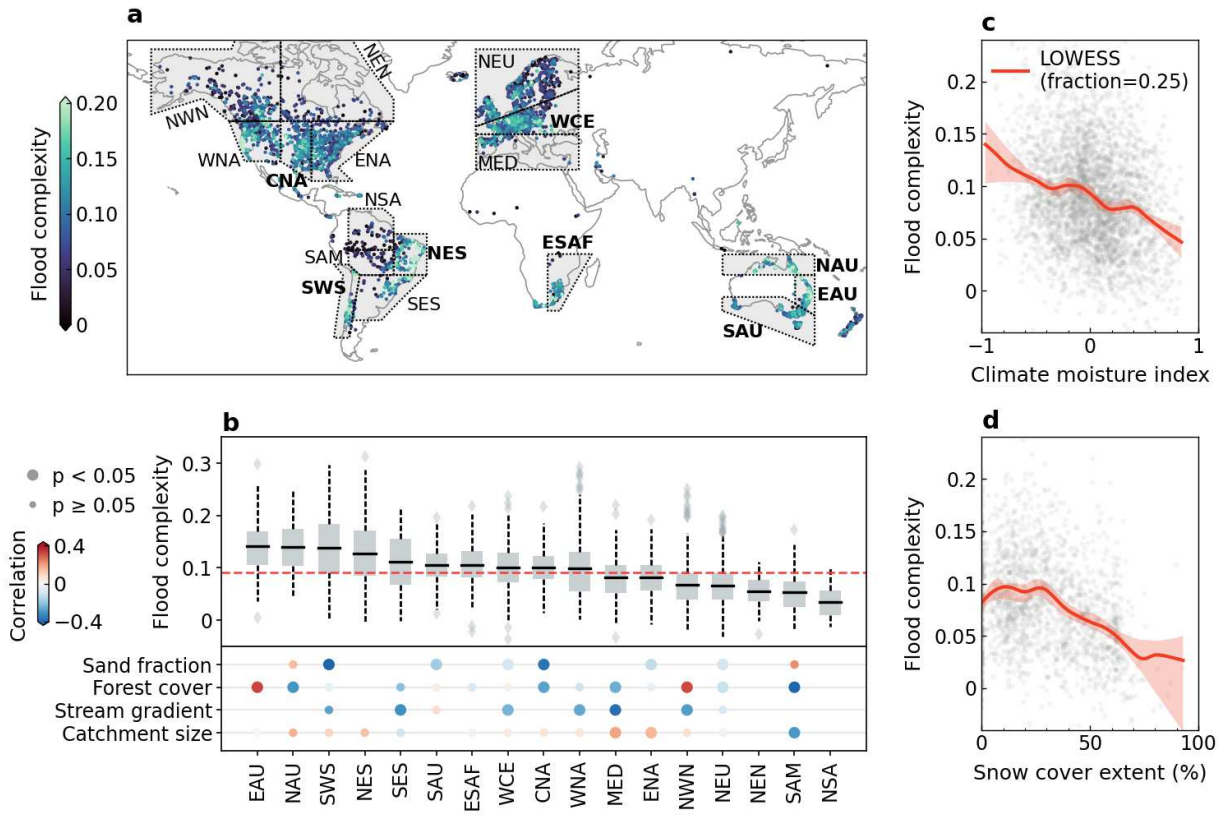
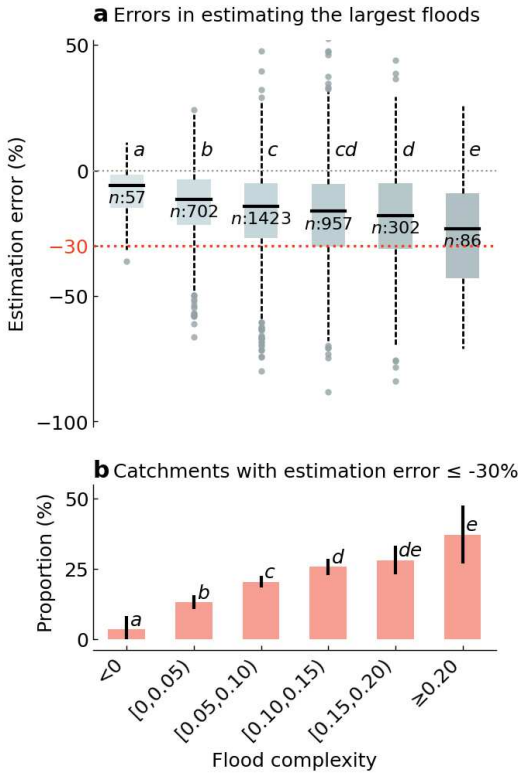


Fig. 4 | Flood complexity and its relationship to catchment attributes. a, Spatial distribution of flood complexity for individual catchments (median across 100 replicates). The gray polygons represent the IPCC climate reference regions[33], each of which contains at least 50 study catchments. The bold abbreviation indicates the region has an average flood complexity significantly higher than the global average (one-sided *t*-test, α=0.001). **b**, Flood complexity of catchments in IPCC reference regions and correlations with catchment characteristics. The red dashed line indicates the average flood complexity across all catchments. Box plots show the median, 25th percentile, 75th percentile, and 1.5x interquartile range of the data. Correlations in the lower panel represent Spearman's rank correlations (within each region) between flood complexity and catchment size, catchment-average stream gradient, forest cover extent, and sand fraction in soil. We only display the correlation results for reference regions where catchment attributes have sufficient variability, e.g., the interdecile range of a catchment attribute in the region should be at least 60% of the interdecile range of the same attribute across all catchments. **c**-**d**, The relationship between flood complexity and catchment-average snow cover extent (**c**) and climate moisture index (**d**). The red line shows the LOWESS (locally weighted scatterplot smoothing) of the points and the shaded area indicates the 95%

confidence interval based on 1,000 bootstraps. Only wet catchments (climate moisture index ≥ 0) are considered in panel **d**.

To assess the potential errors in estimating extreme events in catchments with varying degrees of flood complexity, we conducted flood frequency analysis based on all annual flood events except the largest one within individual catchments. We find that estimates of the magnitude of the largest flood is severely underestimated in catchments with higher flood complexity (Fig. 5). For example, in the 388 catchments with a flood complexity > 0.15 (the green and white points in Fig. 4a), the estimated magnitude of the largest floods is, on average, 20.3% lower than the observed values. This serious underestimation can pose an enormous risk to practical hydrological design and flood management[3,36,37]. Our results suggest that catchments without recorded extreme flood events may be subject to unexpectedly large events, and such a case could be exacerbated if the physical processes in flood generation are heterogeneous in the catchment due to compounding effects[38,39]. Sensitivity analyses show that the conclusion is robust even if only catchments with a longer observation period are considered (Extended Data Fig. 10).



289

290 **Fig. 5 | Impact of flood complexity on estimating large flood magnitudes. a**, Estimation error in
291 the magnitude of the largest observed floods per catchment against different levels of flood
292 complexity. The estimation error in each catchment is calculated as the relative error between the
293 estimated magnitude extrapolated from all other flood events and the observed magnitude of the
294 largest flood in the observations (see **Methods**). A negative error indicates an underestimation of the
295 largest flood magnitude. Box plots show the median, 25th percentile, 75th percentile, and 1.5x
296 interquartile range across catchments within different flood complexity bins (x-axis). The number
297 indicates the sample size of each bin, and the letters are assigned based on a one-sided *t*-test with a
298 significance level of 0.05 (different letters indicate statistically significant differences in the mean
299 estimation error). **b**, Proportions of catchments where the estimated largest flood is 30% lower than
300 observed (i.e., below the red dashed line in panel **a**). The error bar indicates the 95% confidence
301 interval, which is approximated as $\hat{p} \pm 1.96\sqrt{\frac{\hat{p}(100-\hat{p})}{n}}$ ($\hat{p}$ is the estimated proportion and n is the
302 sample size). The letters above each bar indicate the significance of the difference between
303 proportions (one-sided z-test, α=0.05).

Discussion

Recent literature has increasingly focused on the compounding effects of drivers of river floods, which can potentially improve our understanding of flood extremes under historical conditions and enhance predictive capabilities for future flood risks[5]. In particular, the way in which increased extreme precipitation will translate into changes in river flooding remains controversial[12,13,40], and therefore we may benefit from learning more about the interplay between soil moisture, precipitation, evapotranspiration, and snowmelt[41]. Despite the theoretical hypotheses, there have been few large-scale investigations of the compounding drivers of river floods and the impact of such compounding effects on the severity of river flood events, which requires reliably capturing the nonlinear interactions and dependencies between variables[42].

Based on advanced explainable ML techniques, this study developed a new approach to quantifying the compounding effects between river flood drivers in a unified framework (Fig. 1). Our results demonstrate that the compounding effects were prevalent in river floods around the globe (Fig. 2e) and show their relevance for extreme river floods in most catchments (Fig. 3b). Compounding drivers can make river floods more severe (Fig. 3a) and even moderate drivers can cause potentially extreme floods when interacting with each other (Fig. 3c-e), which highlights the need to understand river flood risks from a compound event perspective. We also examined the flood complexity in catchments resulting from the compounding effects, which is found to be highly variable across catchments with different climatic and local conditions (Fig. 4). Higher flood complexity implies that the generation processes are more heterogeneous across flood magnitudes, thereby suggesting less reliability in extrapolating from moderate floods to the estimation of extreme flood risk (Fig. 5). Our results suggest that flood risk assessment and management in arid and snow-rain mixed catchments in particular should be approached with caution.

The origins of large (extreme) floods have been the subject of debate as to whether the extremes are distinct from small floods or arise from the same distributions as small floods but are simply associated with different magnitudes of the same process[37,43,44]. Understanding these origins is important for selecting and designing appropriate flood estimation methods, most of which assume

homogeneous flood samples and are widely used in engineering practice[39]. Our results suggest that the disagreement can probably be reconciled by considering different catchment conditions that may affect the compounding nature of river floods. In regions with high flood complexity (e.g., arid and snow-rain mixed catchments), conventional statistical methods based on the homogeneity assumption may strongly underestimate extreme flood magnititudes[3].

The ubiquitous compounding effects in river floods also pose a huge challenge to flood risk management in a warming climate, especially considering that different trends in flood drivers will alter the magnitudes and associated probabilities of extreme floods[10,11], thereby challenging current risk management measures. Moreover, climate change is likely to increase the currently low flood complexity in high-latitude regions because the present snowmelt-dominated flood generation will probably become more mixed with rainfall[45,46]. These regions may therefore be underprepared for potential flood risk increases because flood magnitudes have decreased in recent decades concurrent with decreasing snowmelt. Although snowmelt so far has remained the dominant driver in those catchments, the expected increase in precipitation extremes may soon break the stationarity and lead to unprecedented flood disasters[12].

Methods

Historical observations and simulations

Daily discharge observations in 3,527 catchments around the world (Extended Data Fig. 1) from 1981 to 2020 were obtained from the Global Runoff Data Center (https://www.bafg.de/GRDC, accessed August 1, 2022), which consists of discharge records from approximately 10,000 catchments worldwide. The 3,752 catchments were selected using the following criteria. First, the catchment should have daily discharge records for at least 20 years during 1981 to 2020. Second, the catchments are larger than 100 km^2 to encompass at least one grid cell of the meteorological datasets (at ~9 km spatial resolution) and smaller than 100,000 km^2. Third, the predictive relationship between hydrometeorological data and identified peak flows should be well captured by the ML model.

Therefore, the catchments with an average R^2 regression score in the test periods across all replicated cross-validations below 0.3 were excluded.

Global daily precipitation and air temperature from 1979 to 2020 were obtained from the MSWEP (Multi-Source Weighted-Ensemble Precipitation) and MSWX (Multi-Source Weather) datasets[47,48], respectively, with a spatial resolution of 0.1°. Daily rainfall, soil moisture storage, and snowpack were estimated by a gridded implementation of the HBV model[49] using daily precipitation and temperature. The model has been well calibrated with daily observed discharge from over 4,000 catchments worldwide using the same MSWEP precipitation product[50]. The simulations between 1979 and 1980 were excluded to account for model warm-up processes. The gridded rainfall, temperature, soil moisture, and snowpack were aggregated to individual catchments (Extended Data Fig. 2), with catchment boundaries delineated using an automated outlet relocation algorithm[51]. Overall, the spatially aggregated runoff outputs for each catchment simulate the observed discharge well, with a median Kling-Gupta Efficiency of 0.47 and a median Pearson correlation coefficient of 0.76 across the 3,527 catchments, which are comparable to the model performance reported[50]. Importantly, the simulation data would be reconsidered in ML models in terms of their predictive relationship to the discharge observations. Catchments with underperformed predictions due to potentially poor simulation data were excluded.

Catchment attributes

We chose climate moisture index, snow cover extent, catchment size, stream gradient, forest cover extent, and sand fraction in soil to reflect certain aspects of a catchment's climatology, physiography, vegetation, and soil attributes, which considers both their representativeness and relevance to flood generation as reported in the literature[34,35]. These catchment attributes (except catchment size) were derived from the well-established HydroATLAS dataset[52], which provides a compendium of descriptive hydro-environmental information for all (sub)basins worldwide with nested levels, each representing consistently sized polygons. We used the highest spatial resolution level (level 12), which has a scale of approximately tens of square kilometers. We then derived the relevant attributes

for the 3,527 catchments in our study using an area-weighted aggregate based on the coverage (Extended Data Fig. 3). Among these attributes, the climate moisture index (CMI) was calculated from annual precipitation (P) and potential evapotranspiration (PET) using the equation: [CMI = (P / PET) - 1 if P < PET] or [CMI = 1 - (PET / P) if P ≥ PET]. The CMI ranges from -1 to 1, with a higher value indicating wetter conditions. Catchment size was estimated directly from the catchment boundary. Stream gradient refers to the ratio of the slope within the stream reach to the length of the reach.

Preprocessing of training samples

The training targets for the ML models used in this study are identifiable discharge peaks, regardless of their extremeness, in the daily discharge series of each catchment (Extended Data Fig. 2). The identification follows the procedure recommended by the guidelines of the US Water Resources Council, which has been widely adopted in many studies[53,54]. First, all local peaks with a minimum distance $T = 5$ days + log(A) between neighboring peaks were selected, where A is the basin area in square miles and T is rounded to an integer. Then, the criterion that the minimum discharge between two consecutive peaks should be less than 75% must be satisfied, otherwise the smallest peak in such a pair is removed until the condition is fulfilled for all remaining peaks. In total, we identified 1,582,043 discharge peaks for all 3,527 catchments, with an average of 12.2 event peaks per year across the catchments. Note that although only annual maximum discharges are considered in the subsequent analysis, we trained the model on all identifiable peak flows for two main reasons. The first is to increase training samples to better capture diverse runoff processes. The second is to provide an appropriate background for the interpretation of ML models[55], which serves as a reference point for identifying the contributions of features in the input to the prediction.

We used the time series of daily precipitation and 7-day mean temperature in the last 7 days before the discharge peaks, and soil moisture and snowpack on the day before this 7-day synoptic window as inputs to the ML model (Extended Data Fig. 2). The 7-day time window was chosen according to previous studies[9,14,15]. We averaged the temperature within the 7-day synoptic window instead of

using the time series within it to simplify the model complexity for better interpretability and to avoid interpretation instability due to autocorrelation between temperatures in the model inputs.

Training and interpretation of ML models

We used the LightGBM as the ML model, which is based on decision tree algorithms in a gradient boosting framework[24]. In our preliminary experiments, we compared it with other tree-based ML models (e.g., random forest and XGBoost) and eventually chose LightGBM after accounting for both model performance and efficiency. For each catchment, we used repeated five-fold cross-validation for model evaluation and interpretation. Specifically, we repeated the five-fold cross-validation process 100 times for each catchment, meaning that each data sample was independently evaluated and interpreted 100 times. In each iteration of the 100 replicates, the sample set was split five-fold differently and, in turn, the interpretation results in each sample in different test periods were based on the model trained on different training samples. Repeated cross-validation can help reduce the randomness and potential bias of the interpretation baseline.

As prior knowledge, we enforced monotonicity constraints on the input features of rainfall, soil moisture, and snowpack in the model, meaning that these features have a monotonically increasing relationship to the discharge response. Our preliminary experiments showed that such constraints improve the predictive performance of the model. We also disabled the interactions between the input features of rainfall and temperature in the model to ensure the interpretability of the compounding effects are consistent with our domain knowledge. In each training process, the hyperparameters of the LightGBM model were automatically searched by optimizing the model performance on a subset of the corresponding training samples. The candidate hyperparameters are listed in Extended Data Table 1.

The SHAP interaction values[25] were used to explain the model outputs in terms of the predictive contributions of the pairwise interactions between input features. The SHAP interaction values for each sample consist of a matrix of feature attributions (interaction effects on the off-diagonal and main effects on the diagonal, as shown in Fig. 1b). In this study, we calculated SHAP interaction

values for all test samples for each of the catchments. Intuitively, the SHAP interaction values explain why the prediction was different from the expected output (i.e., the average of the training targets). As an additive feature attribution method, the SHAP interaction values between feature *i* and all features (including feature *i* itself) sum to the predictive contribution of feature *i*. Therefore, we derived the predictive contribution of different types of variables by aggregating all the interaction effects of the corresponding variables.

Determining main drivers and identifying multi-driver floods

For each of the 100 replicates, we can derive the aggregated contribution of recent rainfall, recent temperature, antecedent soil moisture, and antecedent snowpack to every identifiable discharge peak. We then determined a cutoff threshold to distinguish whether the specific driver had a considerable contribution. In the main text, we used the 80th percentile of the aggregated contribution values for all peaks as the threshold, above which the driver is considered a main driver. A flood associated with at least two main drivers is regarded as a multi-driver flood. We can also determine the specific combination for the multi-driver flood (e.g., a flood driven by main contributions from recent rainfall and antecedent soil moisture). For each flood event in the 100 replicates, we can derive a set of Boolean values (100 elements in this case) that indicate whether the event is associated with a particular driver or combination. We took the majority as the final result and tested its significance using the binomial test. With 100 replicates and a significance level of 0.01, a driver or a combination with at least 63 exceedances of the respective thresholds is considered to be significantly associated with the corresponding flood event. Especially, a flood that is associated with multiple drivers at least 63 times in the 100 replicates is regarded as a multi-driver flood. Conversely, a flood that is associated with multiple drivers at most 37 times is regarded as a single-driver flood. Note that a single-driver flood may include cases in which the flood is not associated with any main drivers (threshold for aggregated contribution never exceeded).

458 **Richness of interactions and flood complexity**

459 Similar to the thresholding of aggregated contribution values, we can threshold positive interactions
460 between features in the original SHAP interaction values to identify the main interaction effects that
461 considerably contribute to the prediction of discharge peaks. We define the fraction of the main
462 interaction effects to all possible interaction effects as the richness of interaction for individual flood
463 events (exemplified in Extended Data Fig. 4). In each of the 100 replicates, we used the simple
464 regression model to fit a slope between the non-exceedance probability and the richness of interaction
465 of annual maximum flood events in a catchment. The regression slope, which we defined as the flood
466 complexity of a catchment, indicates the change in the richness of interaction for every 1% increase in
467 the non-exceedance probability of floods. We derived 100 slopes and associated *p*-values that test
468 whether the corresponding slope is significantly positive. We used the median slope as the final slope
469 across the 100 replicates and estimated the combined *p*-value using Fisher's method.

470 **Errors in estimating large flood magnitudes**

471 The estimation error in the magnitude of the largest observed flood in a catchment is calculated as
472 $(Q_{\mathrm{est}} - Q_{\mathrm{obs}})/Q_{\mathrm{obs}}$, where Q_{obs} is the magnitude of the largest observed flood during the study
473 period (from 1980 to 2020) and Q_{est} is the estimated magnitude based on the flood frequency
474 analysis. Specifically, we first fit the available annual maximum discharge events during the study
475 period (the largest is assumed to be unknown and thus not included) using the Generalized Extreme
476 Value (GEV) distribution. For the annual maximum discharge events, only events in a calendar year
477 with at least 200 days of discharge records were considered. The parameters of the GEV distribution
478 are estimated by maximum likelihood. We then calculate Q_{est} for the largest flood from the fitted
479 GEV distribution, given the empirical return period of the largest observed flood.

480 **Data availability**

481 The streamflow records were obtained from the Global Runoff Data Centre
482 (http://www.bafg.de/GRDC). Global daily precipitation is available at http://www.gloh2o.org/mswep
483 and the daily temperature is available at http://www.gloh2o.org/mswx. The HydroATLAS dataset can

be accessed at https://www.hydrosheds.org/hydroatlas. The data used for producing the main figures in the study are deposited in https://doi.org/10.5281/zenodo.7765151.

Code availability

The HBV model and the associated parameter maps used to generate soil moisture and snowpack are available at http://www.gloh2o.org/hbv/. The code for the identification of compounding effects (including model training, interpretation, and analysis) can be found in https://doi.org/10.5281/zenodo.7765151.

Acknowledgements

The authors acknowledge the European COST Action DAMOCLES (CA17109) and the Helmholtz Initiative and Networking Fund (Young Investigator Group COMPOUNDX; grant agreement no. VH-NG-1537). This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101003469 (XAIDA).

Author information

Author contribution

SJ and JZ conceived the study. SJ performed all analyses and wrote the initial draft. All authors substantially contributed to the final draft.

Correspondence

Correspondence to Shijie Jiang (shijie.jiang@ufz.de)

Competing interests

The authors declare that they have no conflict of interest.

References

1 Kreibich, H. *et al.* The challenge of unprecedented floods and droughts in risk management. *Nature* **608**, 80-86 (2022). https://doi.org:10.1038/s41586-022-04917-5

2 Merz, B. *et al.* Causes, impacts and patterns of disastrous river floods. *Nature Reviews Earth & Environment* **2**, 592-609 (2021). https://doi.org:10.1038/s43017-021-00195-3

3 Merz, B. *et al.* Understanding Heavy Tails of Flood Peak Distributions. *Water Resources Research* **58**, e2021WR030506 (2022). https://doi.org:https://doi.org/10.1029/2021WR030506

4 Zscheischler, J. *et al.* Future climate risk from compound events. *Nat. Clim. Chang.* **8**, 469-477 (2018). https://doi.org:10.1038/s41558-018-0156-3

5 Brunner, M. I., Slater, L., Tallaksen, L. M. & Clark, M. Challenges in modeling and predicting floods and droughts: A review. *WIREs Water* **8**, e1520 (2021). https://doi.org:https://doi.org/10.1002/wat2.1520

6 Dottori, F. *et al.* Increased human and economic losses from river flooding with anthropogenic warming. *Nat. Clim. Chang.* **8**, 781-+ (2018). https://doi.org:10.1038/s41558-018-0257-z

7 Berghuijs, W. R., Woods, R. A., Hutton, C. J. & Sivapalan, M. Dominant flood generating mechanisms across the United States. *Geophysical Research Letters* **43**, 4382–4390 (2016). https://doi.org:10.1002/2016gl068070

8 Tarasova, L. *et al.* Causative classification of river flood events. *Wiley Interdisciplinary Reviews-Water* **6** (2019). https://doi.org:10.1002/wat2.1353

9 Blöschl, G. *et al.* Changing climate shifts timing of European floods. *Science* **357**, 588-590 (2017). https://doi.org:10.1126/science.aan2506

10 Blöschl, G. *et al.* Changing climate both increases and decreases European river floods. *Nature* **573**, 108-111 (2019). https://doi.org:10.1038/s41586-019-1495-6

11 Tarasova, L. *et al.* Shifts in flood generation processes exacerbate regional flood anomalies in Europe. *Communications Earth & Environment* **4**, 49 (2023). https://doi.org:10.1038/s43247-023-00714-8

12 Brunner, M. I. *et al.* An extremeness threshold determines the regional response of floods to changes in rainfall extremes. *Communications Earth & Environment* **2** (2021). https://doi.org:10.1038/s43247-021-00248-x

13 Zhang, S. *et al.* Reconciling disagreement on global river flood changes in a warming climate. *Nat. Clim. Chang.* **12**, 1160-1167 (2022). https://doi.org:10.1038/s41558-022-01539-7

14 Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J. & Kirchner, J. W. The relative importance of different flood-generating mechanisms across Europe. *Water Resources Research* **55**, 4582–4593 (2019). https://doi.org:10.1029/2019wr024841

15 Stein, L., Pianosi, F. & Woods, R. Event-based classification for global study of river flood generating processes. *Hydrological Processes* **34**, 1514-1529 (2020). https://doi.org:10.1002/hyp.13678

16 Merz, R. & Blöschl, G. A process typology of regional floods. *Water Resources Research* **39** (2003). https://doi.org:10.1029/2002wr001952

17 Sikorska, A. E., Viviroli, D. & Seibert, J. Flood-type classification in mountainous catchments using crisp and fuzzy decision trees. *Water Resources Research* **51**, 7959-7976 (2015). https://doi.org:10.1002/2015wr017326

18 Jiang, S., Bevacqua, E. & Zscheischler, J. River flooding mechanisms and their changes in Europe revealed by explainable machine learning. *Hydrol. Earth Syst. Sci.* **26**, 6339-6359 (2022). https://doi.org:10.5194/hess-26-6339-2022

19 Zscheischler, J. *et al.* A typology of compound weather and climate events. *Nature Reviews Earth & Environment* **1**, 333-347 (2020). https://doi.org:10.1038/s43017-020-0060-z

20 Rössler, O. *et al.* Retrospective analysis of a nonforecasted rain-on-snow flood in the Alps – a matter of model limitations or unpredictable nature? *Hydrol. Earth Syst. Sci.* **18**, 2265-2285 (2014). https://doi.org:10.5194/hess-18-2265-2014

21 Wahl, T., Jain, S., Bender, J., Meyers, S. D. & Luther, M. E. Increasing risk of compound flooding from storm surge and rainfall for major US cities. *Nat. Clim. Chang.* **5**, 1093-1097 (2015). https://doi.org:10.1038/nclimate2736

22 Bevacqua, E. *et al.* Higher probability of compound flooding from precipitation and storm surge in Europe under anthropogenic climate change. *Science Advances* **5**, eaaw5531 (2019). https://doi.org:doi:10.1126/sciadv.aaw5531

23 Li, D., Lettenmaier, D. P., Margulis, S. A. & Andreadis, K. The Role of Rain-on-Snow in Flooding Over the Conterminous United States. *Water Resources Research* **55**, 8492-8513 (2019). https://doi.org:https://doi.org/10.1029/2019WR024950

24 Ke, G. *et al.* in *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)* Vol. 30 3149–3157 (Long Beach, USA, 2017).

25 Lundberg, S. M. *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* **2**, 56-67 (2020). https://doi.org:10.1038/s42256-019-0138-9

26 Wasko, C. & Nathan, R. Influence of changes in rainfall and soil moisture on trends in flooding. *Journal of Hydrology* **575**, 432-441 (2019). https://doi.org:10.1016/j.jhydrol.2019.05.054

27 Guo, J. *et al.* Links between flood frequency and annual water balance behaviors: A basis for similarity and regionalization. *Water Resources Research* **50**, 937-953 (2014). https://doi.org:https://doi.org/10.1002/2013WR014374

28 Chagas, V. B. P., Chaffe, P. L. B. & Blöschl, G. Process Controls on Flood Seasonality in Brazil. *Geophysical Research Letters* **49**, e2021GL096754 (2022). https://doi.org:https://doi.org/10.1029/2021GL096754

29 Maneta, M. P., Soulsby, C., Kuppel, S. & Tetzlaff, D. Conceptualizing catchment storage dynamics and nonlinearities. *Hydrological Processes* **32**, 3299-3303 (2018). https://doi.org:https://doi.org/10.1002/hyp.13262

30 Pande, S. & Moayeri, M. Hydrological Interpretation of a Statistical Measure of Basin Complexity. *Water Resources Research* **54**, 7403-7416 (2018). https://doi.org:https://doi.org/10.1029/2018WR022675

31 Parajka, J. *et al.* Comparative assessment of predictions in ungauged basins – Part 1: Runoff-hydrograph studies. *Hydrol. Earth Syst. Sci.* **17**, 1783-1795 (2013). https://doi.org:10.5194/hess-17-1783-2013

32 Botter, G., Basso, S., Rodriguez-Iturbe, I. & Rinaldo, A. Resilience of river flow regimes. *Proceedings of the National Academy of Sciences* **110**, 12925-12930 (2013). https://doi.org:doi:10.1073/pnas.1311920110

33 Iturbide, M. *et al.* An update of IPCC climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth Syst. Sci. Data* **12**, 2959-2970 (2020). https://doi.org:10.5194/essd-12-2959-2020

34 Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F. & Woods, R. A. How do climate and catchment attributes influence flood generating processes? A large-sample study for 671 catchments across the contiguous USA. *Water Resources Research* **57**, 21 (2021). https://doi.org:10.1029/2020wr028300

35 Macdonald, E. *et al.* Event and Catchment Controls of Heavy Tail Behavior of Floods. *Water Resources Research* **58**, e2021WR031260 (2022). https://doi.org:https://doi.org/10.1029/2021WR031260

36 Wietzke, L. M. *et al.* Comparative analysis of scalar upper tail indicators. *Hydrological Sciences Journal* **65**, 1625-1639 (2020). https://doi.org:10.1080/02626667.2020.1769104

37 Smith, J. A., Cox, A. A., Baeck, M. L., Yang, L. & Bates, P. Strange Floods: The Upper Tail of Flood Peaks in the United States. *Water Resources Research* **54**, 6510-6542 (2018). https://doi.org:https://doi.org/10.1029/2018WR022539

38 Villarini, G. & Smith, J. A. Flood peak distributions for the eastern United States. *Water Resources Research* **46** (2010). https://doi.org:https://doi.org/10.1029/2009WR008395

39 Yu, G., Wright, D. B. & Davenport, F. V. Diverse Physical Processes Drive Upper-Tail Flood Quantiles in the US Mountain West. *Geophysical Research Letters* **49**, e2022GL098855 (2022). https://doi.org:https://doi.org/10.1029/2022GL098855

40 Sharma, A., Wasko, C. & Lettenmaier, D. P. If precipitation extremes are increasing, why aren't floods? *Water Resources Research* **54**, 8545-8551 (2018). https://doi.org:10.1029/2018wr023749

41 Blöschl, G. *et al.* Increasing river floods: fiction or reality? *WIREs Water* **2**, 329-344 (2015). https://doi.org:https://doi.org/10.1002/wat2.1079

42 Bevacqua, E. *et al.* Guidelines for studying diverse types of compound weather and climate events. *Earths Future* **9** (2021). https://doi.org:10.1029/2021ef002340

43 Tarasova, L., Basso, S. & Merz, R. Transformation of Generation Processes From Small Runoff Events to Large Floods. *Geophysical Research Letters* **47**, e2020GL090547 (2020). https://doi.org:https://doi.org/10.1029/2020GL090547

44 Zorzetto, E., Botter, G. & Marani, M. On the emergence of rainfall extremes from ordinary events. *Geophysical Research Letters* **43**, 8076-8082 (2016). https://doi.org:https://doi.org/10.1002/2016GL069445

45 S Chegwidden, O., Rupp, D. E. & Nijssen, B. Climate change alters flood magnitudes and mechanisms in climatically-diverse headwaters across the northwestern United States. *Environmental Research Letters* **15**, 094048 (2020). https://doi.org:10.1088/1748-9326/ab986f

46 Musselman, K. N. *et al.* Projected increases and shifts in rain-on-snow flood risk over western North America. *Nat. Clim. Chang.* **8**, 808-812 (2018). https://doi.org:10.1038/s41558-018-0236-4

Methods references

47 Beck, H. E. *et al.* MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bulletin of the American Meteorological Society* **100**, 473-500 (2019). https://doi.org:10.1175/bams-d-17-0138.1

48 Beck, H. E. *et al.* MSWX: Global 3-Hourly 0.1° Bias-Corrected Meteorological Data Including Near-Real-Time Updates and Forecast Ensembles. *Bulletin of the American Meteorological Society* **103**, E710-E732 (2022). https://doi.org:10.1175/bams-d-21-0145.1

49 Seibert, J. & Vis, M. J. P. Teaching hydrological modeling with a user-friendly catchment-runoff-model software package. *Hydrol. Earth Syst. Sci.* **16**, 3315-3325 (2012). https://doi.org:10.5194/hess-16-3315-2012

50 Beck, H. E. *et al.* Global Fully Distributed Parameter Regionalization Based on Observed Streamflow From 4,229 Headwater Catchments. *Journal of Geophysical Research: Atmospheres* **125**, e2019JD031485 (2020). https://doi.org:https://doi.org/10.1029/2019JD031485

51 Xie, J., Liu, X., Bai, P. & Liu, C. Rapid Watershed Delineation Using an Automatic Outlet Relocation Algorithm. *Water Resources Research* **58**, e2021WR031129 (2022). https://doi.org:https://doi.org/10.1029/2021WR031129

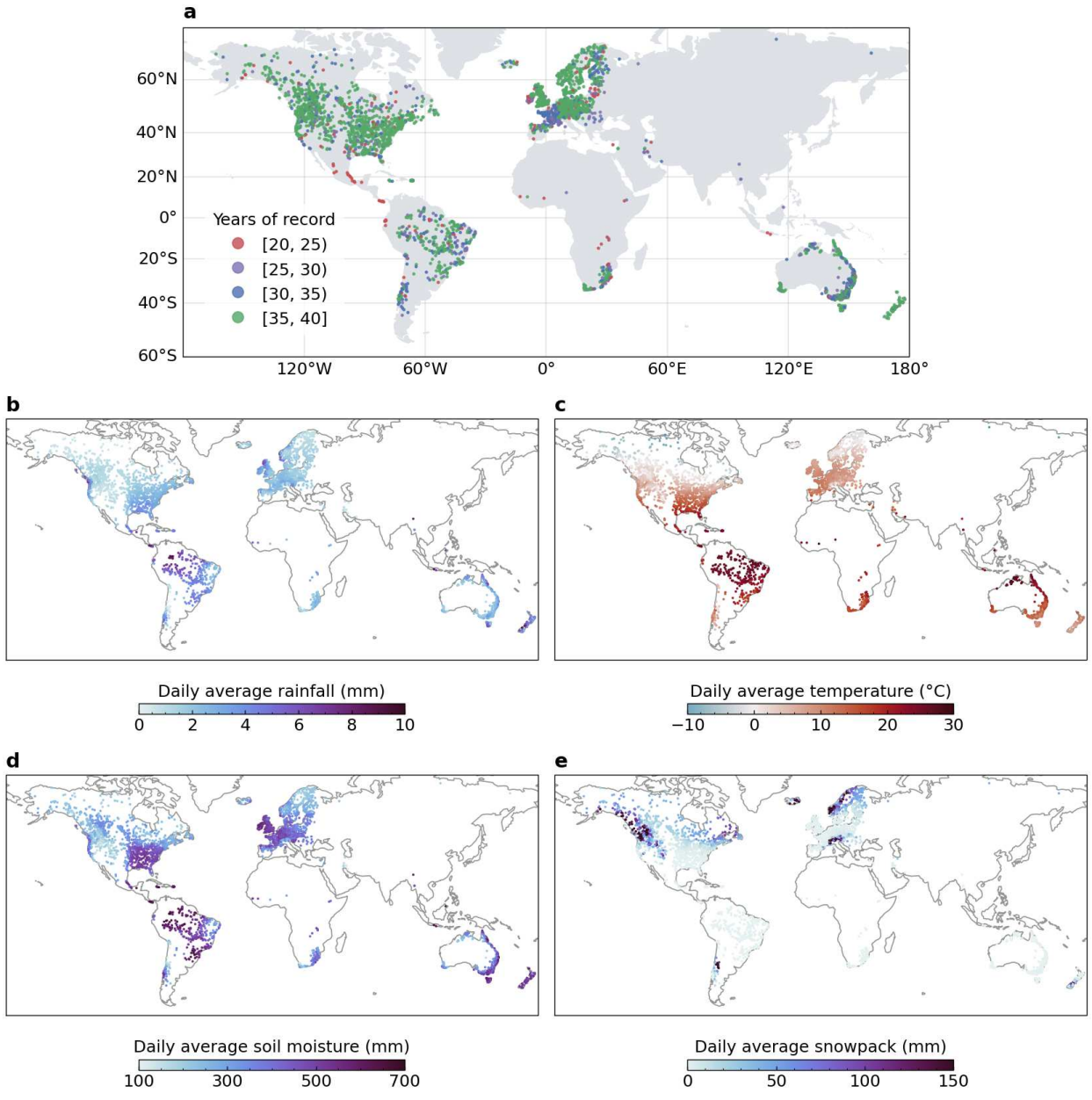
52 Linke, S. *et al.* Global hydro-environmental sub-basin and river reach characteristics at high spatial resolution. *Scientific Data* **6**, 283 (2019). https://doi.org:10.1038/s41597-019-0300-6

53 Mushtaq, S., Miniussi, A., Merz, R. & Basso, S. Reliable estimation of high floods: A method to select the most suitable ordinary distribution in the Metastatistical extreme value framework. *Advances in Water Resources* **161**, 104127 (2022). https://doi.org:https://doi.org/10.1016/j.advwatres.2022.104127

54 Lang, M., Ouarda, T. & Bobee, B. Towards operational guidelines for over-threshold modeling. *Journal of Hydrology* **225**, 103-117 (1999). https://doi.org:10.1016/s0022-1694(99)00167-5

55 Mamalakis, A., Barnes, E. A. & Ebert-Uphoff, I. Carefully Choose the Baseline: Lessons Learned from Applying XAI Attribution Methods for Regression Tasks in Geoscience. *Artificial Intelligence for the Earth Systems* **2**, e220058 (2023). https://doi.org:10.1175/AIES-D-22-0058.1

662 **Extended Data**



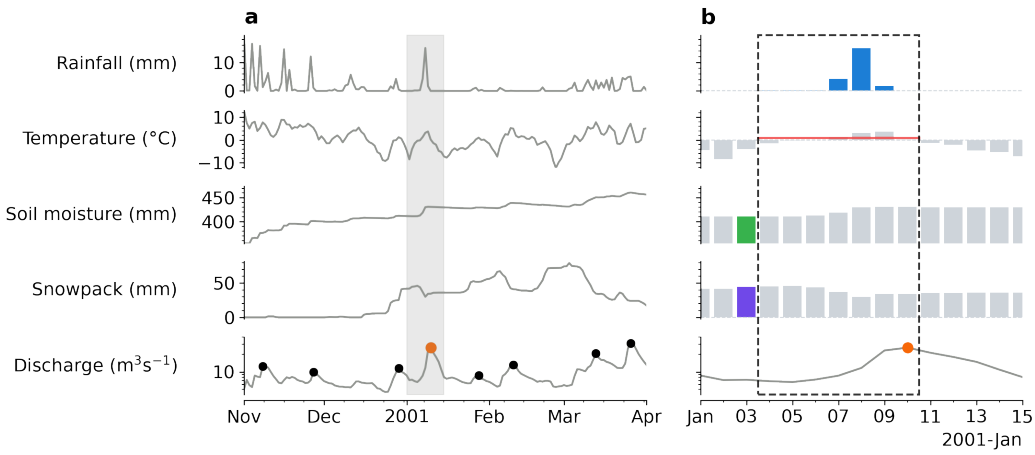
663

664 **Extended Data Fig. 1 | Locations and hydrometeorological conditions of the 3,527 catchments**

665 **used in this study. a**, the length of discharge records in individual catchments. **b-e**, The daily average

666 of rainfall (**b**), temperature (**c**), soil moisture (**d**), and snowpack (**e**) in 1981 to 2020.

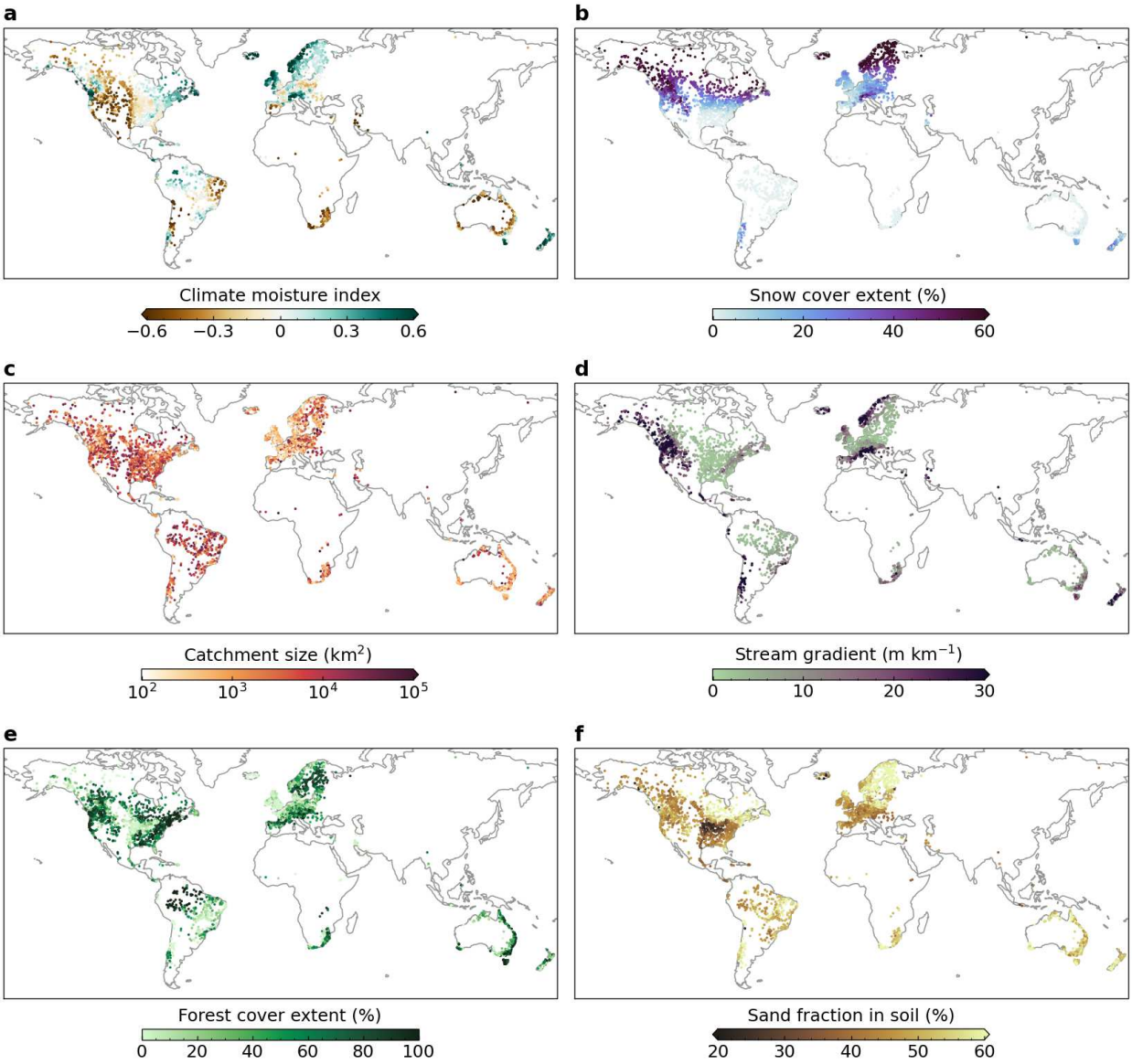
667



668

669 **Extended Data Fig. 2 | Illustration of training samples.** A catchment in Slovenia with an outlet
670 located at 49.07°N, 18.91°E is used as an example. The solid points in the lower panel in **a** show the
671 identifiable discharge peaks used as training targets. The gray shadow (zoomed in panel **b**) highlights
672 the 7-day synoptic window for determining model inputs to the model target indicated by the orange
673 point. In panel **b**, the colored bars show the input features used; for precipitation, we used the 7 days
674 before the discharge peak; for temperature, we used the average over those 7 days.

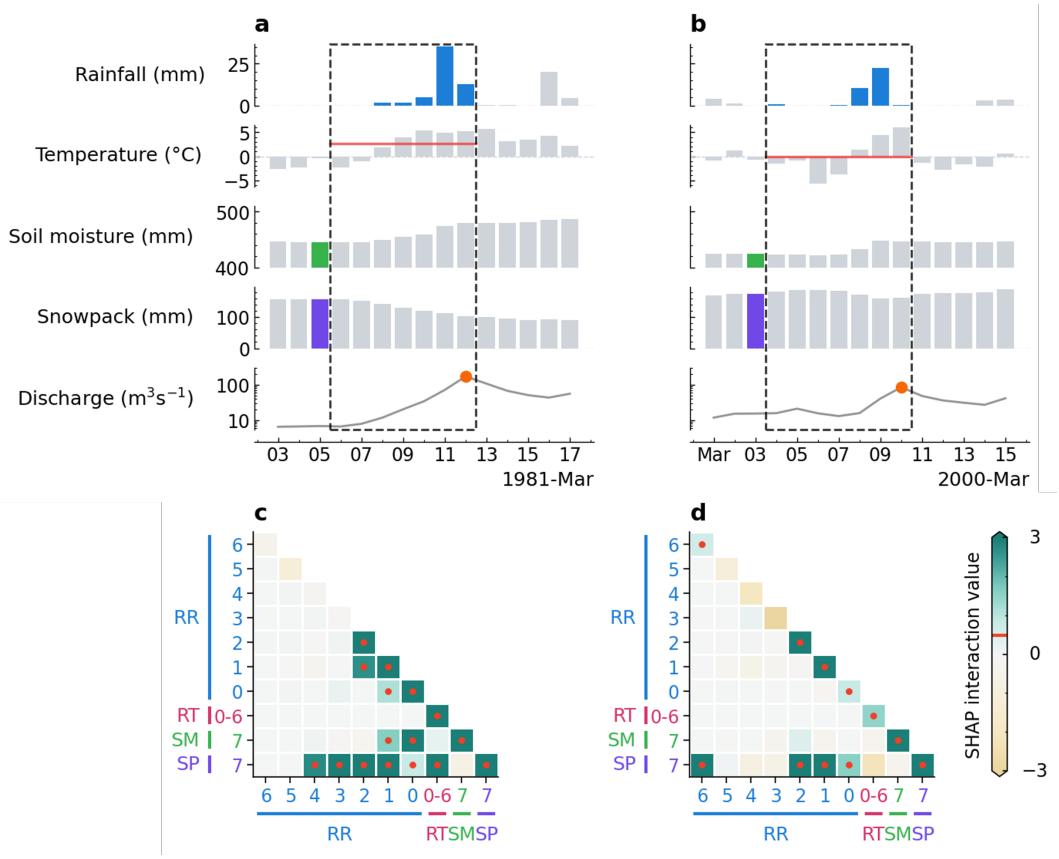
675



676

677 **Extended Data Fig. 3 | Physio-climatic attributes of the 3,527 catchments used in this study. a**,
678 Climate moisture index. **b**, Snow cover extent. **c**, Catchment size, which is estimated based on the
679 catchment boundary. **d**, Stream gradient. **e**, Forest cover extent. **f**, Sand fraction in soil.

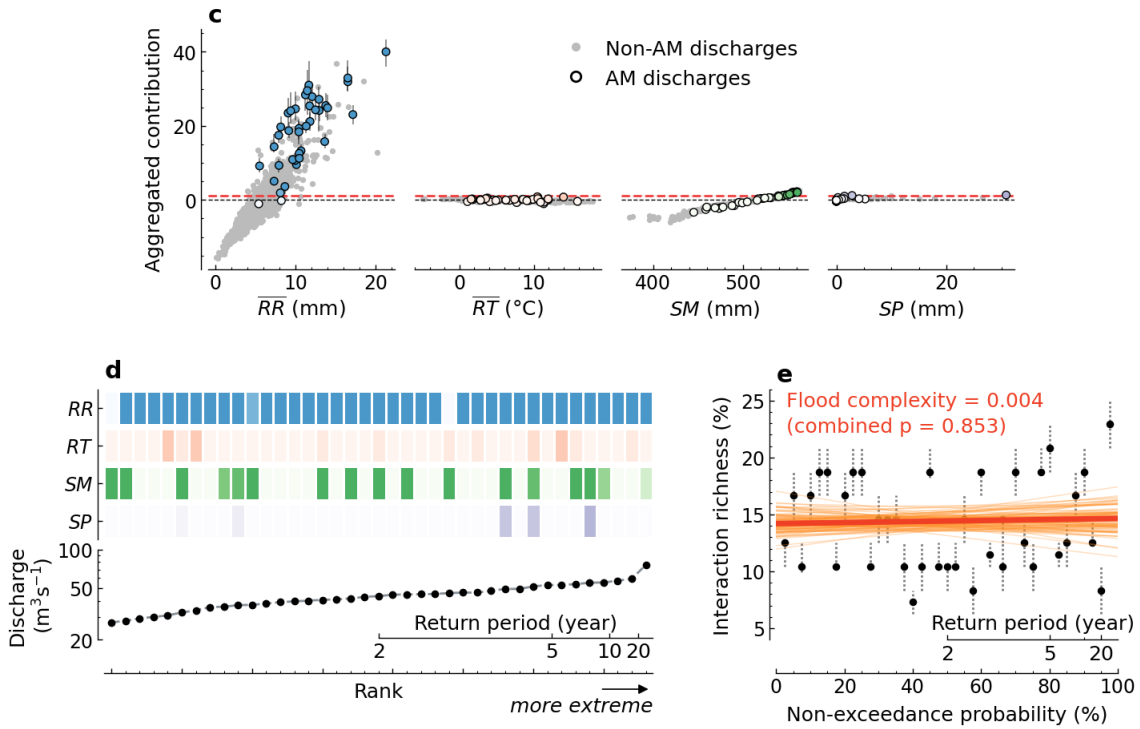
680



681

682 **Extended Data Fig. 4 | Illustration of the interaction richness of two flood event samples in one**
683 **catchment. a-b**, The input features (colored bars) and model output (orange point) of two flood
684 samples. **c**, The pairwise interaction effects between features (including the main effects of the
685 features in the diagonal) colored by the SHAP interaction values for the flood sample in panel **a**. The
686 red dots highlight the main interactions where the SHAP interaction value exceeds the threshold
687 (indicated by the red line in the color bar). Here, the threshold is calculated as the 80th percentile of
688 the positive interaction values between features (including the main effects of the features) across all
689 the samples in the catchment. In this case, the number of main interactions is 16, so the interaction
690 richness is 16/48 = 33.3%, where 48 is the number of all potential interactions in the model (note that
691 we have disabled the interactions between the input features of rainfall and temperature in the model).
692 **d**, The pairwise interaction effects between features for the flood sample in panel **b**, for which the
693 interaction richness is 23.0%.

694



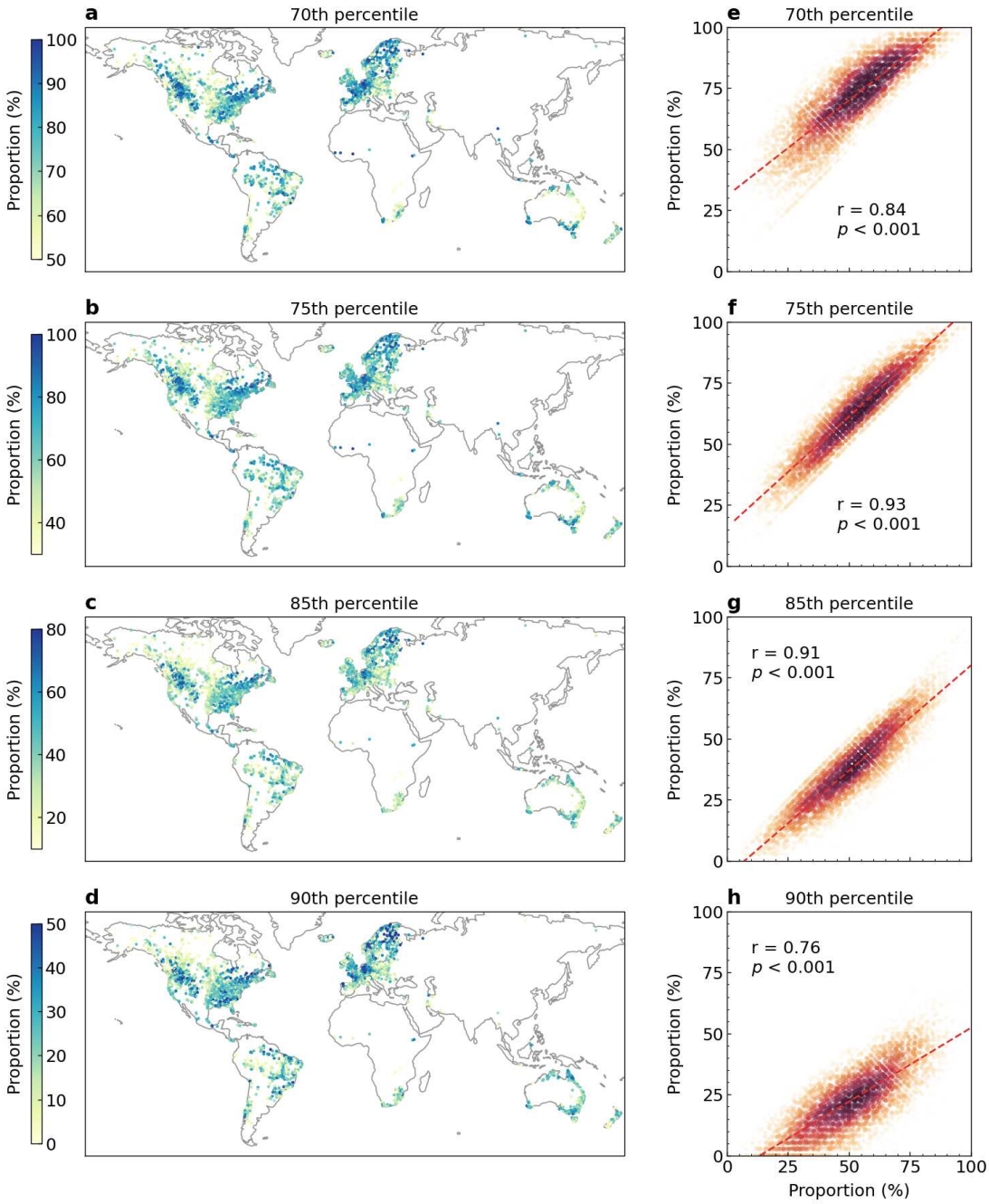
695

696 **Extended Data Fig. 5 | Compounding effects in river flood drivers.** Same as Fig. 1 but using a

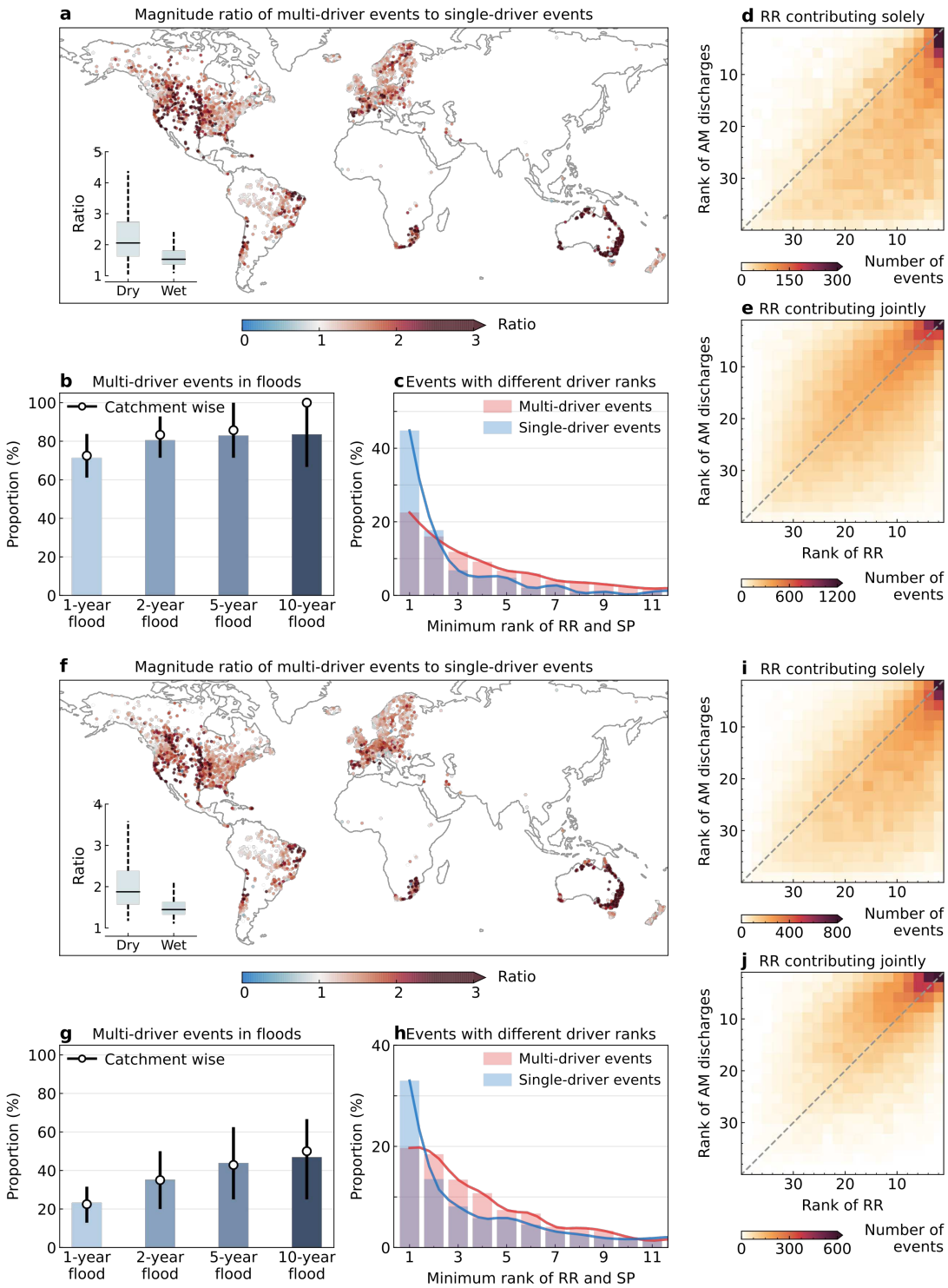
697 catchment in the United Kingdom with an outlet at 54.44°N, 3.53°W as an example.

698

699



Extended Data Fig. 6 | Sensitivity of using different thresholds to identify main drivers. a-d, Proportions of multi-driver floods when using the 70th, 75th, 85th, and 90th percentiles of the aggregated contributions as cutoffs, respectively. **e-h**, Correlations between the proportion of multi-driver floods when using the 80th percentile as the threshold (x-axis) and the proportions when using other thresholds (y-axis).

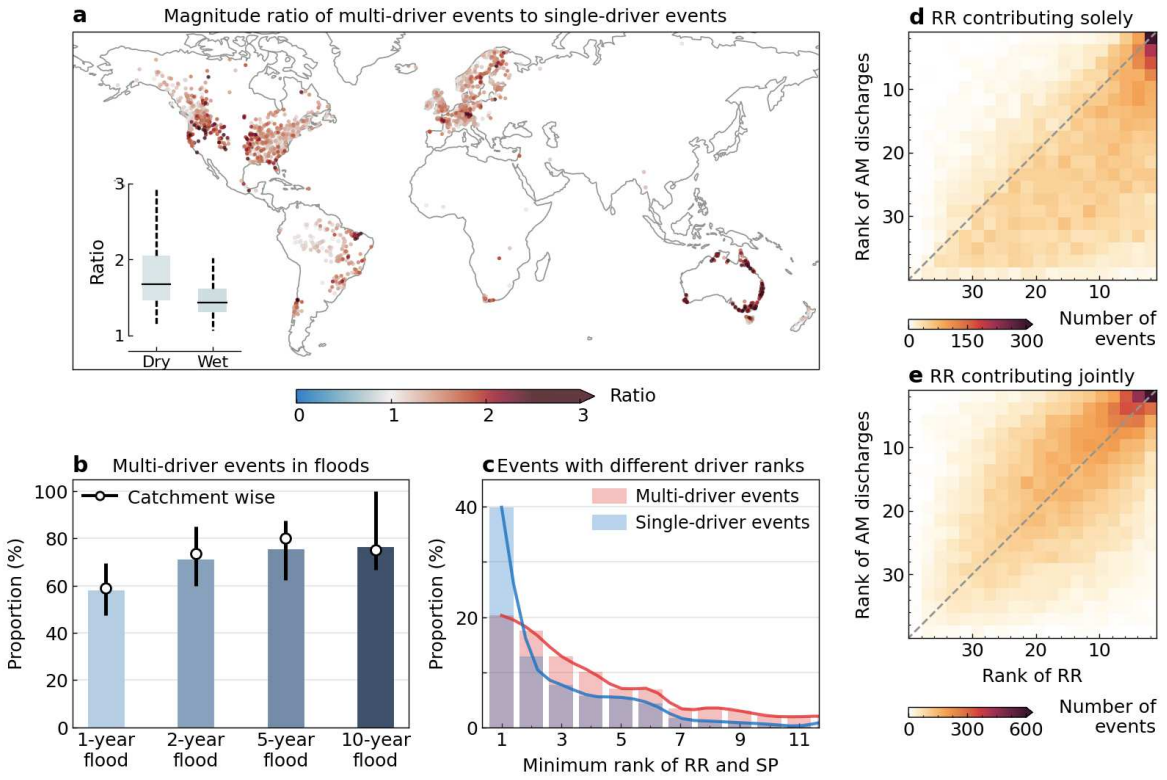


706

707 **Extended Data Fig. 7 | Impact of compounding drivers on river flood events. a-e**, Same as Fig. 3,

708 but using the 70th percentile as the threshold to identify main drivers. **f-j**, Same as Fig. 3, but using

709 the 90th percentile as the threshold to identify main drivers.



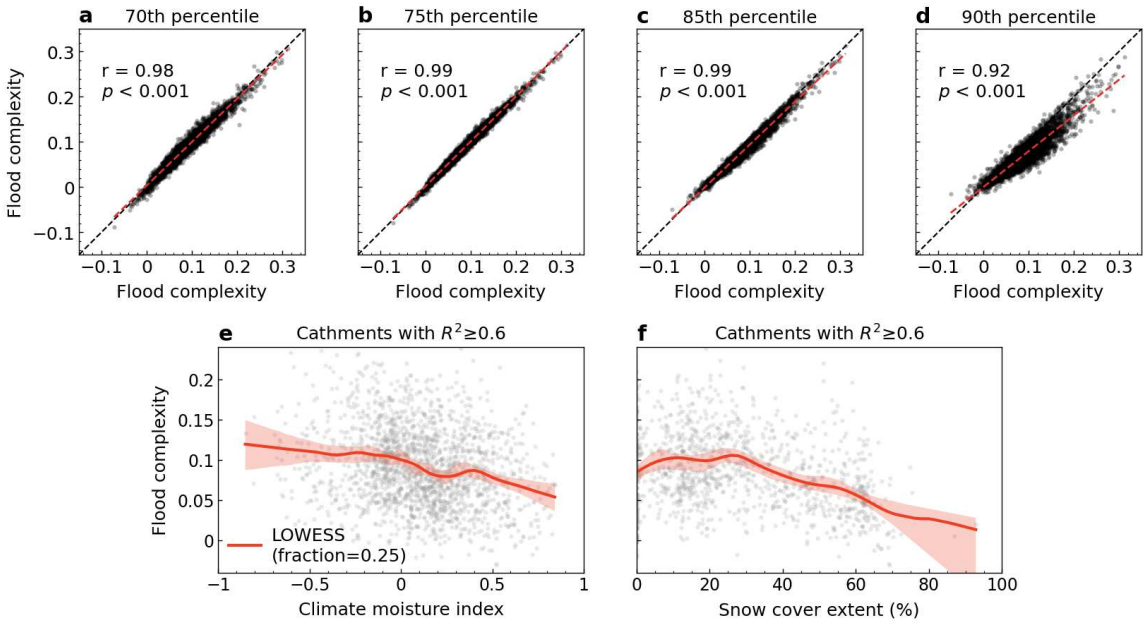
710

711 **Extended Data Fig. 8 | Impact of compounding drivers on river flood events.** Same as Fig. 3, but

712 only 1,886 catchments with the average R^2 regression value in repeated cross-validation greater than

713 0.6 were considered.

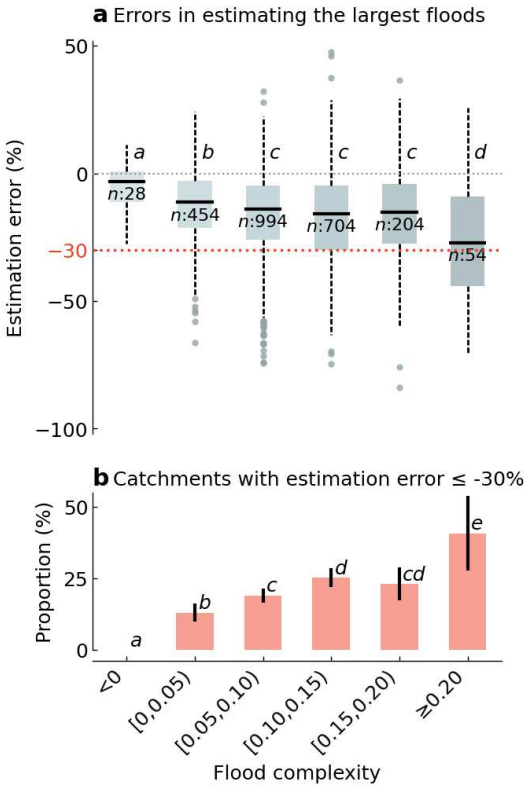
714



715

716 **Extended Data Fig. 9 | Sensitivity of using different thresholds to identify main interaction**
717 **effects and a stricter criterion for the predictive performance of ML algorithms. a-d**,
718 Correlations between the flood complexity of catchments when using the 80th percentile as the
719 threshold (x-axis) and the proportions when using other thresholds (y-axis). **e-f**, Same as Fig. 4c-d,
720 but only 1,886 catchments with the average R^2 regression value in repeated cross-validation greater
721 than 0.6 were considered.

722



723

724 **Extended Data Fig. 10 | Impact of flood complexity on estimating large flood magnitudes.** Same

725 as Fig. 5, but only 2,438 catchments with at least 35 observations of annual maximum discharge were

726 considered.

727

Extended Data Table 1 | Candidate values for hyperparameters of the LightGBM models

The hyperparameters to be determined	The meaning of the hyperparameters	Candidate values
learning_rate	Learning rate	[0.01, 0.02, 0.03, 0.04, 0.05]
n_estimators	the number of boosting iterations	[50, 100, 150, 200]
subsample	the fraction of data to be used for each iteration (tree)	[0.3, 0.5, 0.7]
colsample_bytree	the fraction of features on each iteration (tree)	[0.3, 0.5, 0.7]
max_bin	The max number of bins that feature values will be bucketed in	[8, 16, 24, 32, 64]
min_child_samples	The minimal number of data in one leaf	[3, 5, 7]