Hydrology Research

Hydrology Research Vol 54 No 10, 1181 doi: 10.2166/nh.2023.058

Unveiling flood-generating mechanisms using circular statistics-based machine learning approach without the need for discharge data during inference

Zhi Zhang, Dagang Wang*, Xinxin Wu, Yiwen Mei, Jianxiu Qiu and Jinxin Zhu

School of Geography and Planning, Sun Yat-sen University, Guangzhou, China
*Corresponding author. E-mail: wangdag@mail.sysu.edu.cn

ZZ, 0000-0002-2030-076X; DW, 0000-0002-6424-6398

ABSTRACT

Understanding the drivers of flooding is essential for flood disaster prevention. However, conventional flood prediction methods are hindered by their reliance on local discharge data, which can be constrained by limited spatial resolution. To address this limitation, we present a machine learning model that can categorize floods without requiring discharge data during inference. We first use circular statistics to calculate the relative importance of three candidate flood-generating mechanisms. Global land areas are classified into three primary categories and eight sub-categories based on the proportion of relative importance. A random forest model is then applied to identify the flood types by assuming that the discharge data is unavailable. The findings from circular statistics highlight that globally, soil moisture excess is the most influential driver of floods followed by extreme precipitation and snowmelt, with an average relative importance of 0.535, 0.387, and 0.078, respectively. The RF model performs well in resembling the three primary flood categories with an accuracy of 0.701 and a F1-score of 0.692 in 10-fold cross-validation. The trained gridded-based model provides a swift and efficient approach for analyzing flood mechanisms, even in limited discharge scenarios, allowing for rapid insights.

Key words: circular statistic, flood mechanism, flood type prediction, limited discharge data, random forest

HIGHLIGHTS

- We developed a global machine learning model for predicting flood mechanisms, utilizing precipitation, soil moisture, and snowmelt characteristics as flood drivers – without the need for discharge data.
- We applied circular statistics to ascertain the flood generation mechanisms.
- The freely available trained model is adaptable to any scale, serving as a fast and practical tool for analyzing flood mechanisms.

1. INTRODUCTION

Flooding is one of the most destructive natural disasters in the world, which leads to serious fatalities and massive property losses (Jonkman 2005; Dottori *et al.* 2018; Koç & Thieken 2018). In the context of climate warming, the frequency and intensity of extreme precipitation are expected to increase, thereby increasing the risk of flooding (Kendon *et al.* 2014; Alfieri *et al.* 2015; Mallakpour & Villarini 2015; Arnell & Gosling 2016; Esposito *et al.* 2018; Yin *et al.* 2018; Hounkpè *et al.* 2019). The impact of climate change on flood characteristics is strongly influenced by the specific and nuanced details of global changes (Kundzewicz *et al.* 2019), which encompass a complex interplay of factors such as changes in precipitation patterns, temperature, land use, and soil moisture regimes, among others. Understanding the main drivers of flooding is thus essential for flood hazard mitigation and water resources management (Winsemius *et al.* 2016; Keller *et al.* 2018).

Extensive studies have been conducted on changes in the frequency of flood events (e.g., Slater & Villarini 2016; Wu & Qian 2017; Najibi & Devineni 2018). On this basis, there is a growing interest in attributing floods to their causal factors, which is crucial to gaining a better understanding of the underlying physical mechanisms (Neri *et al.* 2019). The Budyko framework (Budyko 1974) is frequently employed to assess the sensitivity of hydrological components such as discharge, evapotranspiration, and soil moisture to various impact factors (Gudmundsson *et al.* 2016, 2017). Patterson *et al.* (2013) used Budyko curves to ascribe changes in streamflow and found that human impacts were equivalent to or even greater than

climate impacts over 27% of the South Atlantic basins, USA. Berghuijs *et al.* (2017) developed an improved Budyko-based method to assess the sensitivity of average annual runoff, and the results indicate that runoff is the most sensitive to changes in precipitation for 83% of land areas of the globe. Milly *et al.* (2018) revealed that temperature can affect flood peaks through evapotranspiration and the snow-related processes based on extensions of the Budyko water-balance hypothesis. Statistical methods are also used to identify flood-generating schemes in complement with the Budyko framework.

However, the Budyko framework is limited by assuming a linear relationship between precipitation and evaporation, which may not hold true for certain climatic conditions in watersheds. Additionally, the framework only considers mean annual precipitation and evapotranspiration, ignoring their temporal variabilities. To address these limitations, many studies employed various methods, such as incorporating nonlinear relationships into models, considering seasonal changes in water balance, and utilizing the remote sensing data to better capture the spatiotemporal variability. Slater & Villarini (2017) modeled streamflow as a function of precipitation, temperature, antecedent wetness, agricultural land cover and population using the Generalized Additive Model for the U.S. results showed that rainfall is the dominant source of discharge in rain-on-snow events along the West Coast, Cascades and the Sierra Nevada, while snowmelt dominates streamflow in the other regions. Kundzewicz *et al.* (2019) used the Generalized Extreme Value distribution to assess flood risk caused by rain-on-snow events over the conterminous United States. Circular statistics (Burn 1997) are widely applied to investigate the consistency of hypothesized flood-generating processes and maximum annual flow (Berghuijs *et al.* 2016; Slater & Villarini 2017; Berghuijs *et al.* 2019). Process-based models were also adopted to investigate the factors affecting discharge characteristics. For example, Wang *et al.* (2017) found that the human-induced changes in land use/land cover can alter the pattern and the magnitude of streamflow based on the Soil and Water Assessment Tool (SWAT) model.

Discharge data are indispensable when the aforementioned statistical methods or process-based models are used to identify flood-generating mechanisms. However, it is important to note that such data may not always be readily available. Discharge data are typically obtained from monitoring stations or derived from models, which can limit the spatial resolution of the data. In situations where critical variables are missing, traditional statistical methods are often unable to perform effective analysis. Machine learning algorithms use the power of big data to learn and optimize through automatic learning and iteration, enabling them to evaluate the relationships between various factors and perform effective prediction and analysis even in the absence of certain variables. Parajka *et al.* (2010) identified the different seasonal characteristics of flood regimes by the k-mean clustering technique using annual maximum of flow and precipitation. Curran & Biles (2021) estimated the metrics of mean monthly streamflow to guide the identification of seasonal flow drivers by principal component analysis. Stein *et al.* (2021) employed a random forest (RF) model with accumulated local effects to investigate the influence of climate and catchment attributes on the spatial distribution of flood-generating processes.

In this study, we present a machine learning framework based on circular statistics for determining flood-generating mechanisms even in situations with limited discharge data. This study seeks to evaluate the potential of machine learning in identifying flood-generating mechanisms without relying on discharge information, contributing to a more effective and comprehensive approach to understanding and managing flood risk.

2. DATA

We utilized hydrological variables from the Catchment Land Surface Model of the Global Land Data Assimilation System (GLDAS) version 2.0 at the 0.25° spatial scale, including precipitation, soil moisture, and snowmelt to maintain consistency among these variables and their inherent interrelations. Meanwhile, the discharge data is obtained from the Global Reach-scale A priori Discharge Estimates for Surface Water and Ocean Topography (GRADES), as GLDAS does not have this variable. GRADES is a model-derived 0.25° daily discharge database based on the hydrography data from Multi-Error-Removed-Improved-Terrain Hydro. The modeling chain used to derive the dataset includes the Variable Infiltration Capacity land surface model and Routing Application for Parallel computatIon of Discharge river routing model. More detailed information about GRADES can be found in Lin *et al.* (2019). A 35-year period from 1979 to 2013 was considered to provide a sufficiently long temporal overlap between GLDAS and GRADES.

3. MATERIALS AND METHODS

Schematics of the framework proposed to identify and predict the flood-generating mechanisms are shown in Figure 1. The first component employs circular statistics to estimate the relative importance among precipitation, soil moisture, and

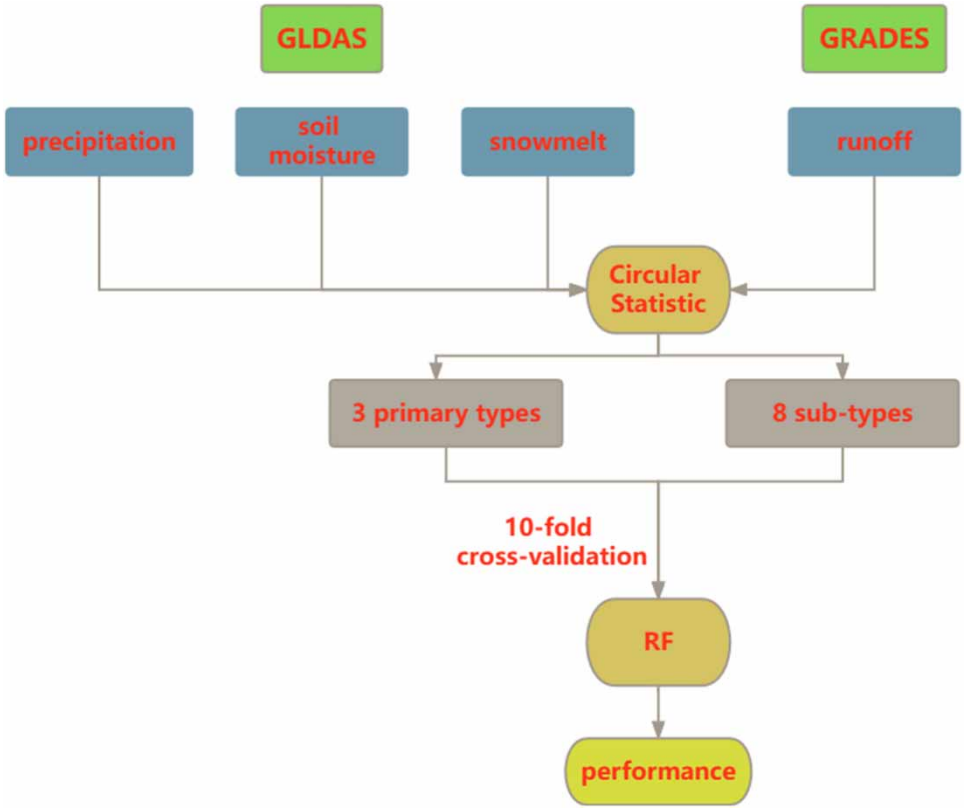

Figure 1 | Workflow diagram of the proposed method, including estimation of flood types based on circular statistics, and prediction of these types based on the RF model.

snowmelt (Section 3.1). Based on the relative importance, three primary flood types and eight sub-types are defined (Section 3.2) and are adopted as the reference during the model training process. Then, a RF classification model was trained to predict the flood types (Section 3.3). Apart from this, evaluation metrics are introduced in Section 3.4.

3.1. Circular statistics

The definition of flooding occurrence of each year is the maximum annual discharge. Three major mechanisms, namely precipitation, soil moisture and snowmelt, are considered the causes of the flooding occurrence following Berghuijs *et al.* (2019). For the first mechanism, floods are caused by extreme precipitation events. In other words, the maximum annual flow (MAF) and the largest precipitation event occurred at the same time of the year. Regulated by the second mechanism, the MAF occurred simultaneously with the largest soil moisture event. When the soil moisture is high, even light rainstorms can cause floods. This hypothesis suggests that the antecedent soil moisture storage controlled by seasonal rainfall and evaporation is the primary controlling factor for the streamflow generation in a flood event. Under the last mechanism, the floods are caused by the largest snowmelt event, which means that the occurrence dates for the MAF and the largest snowmelt are the same. The proposed mechanisms serve as the basis for our analyses on the drivers of flood generation.

Following the defined mechanisms, circular statistics is used to compute the seasonality of each process (i.e., flood, extreme precipitation, excess soil moisture and snowmelt). Circular statistics is a branch of statistics that deals with data that have periodic properties. The basic idea of circular statistics is to represent circular data using a two-dimensional vector. In the context of this study, the seasonality (mean date of occurrence) for each process was determined using circular statistics and expressed as a two-dimensional vector (r, θ). The length of the vector r represents the strength or magnitude of the process, while the direction of the vector θ represents its phase or timing within the circle. The calculation of (r, θ) is not commonly done using conventional methods, since it is expressed in polar coordinates. Therefore, it is typically converted

to the Cartesian coordinate system ($\overline{x_i}$, $\overline{y_i}$) and can be computed using the following formula:

$$\overline{x_i} = \frac{1}{n}\sum_{j=1}^{n}\cos\left(d_{i,j} \times \frac{2\pi}{m_j}\right) \tag{1}$$

$$\overline{y_i} = \frac{1}{n}\sum_{j=1}^{n}\sin\left(d_{i,j} \times \frac{2\pi}{m_j}\right) \tag{2}$$

The term $d_{i,j}$ is the date of occurrence of process i in each individual year j; m_j equals 365 days (366 for leap years), and n is the number of years of the data (35 years for this study).

If floods are caused by these three mechanisms, the sum of their cosine and sine components of the seasonality vector $\overline{x_i}$ and $\overline{y_i}$ should be approximately equal to those of flooding (Berghuijs *et al.* 2019). The relative importance of each flood driver can be solved by the following set of linear equations:

$$\alpha_p\overline{x_p} + \alpha_m\overline{x_m} + \alpha_s\overline{x_n} = \overline{x_f} \tag{3}$$

$$\alpha_p\overline{y_p} + \alpha_m\overline{y_m} + \alpha_s\overline{y_n} = \overline{y_f} \tag{4}$$

$$\alpha_p + \alpha_m + \alpha_n = 1 \tag{5}$$

where α_i denotes the weighting contribution of each flood driver (with $0 \leq \alpha_i \leq 1$). An α_i value of 0 indicates that the mechanism does not influence flooding, whereas an α_i value of 1 signifies that flooding is totally caused by the mechanism. More detailed information about circular statistics can be found in Bloschl *et al.* (2017) and Berghuijs *et al.* (2019).

3.2. Flood type definition

Three different types of flooding occurrences are classified by the calculated relative importance: P type (dominated by precipitation), M type (dominated by soil moisture excess), and N type (dominated by snowmelt). The dominant type is defined by the flood-generating mechanism with the highest relative importance.

Besides the dominant mechanism, other mechanisms may also regulate flood generations for a substantial portion of flooding occurrences, which leads to multiple mechanisms in areas of interest. Therefore, a sub-type flooding classification is needed to better understand the complex hydrological processes. Hence, to classify flood sub-types, we employed the k-means clustering algorithm. The relative importance (α_p, α_m , and α_n) for each grid will be input into the k-means clustering model. The model is implemented by the Scikit-learn software package to classify flood mechanisms into sub-types. To determine the optimal number of flood sub-types, two contour metrics, Silhouette Coefficient (SC) and Calinski-Harabasz Index (CI), were calculated to evaluate the performance of the clustering. SC measures the average distance of each sample from other samples, while CI assesses the covariance between different classes. In simpler terms, higher SC and CI values indicate that the clustering results are more effective and accurate.

3.3. Flood type identification model

An RF model is trained to predict the flood type. RF is a supervised ensemble learning algorithm to solve problems by creating and combining regression trees (Breiman 2001). It can tackle high dimensional nonlinear relationship (Addor *et al.* 2018), which is suitable for the investigation of flood generation with other factors. A more detailed description of the RF algorithm can be found in (Wiener 2002). This model has been widely applied into the hydrometeorological field (Turini *et al.* 2019; Xu *et al.* 2019; Zhang & Shi 2019).

The inputs of the RF model are set as longitude, latitude, and the occurrence day of the highest precipitation/soil moisture excess/snowmelt of each year during the period of 1979–2013, and the output is the flood type. The selection of the predictor variables is crucial for gaining insights into the factors that contribute to the occurrence of floods. The inclusion of longitude and latitude enables the model to identify the location and its associated flood type with great accuracy. This is because different areas are prone to different types of flooding due to their unique topographical and geographical characteristics. In addition, floods are often caused by excessive water exceeding the holding capacity of soil due to heavy rainfall, rapid snowmelt, and high soil moisture excess. By incorporating the dates on which the highest precipitation, soil moisture excess and snowmelt occurred in the model, it becomes possible to isolate the underlying causes of the flooding event and better predict

the expected flood type. However, the original information of the occurrence days of the highest precipitation, soil moisture, and snowmelt during a period need to be transformed before being inputted into the model. These days are represented as numbers between 1 and 365 (366 for leap years) and their statistics are calculated, including minimum, maximum, mean, standard deviation, range, peak-to-peak, and 7%iles from 20th to 80th. Anomaly for each percentile is also calculated by subtracting the mean from the percentile value. Consequently, 20 statistics are generated for each variable of precipitation, soil moisture, and snowmelt. In total, 62 features are fed into the model for training and prediction.

To minimize spatial contiguity between the test and training sets, a 10-fold cross-validation with a longitude-based partition was employed (e.g., areas between 0°E and 36°E for testing and other areas for training) instead of a random split. This method was implemented to ensure that the model's performance accurately reflects its generalizability to regions that are geographically distant from the training data. The dataset is divided along longitude instead of latitude to ensure sufficient samples for each of the flood types in both training and testing datasets and therefore increases robustness of the RF mode. By partitioning based on longitude, the training and testing sets are more representative of the geographical diversity of the region under investigation, thereby potentially increasing the robustness of the findings.

3.4. Evaluate metrics

Four metrics based on the confusion matrix were used to evaluate the classification performance. These metrics are precision, recall, F1-score and accuracy. Accuracy is defined as the ratio of correctly classified samples to the total number of samples in the test set, while precision represents the ratio of true positives to all predicted positive samples, and recall represents the ratio of true positives to all actual positive samples. The F1-score represents a weighted average of precision and recall, which evaluates the balance between these two measures. It should be noted that for each of these four metrics, an optimal value is 1, representing ideal performance of the classifier. They are calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{6}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{7}$$

$$\text{F1\_score} = 2\frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}} \tag{8}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}} \tag{9}$$

where for the *i*th category, True Positive (TP) represents the number of samples in the *i*th category that were correctly predicted to belong to that category; False Positive (FP) represents the number of samples not in the *i*th category that were incorrectly predicted to belong to that category; False Negative (FN) represents the number of samples in the *i*th category that were incorrectly predicted as not belonging to that category; and True Negative (TN) represents the number of samples not in the *i*th category that were correctly predicted as not belonging to that category.

Furthermore, the feature importance of the RF model is calculated based on how much each feature contributes to the reduction in impurity when building the decision trees that make up the forest. When building an RF, each decision tree is constructed using a randomly selected subset of features. This helps to prevent overfitting and increase accuracy. For each decision tree, the importance of each feature is calculated by measuring how much it reduces impurity across all of its nodes. The resulting values are then averaged across all trees in the forest to obtain the final feature importance. This feature importance can be used to identify which features are most important for making accurate predictions with the RF model. This information can be useful for understanding the underlying dynamics of the data being modeled.

4. RESULTS

4.1. Spatial distribution of relative importance of the three flood-generating mechanisms

The spatial distribution of relative importance of the three flood-generating mechanisms (extreme precipitation, soil moisture excess, and snowmelt) is presented in Figure 2. In total, 234,522 grids of data are obtained for the global land. Note that grids

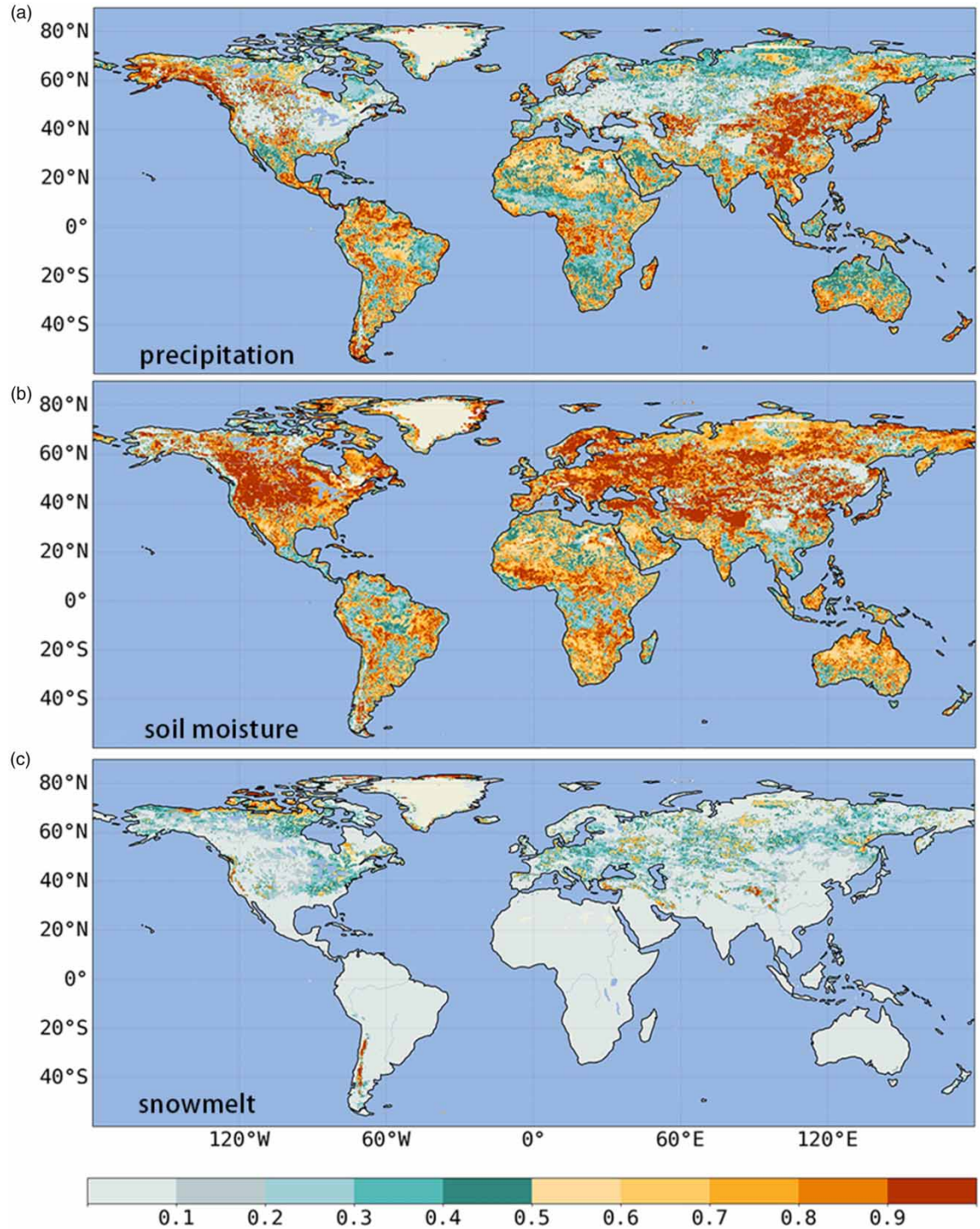

Figure 2 | Spatial pattern of the relative importance of (a) extreme precipitation floods, (b) soil moisture excess floods, and (c) snowmelt floods over the period 1979 to 2013. Grids covered by water and less than 20 years GRADES or GLDAS data are set to be the background color (ocean blue).

with GRADES or GLDAS data less than 20 years are excluded from the calculation, and the corresponding color is set to the background color (ocean blue) in Figure 2.

Globally, maximum annual precipitation is an important flood driver with the global mean of 0.387 (Figure 2(a)). In this study, a mechanism driver is deemed to be the overwhelming (respectively, insignificant) one if its relative importance score

exceeds 0.8 (respectively, falls below 0.2). We adopted the threshold values (0.8 and 0.2) by referring to several studies, including Wang *et al.* (2021), Kim *et al.* (2018), Nguyen *et al.* (2017), and Zhou *et al.* (2019). Extreme precipitation takes the overwhelming control ($\alpha_p \geq 0.8$) on flood generation for 12% of the global land area. These areas are primarily located in and near Southeast Asia, Central Africa, Northern South America, Southern North America, and Northwestern North America. In contrast, maximum annual precipitation is not a significant flood driver ($\alpha_p \leq 0.2$) for approximately one-third of the globe; these areas are mainly located in Europe, the U.S., and Western Asia.

Among the three flood-generating mechanisms, soil moisture excess is the most important flood driver with the global mean of 0.535 (Figure 2(b)). This driver overwhelms ($\alpha_m \geq 0.8$) flood generation in nearly one-quarter of global land areas, which exceeds the areas overwhelmed by the other two drivers combined. These regions are predominantly located in Western Asia, Northern Asia, Europe, Central America, and Central Africa. Soil moisture excess is an insignificant mechanism ($\alpha_m \leq 0.2$) for flooding in such areas as Southern Asia, Northeast Asia, Central Africa, and Northwestern North America, which accounts for 20% of global land areas.

Although snowmelt has a low global average value of 0.078 as a flood driver, its significance can vary greatly depending on location and local conditions (Figure 2(c)). Therefore, while it may not be the most significant flood driver globally, it should not be overlooked in specific regions. Flood is overwhelmingly controlled by snowmelt in a few areas, and the relative importance is insignificant ($\alpha_n \leq 0.2$) in nearly 85% of global land areas. Snowmelt has a negligible impact on the flood formation in low-latitude regions and the Southern Hemisphere. By contrast, it has a more significant influence in the areas with high latitudes and high altitudes, such as Southwestern Africa, along the west coast of the U.S. (the Cascade Range and Sierra Nevada Mountains), Northern North America, and High Mountain Asia.

The above results are similar to the studies of Berghuijs *et al.* (2019) in Europe, and of Berghuijs *et al.* (2016) in the United States with some differences. For example, a lower relative importance of snowmelt (0.4–0.5) in Northeastern Europe is found in this study, whereas Berghuijs *et al.* (2019) concluded that 31% of the catchments in Northeastern Europe are dominated by snowmelt with the relative importance larger than 0.5. The difference may be explained by the following: (1) Data sources such as the GRADES and GLDAS data are used in this study, whereas the European Flood Database (Hall *et al.* 2015) is adopted in Berghuijs *et al.* (2019) and (2) we modeled snowmelt using the GLDAS while the snowmelt described in the study by Berghuijs *et al.* (2019) was defined as the sum of daily liquid precipitation and snowmelt during the melting days.

4.2. Primary classification of flood types and its identification

For simplicity, three different types are divided directly as P type, M type, and N type. Such classification is to identify the most important driver whose proportion is the largest among the three mechanisms in controlling the flood occurrence. The distribution of each flood type is shown in Figure 3(a). Globally, soil moisture excess is the most important driver, as it accounts for 59% of total land areas. The M type is mainly concentrated in the middle of each continent. The other two drivers, extreme precipitation and snowmelt, account for 36 and 5%, respectively. Most P-type areas are the tropical regions under the control of the low-pressure system that causes heavy precipitation. Among other P-type areas, the west coast of North America between 40°N and 60°N is the area with strong westerly winds all year round, which leads to heavy precipitation. The same climate type causes the P type flood in the west coast between 40°S and 60°S of South America. Most of the monsoon regions in Asia also belong to the P type as the Eastern and Southern Asia Monsoon bring in moisture, causing heavy precipitation. The N type is mainly distributed in high latitudes of North America and Asia. It also occurs in mountainous regions such as Tibet.

The evaluated metrics of the RF model for the primary classification are shown in Table 1. Due to the large difference in the proportion of each type, the weighted average values are used to assess the overall performance. The RF model shows a precision value of 0.692 and a recall value of 0.701. The F1-score, the harmonic average of precision and recall also reaches 0.692. Of the three types of mechanisms, the RF model performs the best for the M type followed by the P type and the N type in terms of the F1-score value. It is more difficult to accurately identify the types of P and N compared to the M type due to the lower number of training samples. Nevertheless, the RF model maintains an accuracy of 0.701 for all three types.

Figure 3 shows the spatial distributions of three flood types estimated on the basis of the circular statistics and predicted by the RF model in the 10-fold cross-validation. In general, the RF model can capture the spatial patterns of the three flood types. Some areas identified as M type by the circular statistics are classified as P type by the RF predictions. Another evident discrepancy between the circular statistics-based estimations and the RF predictions is that many areas belonging to the N type

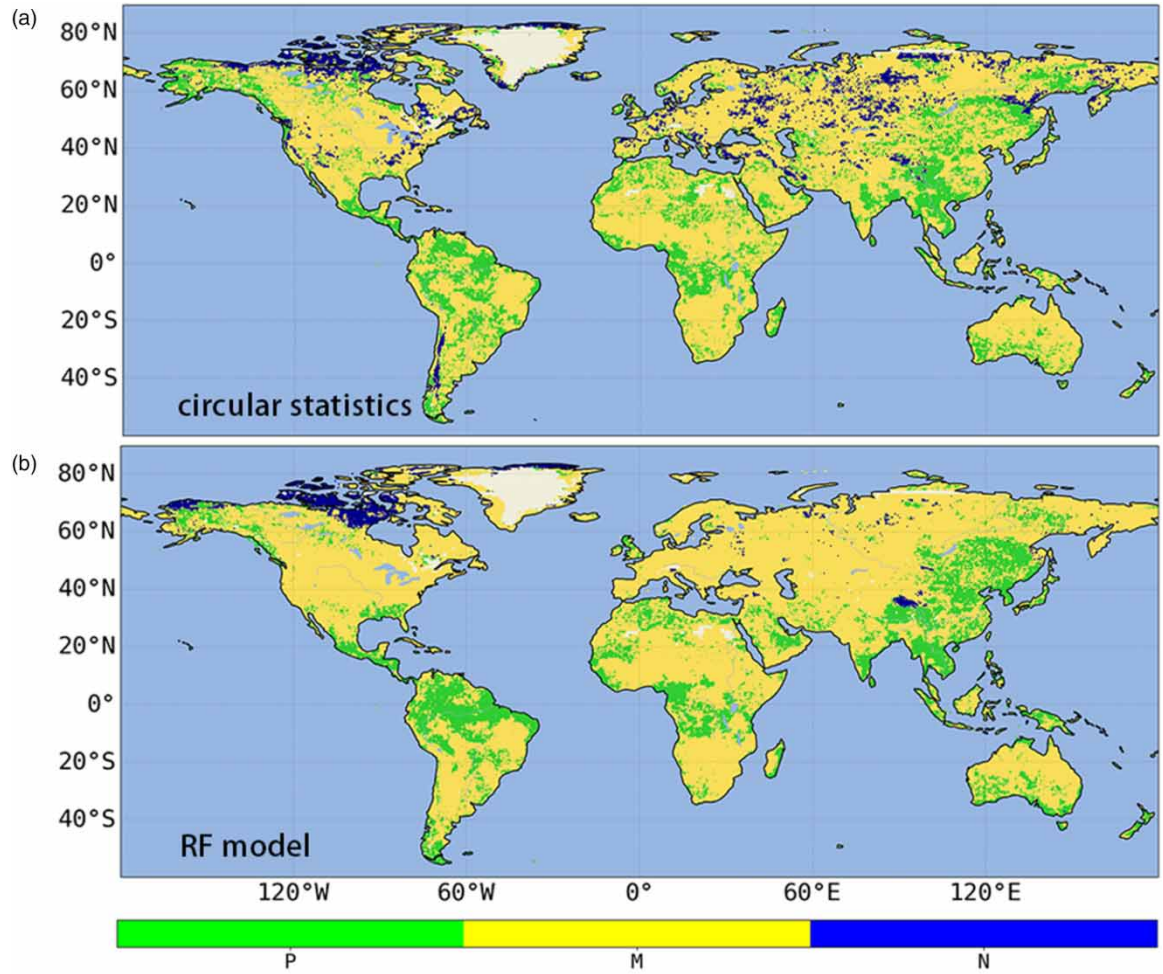

Figure 3 | The global distribution of P-type, M–type, and N-type floods from (a) circular statistics-based estimations and (b) the RF model in the 10-fold cross-validation.

Table 1 | The RF model performance from the test periods in the 10-fold cross-validation and the circular statistics of the relative importance in identifying primary type flood mechanisms

	RF model performance			Relative importance			
Primary type	**Precision**	**Recall**	**F1-score**	**Extreme precipitation**	**Soil moisture excess**	**Snowmelt**	**Number of grids**
P	0.647	0.570	0.606	0.728 (0.340–1.000)	0.220 (0.000–0.499)	0.053 (0.000–0.500)	84227 (36%)
M	0.734	0.818	0.773	0.195 (0.000–0.500)	0.760 (0.333–1.000)	0.046 (0.340–1.000)	138953 (59%)
N	0.510	0.249	0.334	0.216 (0.000–0.499)	0.125 (0.000–0.499)	0.659 (0.347–1.000)	11342 (5%)
Average	0.692	0.701	0.692	–			
Accuracy	0.701						

The values outside parentheses are the averages over the grids. The values inside parentheses are the minimum and maximum.

in North Asia are misclassified into the M type. Since the RF model does not use the discharge data, it cannot distinguish one type from another, when the conditions of two or three among precipitation, soil moisture, and snowmelt are similar. The meteorological background between adjacent grids usually maintains a high spatial continuity, whereas the discharge characteristic is more affected by the local topography. The predicted flood types from the RF model do not show a high spatial heterogeneity. The model also predicts a larger N-type area than circular statistics-based estimations in Tibet. The

complicated topography in this area substantially affects the hydrological process. The limited number of input variables used in the RF model cannot capture such a topographic-induced process, which is partially responsible for the poor model performance. In Eastern Asia, Central Africa, and Northern South America, some areas that originally belong to the M type are classified as P type by the model. This is because extreme rainfall and excess soil moisture in these areas have similar contributions to flooding (Figure 2(a) and 2(b)), which makes the model more difficult to correctly predict the flood type. For example, those areas misclassified as the P type in Northern Africa show the relative importance of extreme precipitation is 0.5–0.6 while the relative importance of soil moisture excess is 0.3–0.5.

4.3. Flood sub-type classification

It is important for an area of interest to identify the dominant flood-generating mechanism that represented by one of the three aforementioned flood types. However, only identifying the dominant one might be insufficient to understand local hydrology when other drivers also play significant roles in flood generation. To further classify flood type including the information on multiple mechanisms and their ranks of importance, we define flood sub-types under the P, M, and N types by using the k-means based aggregation. The first step is to determine how many sub-types need to be specified. The SC and CI metrics described in Section 3.2 are used to assess the clustering effect. Figure 4 shows the SC and CI values for different number of clusters. After analyzing the data, a total of eight categories were chosen as (a) the SC value peaks at eight and (b) the CI value remains stable after 8. Table 2 shows the details of the eight flood types and Figure 5(a) shows their spatial distributions. Specifically, the clusters are named using a two-part system that specifies the key variables used in the clustering process. The first part of the name indicates the primary variable of importance, while the second part denotes the secondary variable. If the secondary variable has relatively little impact, an asterisk (*) is included in its place. It is noted that if two factors are of approximately equal importance, they will be grouped together in the same part of the subcategory name. Any lower priority variables are omitted from the name to ensure clear and consistent interpretation of the clustering results.

Specifically, the P type can be further classified into P-* type, P-M type, and PN-* type. The P-* type indicates extreme precipitation overwhelmingly dominates the flood generation, while soil moisture excess and snowmelt play little role in the process. The areas with this sub-type are rare, and they are mainly distributed in southwest China, north South America and central Africa. The P-M type is mainly distributed in southern Asia, central Africa and northern South America. The floods occurring in these areas are most likely influenced by extreme precipitation and soil moisture excess. It is interesting to find that the P-* type is often surrounded by the P-M-type areas. Different from the P-M type, the PN-* type represents that

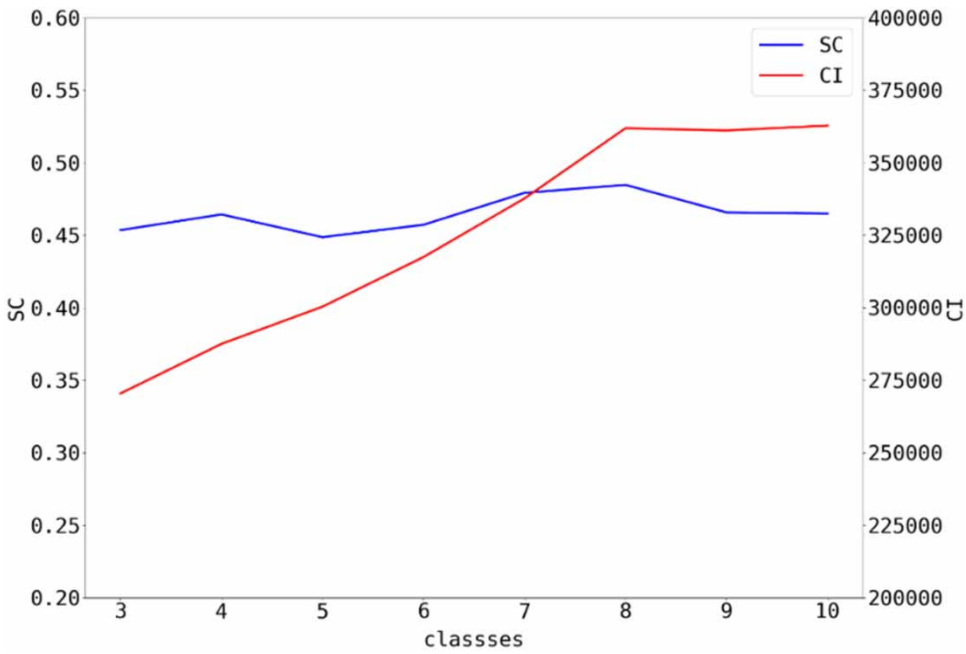

Figure 4 | Performance with different classes when the K-means cluster algorithm is used to classify flood sub-types.

Table 2 | The RF model performance from the test periods in the 10-fold cross-validation and the circular statistics of the relative importance in identifying sub-type flood mechanisms

	RF model performance			Relative importance			
Sub-type	**Precision**	**Recall**	**F1-score**	**Extreme Precipitation**	**Soil Moisture Excess**	**Snowmelt**	**Number of grids**
P-*	0.488	0.407	0.444	0.908 (0.682–1.000)	0.049 (0.000–0.192)	0.043 (0.000–0.257)	30434 (13%)
P-M	0.388	0.332	0.358	0.678 (0.444–0.807)	0.319 (0.119–0.425)	0.002 (0.000–0.263)	29819 (13%)
PN-*	0.556	0.405	0.469	0.546 (0.261–0.742)	0.007 (0.000–0.335)	0.448 (0.214–0.657)	12402 (5%)
M-PN	0.501	0.580	0.538	0.032 (0.000–0.156)	0.938 (0.757–1.000)	0.031 (0.000–0.195)	50307 (21%)
M-P	0.430	0.443	0.437	0.269 (0.098–0.369)	0.729 (0.530–0.843)	0.002 (0.000–0.218)	41731 (18%)
M-N	0.428	0.274	0.334	0.008 (0.000–0.291)	0.648 (0.322–0.804)	0.344 (0.157–0.599)	16356 (7%)
MP-*	0.483	0.587	0.530	0.468 (0.261–0.574)	0.529 (0.297–0.630)	0.000 (0.000–0.000)	48423 (21%)
N-PM	0.400	0.328	0.361	0.105 (0.000–0.342)	0.113 (0.000–0.400)	0.782 (0.448–1.000)	5050 (2%)
Average	0.464	0.467	0.461	–			
Accuracy	0.467						

The values outside parentheses are the averages over the grids. The values inside parentheses are the minimum and maximum.

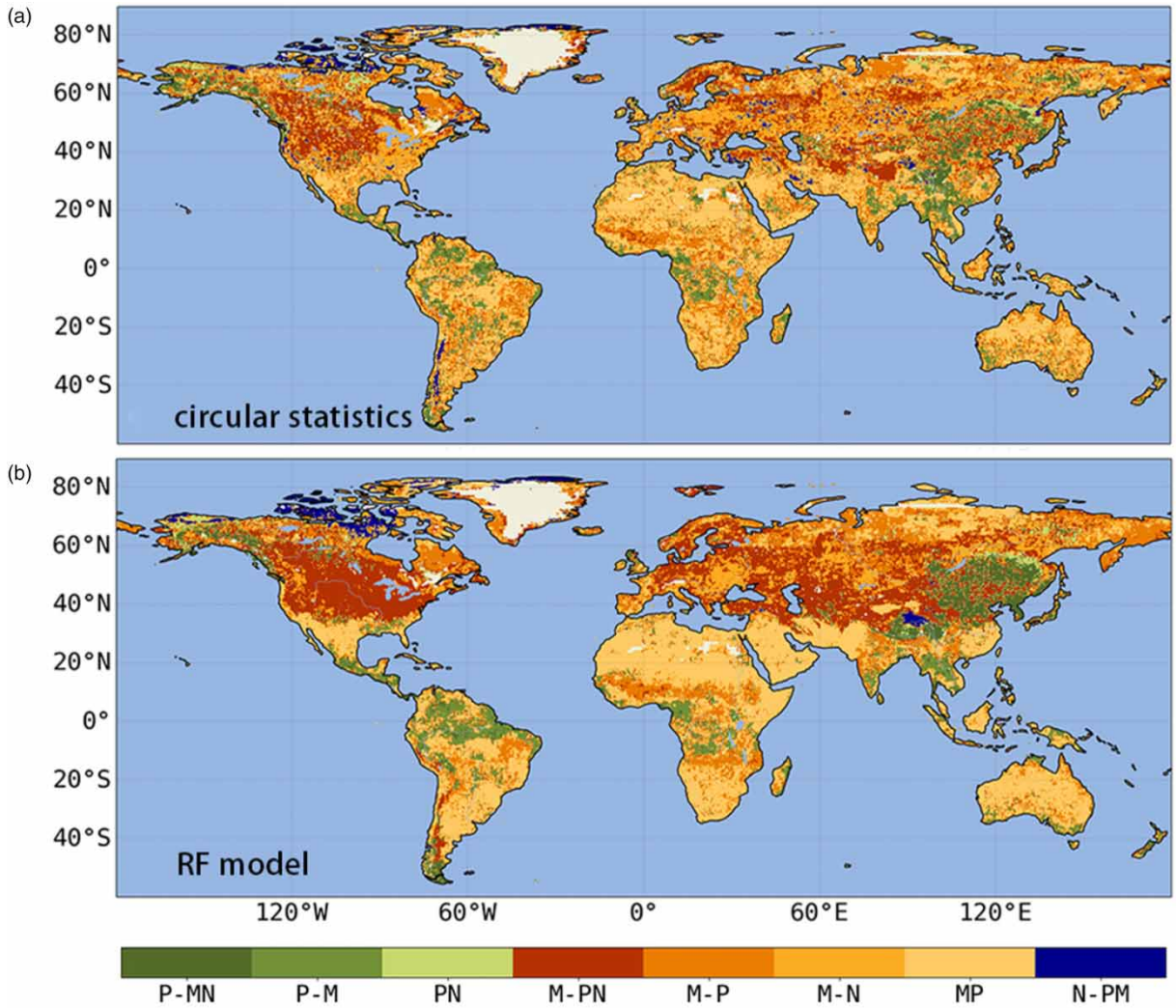

Figure 5 | The spatial distribution of eight sub-types of global floods from (a) circular statistics-based estimations and (b) the RF model in the 10-fold cross-validation.

soil moisture excess accounts for a negligible portion of flooding occurrences while both extreme precipitation and snowmelt play significant roles in flood generation.

The M type can be further divided into four sub-types. The M-PN type indicates that soil moisture excess overwhelmingly dominates the flood with a mean α_m value of 0.938. This flood type is widely distributed over the globe but quite scattered. The M-P type represents that both extreme precipitation and soil moisture excess substantially contribute to the flood generation with a mean relative importance value of 0.729 for the former and a mean value of 0.269 for the later. This type occurs in large swaths of high latitudes in the Northern Hemisphere. It is cold all year round in these areas, and snowmelt has little impact on discharge. Different from the M-P type, the M-N type indicates that snowmelt makes a substantial contribution to flood generation and the mean relative importance of this type is approximately 0.344. This type overwhelms other types in North America and Europe as well as some areas of central and western Asia. The MP-* type represents a similar contribution of soil moisture excess and extreme precipitation to the flood generation but little contribution of snowmelt, and is widely distributed in Northern Africa, central South America, and Australia.

As for the N type, it corresponds to only one type N-PM, showing snowmelt overwhelms flood drivers with a mean α_m of 0.782. The N-PM type represents areas under the primary control of snowmelt and the low influence of precipitation and soil moisture (with a mean value of 0.105 and 0.113, respectively). It should be noted that some overlap was observed between the newly identified clusters and originally three classes. For example, some data points in the newly allocated P-MN category may have originally belonged to the M or N classes. This overlap was not prevented by the lack of inherent structure in the k-means algorithm. Specifically, N type represents the areas with the highest influence from snowmelt no matter how similar its relative importance is to the other two factors. Also, the number of grids with the N-PM type is 5, 050, while the number of the N type is almost doubled.

We also attempt to use the RF model to predict these flood sub-types. Following the same approach as predicting the primary types of flooding, the RF models for predicting sub-types of flooding are also trained using longitude, latitude, and the occurrence day of the highest levels of precipitation, soil moisture, and snowmelt after applying the unordered transformation as the predictors. However, the models do not perform well with 0.467 accuracy and 0.461 F1-score (Table 1) in the 10-fold cross-validation due to insufficient training samples and input features. Similar to the spatial distribution of the three predicted types in Figure 5(b), the sub-types predictions also show a great spatial continuity compared with the scattered distribution in circular statistics-based estimations (Figure 5(a) and 5(b)).

4.4. Model feature importance

Feature importance is an essential tool in machine learning that helps us understand which variables carry the most weight in predicting the outcome of interest. It measures how much each feature contributes to the accuracy of the model prediction. By understanding feature importance, we can gain insights into the underlying factors that drive the prediction results. In this study, the RF model uses 62 features to predict flooding types as described in Section 3.3 (Figure 6). The results show that

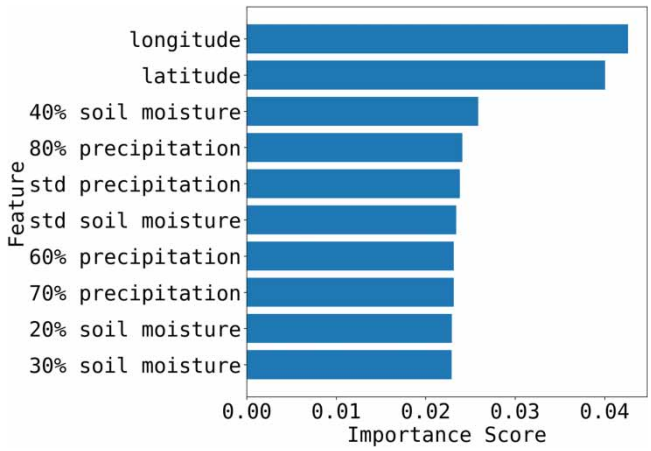

Figure 6 | Feature importance analysis of the RF model for top 10 features. The abbreviation 'std' stands for standard deviation, and the percentage represents the percentile of the variable.

geographic information including longitude and latitude are the most important variables, with an importance score exceeding 0.04. This suggests that the location of a particular area can have a significant impact on the likelihood of flooding. Interestingly, the model also finds that low percentile values (i.e., 20th, 30th, and 40th) for the occurrence day of the highest soil moisture and high percentile values (i.e., 60th, 70th, and 80th) for the occurrence day of the highest precipitation are relatively more important variables. Furthermore, the standard deviation for the highest precipitation and soil moisture are also found to be important features in the model. This suggests that variations in precipitation and soil moisture could significantly affect flooding occurrence.

5. DISCUSSION

The top-down hypothesis explaining the relative importance of floods provides a simple and repeatable method to decipher the first level of understanding of the different flood-generating mechanisms. The RF model proposed in this study offers a simple and direct way to detect the dominant mechanism. The good performance between the RF model predictions and statistics calculations suggests that the flood-generating mechanism can be predicted without discharge input. Compared with other studies that use circular statistic to learn the process controls on floods (e.g., Parajka *et al.* 2010; Slater & Villarini 2017), our trained RF model achieves robust results while not relying on the discharge data. This advantage is particularly important in the areas without discharge data. It seems that our finding that soil moisture is the primary factor influencing global flood occurrences contradicts some of the previous studies. For instance, Brunner & Fischer (2022) and Pechlivanidis *et al.* (2020) concluded that precipitation or snow melt are the main drivers. However, it should be noted that these studies did not consider soil moisture as a candidate factor for flood occurrence, and thus cannot be directly compared to our conclusions. Moreover, Stein *et al.* (2020) highlighted that the local dominant flood generation processes could be varied for different flooding occurrences. These individual contributions from different processes are difficult for the RF model to predict based on the limited number of input attributes. Therefore, we should keep in mind that the uncertainties exist in determining the dominant mechanism for flooding occurrences with this method, though the overall prediction accuracy is high.

To address the issue of imbalanced data in our study sample, we also attempted to use a resampling method, namely the Synthetic Minority Oversampling TEchnique (SMOTE) algorithm for prediction. The SMOTE algorithm is designed to address the issue of class imbalance in machine learning models. It works by oversampling the minority class (which usually has fewer samples) by generating synthetic data points that are similar to the existing samples. This helps to balance the class distribution and therefore may improve the model performance. However, the results showed that after applying the SMOTE algorithm, the average precision, recall, F1-score, and accuracy values were 0.682, 0.685, 0.683, and 0.685, respectively. These values were slightly lower compared to the corresponding values (0.692, 0.701, 0.692, and 0.701, respectively) in the experiments with the original data. Although the SMOTE algorithm slightly improved the predictions for the P and N classes, it resulted in a significant decrease in the prediction accuracy for the M class. For instance, the F1-score dropped from 0.773 to 0.753. In this particular study, the decrease in accuracy after applying the SMOTE algorithm could be due to several reasons. First, the synthetic data generated by SMOTE may not have captured the underlying patterns and relationships between the input features and the target variable accurately. This can result in overfitting or underfitting of the model, leading to a drop in the accuracy. Second, while the SMOTE algorithm can increase the number of samples in the minority class, it does not necessarily guarantee that these samples are the representatives of the true distribution of the class. As a result, the synthetic data points may introduce noise into the model, leading to a lower prediction accuracy. In such cases, oversampling the minority class can actually harm the overall performance of the model. Therefore, it is essential to carefully evaluate the impact of SMOTE on the dataset and model performance before deciding whether to use it or not.

While our study yielded promising results, it is important to note that the generalizability of our findings to other datasets may be limited. As with any machine learning model, the performance of our methodology may be impacted by the specific features and characteristics of the dataset used. Therefore, it is possible that adjustments to the methodology may be necessary to achieve comparable results on different data sources. To address this potential limitation, future research could focus on the generalizability of our methodology across a wide variety of datasets. This would involve replicating our methodology on different data sources and conducting comprehensive sensitivity analyses to evaluate the robustness of the results. Such efforts could provide valuable insights into the performance and generalizability of our methodology in different contexts. Moreover, due to the relatively coarse spatial resolution, topographical effects are not considered. The contributions from

watershed features such as shape and slope are ignored in the implementation. Furthermore, the flood may be driven by mechanisms other than the three drivers investigated in this study. For example, long-term rainfall, glacial outburst (Björnsson 2003) and water resources management activities (such as reservoir operation and irrigation) can also affect flood generation. In addition, flood mechanisms in the same locality may differ depending on the specific events that occur (Bennett *et al.* 2018; Pendergrass 2018), which is not considered in this study.

6. CONCLUSIONS

This study provides a framework of using machine learning models to identify flood-generating mechanisms. We use circular statistics to estimate the seasonality of extreme precipitation, soil moisture excess and snowmelt. Three dominant flood types (P type, M type, and N type) and eight sub-types are identified. RF models are trained to predict the flood types. This global-scale study indicates that soil moisture excess is the most important driver of annual maximum flooding followed by extreme precipitation and snowmelt. During 1979–2013, nearly 44% of the regions have a late trend in annual flood timing with a mean value of 7 days/decade, while 51% of the regions show an earlier trend with a mean value of 6 days/decade. The RF model can identify the dominant driver of flood generation with an overall F1-score value of 0.692 and accuracy of 0.701. The approach presented in this study provides a means to extend the flood type mapping to other regions at a finer spatial scale where discharge data may not be available, thereby enhancing our understanding of the underlying processes driving flood generation.

FUNDING

This research is funded by the National Natural Science Foundation of China (Nos 52079151 and 52111540261).

DATA AVAILABILITY STATEMENT

All relevant data are included in the paper or its Supplementary Information.

CONFLICT OF INTEREST

The authors declare there is no conflict.

REFERENCES

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Le Vine, N. & Clark, M. P. 2018 A ranking of hydrological signatures based on their predictability in space. *Water Resources Research* **54** (11), 8792–8812. https://doi.org/10.1029/2018WR022606.

Alfieri, L., Burek, P., Feyen, L. & Forzieri, G. 2015 Global warming increases the frequency of river floods in Europe. *Hydrology and Earth System Sciences Discussions* **12** (1), 1119–1152. https://doi.org/10.5194/hessd-12-1119-2015.

Arnell, N. W. & Gosling, S. N. 2016 The impacts of climate change on river flood risk at the global scale. *Climatic Change* **134** (3), 387–401. https://doi.org/10.1007/s10584-014-1084-5.

Bennett, B., Leonard, M., Deng, Y. & Westra, S. 2018 An empirical investigation into the effect of antecedent precipitation on flood volume. *Journal of Hydrology* **567**, 435–445. https://doi.org/10.1016/j.jhydrol.2018.10.025.

Berghuijs, W. R., Woods, R. A., Hutton, C. J. & Sivapalan, M. 2016 Dominant flood generating mechanisms across the United States. *Geophysical Research Letters* **43** (9), 4382–4390. https://doi.org/10.1002/2016GL068070.

Berghuijs, W. R., Larsen, J. R., van Emmerik, T. H. M. & Woods, R. A. 2017 A global assessment of runoff sensitivity to changes in precipitation, potential evaporation, and other factors. *Water Resources Research* **53** (10), 8475–8486. https://doi.org/10.1002/2017WR021593.

Berghuijs, W. R., Harrigan, S., Molnar, P., Slater, L. J. & Kirchner, J. W. 2019 The relative importance of different flood-Generating mechanisms across Europe. *Water Resources Research* **55** (11), 8567–8584. https://doi.org/10.1029/2019WR024841.

Björnsson, H. 2003 Subglacial lakes and jökulhlaups in Iceland. *Global and Planetary Change* **35** (3–4), 255–271. https://doi.org/https://doi.org/10.1016/S0921-8181(02)00130-3.

Bloschl, G., Hall, J., Parajka, J., Perdigao, R., Merz, B., Arheimer, B., Aronica, G. T., Bilibashi, A., Bonacci, O., Borga, M., Canjevac, I., Castellarin, A., Chirico, G. B., Claps, P., Fiala, K., Frolova, N., Gorbachova, L., Gul, A., Hannaford, J., Harrigan, S., Kireeva, M., Kiss, A., Kjeldsen, T. R., Kohnova, S., Koskela, J. J., Ledvinka, O., Macdonald, N., Mavrova-Guirguinova, M., Mediero, L., Merz, R., Molnar, P., Montanari, A., Murphy, C., Osuch, M., Ovcharuk, V., Radevski, I., Rogger, M., Salinas, J. L., Sauquet, E., Sraj, M., Szolgay, J., Viglione, A., Volpi, E., Wilson, D., Zaimi, K. & Zivkovic, N. 2017 Changing climate shifts timing of European floods. *Science* **357** (6351), 588–590. https://doi.org/10.1126/science.aan2506.

Breiman, L. 2001 Random forests. *Machine Learning* **5** (45), 32. https://doi.org/10.1023/A:1010933404324.

Brunner, M. & Fischer, S. 2022 Snow-influenced floods are more strongly connected in space than purely rainfall-driven floods. *Environmental Research Letters* **17** (7), 075001. doi:10.1088/1748-9326/ac948f.

Budyko, M. I. 1974 *Climate and Life*. Academic Press, New York, NY.

Burn, D. 1997 Catchment similarity for regional flood frequency analysis using seasonality measures. *Journal of Hydrology* **202** (2), 12–230. https://doi.org/10.1016/S0022-1694(97)00068-1.

Curran, J. H. & Biles, F. E. 2021 Identification of seasonal streamflow regimes and streamflow drivers for daily and peak flows in Alaska. *Water Resources Research* **57** (2). https://doi.org/10.1029/2020WR028425.

Dottori, F., Szewczyk, W., Ciscar, J., Zhao, F., Alfieri, L., Hirabayashi, Y., Bianchi, A., Mongelli, I., Frieler, K., Betts, R. A. & Feyen, L. 2018 Increased human and economic losses from river flooding with anthropogenic warming. *Nature Climate Change* **8** (9), 781–786. https://doi.org/10.1038/s41558-018-0257-z.

Esposito, G., Matano, F. & Scepi, G. 2018 Analysis of increasing flash flood frequency in the densely urbanized coastline of the campi flegrei Volcanic Area, Italy. *Frontiers in Earth Science* **6**. https://doi.org/10.3389/feart.2018.00063.

Gudmundsson, L., Greve, P. & Seneviratne, S. I. 2016 The sensitivity of water availability to changes in the aridity index and other factors – A probabilistic analysis in the Budyko space. *Geophysical Research Letters* **43**, 6985–6994. https://doi.org/10.1002/2016GL069763.

Gudmundsson, L., Greve, P. & Seneviratne, S. I. 2017 Correspondence: flawed assumptions compromise water yield assessment. *Nature Communications* **8**, 14795. https://doi.org/10.1038/ncomms14795.

Hall, J., Arheimer, B., Aronica, G. T., Bilibashi, A., Boháč, M., Bonacci, O., Borga, M., Burlando, P., Castellarin, A., Chirico, G. B., Claps, P., Fiala, K., Gaál, L., Gorbachova, L., Gül, A., Hannaford, J., Kiss, A., Kjeldsen, T., Kohnová, S., Koskela, J. J., Macdonald, N., Mavrova-Guirguinova, M., Ledvinka, O., Mediero, L., Merz, B., Merz, R., Molnar, P., Montanari, A., Osuch, M., Parajka, J., Perdigão, R. A. P., Radevski, I., Renard, B., Rogger, M., Salinas, J. L., Sauquet, E., Šraj, M., Szolgay, J., Viglione, A., Volpi, E., Wilson, D., Zaimi, K. & Blöschl, G. 2015 A European flood database: facilitating comprehensive flood research beyond administrative boundaries. *Proceedings of the International Association of Hydrological Sciences* **370**, 89–95. https://doi.org/10.5194/piahs-370-89-2015.

Hounkpè, J., Diekkrüger, B., Afouda, A. A. & Sintondji, L. O. C. 2019 Land use change increases flood hazard: a multi-modelling approach to assess change in flood characteristics driven by socio-economic land use change scenarios. *Natural Hazards* **98** (3), 1021–1050. https://doi.org/10.1007/s11069-018-3557-8.

Jonkman, S. N. 2005 Global perspectives on loss of human life caused by floods. *Natural Hazards (Dordrecht)* **34** (2), 151–175. https://doi.org/10.1007/s11069-004-8891-3.

Keller, L., Rössler, O., Martius, O. & Weingartner, R. 2018 Delineation of flood generating processes and their hydrological response. *Hydrological Processes* **32** (2), 228–240. https://doi.org/10.1002/hyp.11407.

Kendon, E. J., Roberts, N. M., Fowler, H. J., Roberts, M. J., Chan, S. C. & Senior, C. A. 2014 Heavier summer downpours with climate change revealed by weather forecast resolution model. *Nature Climate Change* **4** (7), 570–576. https://doi.org/10.1038/nclimate2258.

Kim, J. H., Kim, Y. H. & Lee, K. M. 2018 Consumer preferences and willingness to pay for functional rice bread: a choice experiment approach. *Food Quality and Preference* **63**, 73–80. https://doi.org/10.1016/j.foodqual.2017.08.013.

Koç, G. & Thieken, A. H. 2018 The relevance of flood hazards and impacts in Turkey: what can be learned from different disaster loss databases? *Natural Hazards* **91** (1), 375–408. https://doi.org/10.1007/s11069-017-3134-6.

Kundzewicz, Z. W., Kanae, S., Seneviratne, S. I., Handmer, J., Nicholls, N., Peduzzi, P., Mechler, R., Bouwer, L. M., Arnell, N., Mach, K., Li, D., Lettenmaier, D. P., Margulis, S. A. & Andreadis, K. 2019 The role of rain-on-snow in flooding over the conterminous United States. *Water Resources Research* **55** (11), 8492–8513. https://doi.org/10.1029/2019WR024950.

Lin, P., Pan, M., Beck, H. E., Yang, Y., Yamazaki, D., Frasson, R., David, C. H., Durand, M., Pavelsky, T. M., Allen, G. H., Gleason, C. J. & Wood, E. F. 2019 Global reconstruction of naturalized river flows at 2.94 million reaches. *Water Resources Research* **55** (8), 6499–6516. https://doi.org/10.1029/2019WR025287.

Mallakpour, I. & Villarini, G. 2015 The changing nature of flooding across the central United States. *Nature Climate Change* **5** (3), 250–254. https://doi.org/10.1038/nclimate2516.

Milly, P. C. D., Kam, J. & Dunne, K. A. 2018 On the sensitivity of annual streamflow to Air temperature. *Water Resources Research* **54** (4), 2624–2641. https://doi.org/10.1002/2017WR021970.

Najibi, N. & Devineni, N. 2018 Recent trends in the frequency and duration of global floods. *Earth System Dynamics* **9** (2), 757–783. https://doi.org/10.5194/esd-9-757-2018.

Neri, A., Villarini, G., Salvi, K. A., Slater, L. J. & Napolitano, F. 2019 On the decadal predictability of the frequency of flood events across the U.S. Midwest. *International Journal of Climatology* **39** (3), 1796–1804. https://doi.org/10.1002/joc.5915.

Nguyen, T. H., Kant, S. & Ma, X. 2017 Land use/land cover change analysis and its impact on soil erosion potential in the upper catchment of the Mekong River basin. *Geocarto International* **32** (6), 623–636. https://doi.org/10.1080/10106049.2016.1182645.

Parajka, J., Kohnová, S., Bálint, G., Barbuc, M., Borga, M., Claps, P., Cheval, S., Dumitrescu, A., Gaume, E., Hlavčová, K., Merz, R., Pfaundler, M., Stancalie, G., Szolgay, J. & Blöschl, G. 2010 Seasonal characteristics of flood regimes across the Alpine–Carpathian range. *Journal of Hydrology* **394** (1–2), 78–89. https://doi.org/10.1016/j.jhydrol.2010.05.015.

Patterson, L. A., Lutz, B. & Doyle, M. W. 2013 Climate and direct human contributions to changes in mean annual streamflow in the South Atlantic, USA. *Water Resources Research* **49** (11), 7278–7291. https://doi.org/10.1002/2013WR014618.

Pechlivanidis, I. G., Crochemore, L., Rosberg, J. & Bosshard, T. 2020 What are the key drivers controlling the quality of seasonal streamflow forecasts? *Water Resources Research* **56** (5), e2019WR026987. doi:10.1029/2019WR026987.

Pendergrass, A. G. 2018 What precipitation is extreme? *Science (American Association for the Advancement of Science)* **360** (6393), 1072–1073. https://doi.org/10.1126/science.aat1871.

Slater, L. J. & Villarini, G. 2016 Recent trends in U.S. flood risk. *Geophysical Research Letters* **43** (12), 428–436. https://doi.org/10.1002/2016GL068070t.

Slater, L. & Villarini, G. 2017 Evaluating the drivers of seasonal streamflow in the U.S. midwest. *Water* **9** (9), 695. https://doi.org/10.3390/w9090695.

Stein, L., Pianosi, F. & Woods, R. 2020 Event-based classification for global study of river flood generating processes. *Hydrological Processes* **34** (7), 1514–1529. https://doi.org/10.1002/hyp.13678.

Stein, T. V., Tallaksen, L. M., Lyon, S. W., Dumont, J., Dang, T. & Tönshoff, C. 2021 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** (4), e2020WR029359. https://doi.org/10.1029/2020WR029359.

Turini, N., Thies, B. & Bendix, J. 2019 Estimating high spatio-temporal resolution rainfall from MSG1 and GPM IMERG based on machine learning: case study of Iran. *Remote Sensing* **11** (19), 2307. https://doi.org/10.3390/rs11192307.

Wang, H., Sun, F., Xia, J. & Liu, W. 2017 Impact of LUCC on streamflow based on the SWAT model over the Wei River basin on the Loess Plateau in China. *Hydrology and Earth System Sciences* **21** (4), 1929–1945. https://doi.org/10.5194/hess-21-1929-2017.

Wang, H., Wen, X. & Huang, G. 2021 Identifying the driving factors of energy-related CO2 emissions in China using a panel data approach. *Energy Policy* **154**, 112339. https://doi.org/10.1016/j.enpol.2021.112339.

Wiener, A. L. M. 2002 Classification and regression by randomForest. *R News* **2** (3), 18–22.

Winsemius, H. C., Aerts, J. C. J. H., van Beek, L. P. H., Bierkens, M. F. P., Bouwman, A., Jongman, B., Kwadijk, J. C. J., Ligtvoet, W., Lucas, P. L., van Vuuren, D. P. & Ward, P. J. 2016 Global drivers of future river flood risk. *Nature Climate Change* **6** (4), 381–385. https://doi.org/10.1038/nclimate2893.

Wu, H. & Qian, H. 2017 Innovative trend analysis of annual and seasonal rainfall and extreme values in Shaanxi, China, since the 1950s. *International Journal of Climatology* **37** (5), 2582–2592. https://doi.org/10.1002/joc.4866.

Xu, L., Chen, N., Zhang, X., Chen, Z., Hu, C. & Wang, C. 2019 Improving the North American multi-model ensemble (NMME) precipitation forecasts at local areas using wavelet and machine learning. *Climate Dynamics* **53** (1–2), 601–615. https://doi.org/10.1007/s00382-018-04605-z.

Yin, J., Gentine, P., Zhou, S., Sullivan, S. C., Wang, R., Zhang, Y. & Guo, S. 2018 Large increase in global storm runoff extremes driven by climate and anthropogenic changes. *Nature Communications* **9** (1). https://doi.org/10.1038/s41467-018-06765-2.

Zhang, M. & Shi, W. 2019 Systematic comparison of five machine-learning methods in classification and interpolation of soil particle size fractions using different transformed data. *Hydrology and Earth System Sciences discussions* 1–39. https://doi.org/10.5194/hess-2018-584.

Zhou, X., Li, L., Wu, H. & Liu, Z. 2019 Contributions of individual land use types to soil carbon sequestration in the Loess Plateau of China. *Catena* **177**, 7–15. https://doi.org/10.1016/j.catena.2019.02.001.

First received 8 April 2023; accepted in revised form 9 September 2023. Available online 20 September 2023