Hydrological Sciences Journal

ISSN: 0262-6667 (Print) 2150-3435 (Online) Journal homepage: www.tandfonline.com/journals/thsj20

Large-sample characterization of flooding events in India

Sai Kiran Kuntla & Manabendra Saharia

To cite this article: Sai Kiran Kuntla & Manabendra Saharia (30 May 2025): Large-sample characterization of flooding events in India, Hydrological Sciences Journal, DOI: 10.1080/02626667.2025.2513482

To link to this article: https://doi.org/10.1080/02626667.2025.2513482

Publisher: Taylor & Francis & IAHS

Journal: *Hydrological Sciences Journal*

DOI: 10.1080/02626667.2025.2513482

Large-sample characterization of flooding events in India

Sai Kiran Kuntla[1] and Manabendra Saharia*[1]

[1]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

***Corresponding Author**:

Indian Institute of Technology Delhi

New Delhi, India 110016

Office: +91-011-26591260

Email: msaharia@iitd.ac.in

Abstract

Effective flood management requires a robust understanding of past floods. In India, such understanding is largely limited to case studies due to the absence of a standardized observed flood dataset. We address this gap by presenting a national dataset of 7500 flooding events, developed by merging observed streamflow records with official flooding thresholds and augmenting it with multiple catchment-scale variables. Spatial analysis reveals high normalized flood magnitudes along the southwest coast—an area with intense rainfall and mountainous terrain. Temporally, 86% of floods occur during the southwest monsoon. Using a random forest model combined with the game-theoretic SHAP approach, we find that precipitation of the wettest month is the most influential predictor of flood magnitude. Grouped feature importance shows climatology contributes 61% to model performance, while geomorphology accounts for 39%. This comprehensive large-sample study surpasses conventional case studies, providing a more robust understanding of flooding patterns and drivers across India.

Keywords: floods; envelope curves; flood magnitude; random forest; data-driven hydrology

1. Introduction

Widespread floods threaten communities globally, triggering devastating consequences on human and environmental fronts. There is evidence of increasing flood events globally and their consequential economic damage (Kuntla, 2021; Newman & Noy, 2023). In India alone, floods have resulted in 113,390 human casualties between 1975 and 2015 (Saharia et al., 2021). However, floods are an intricate phenomenon resulting from the complex interaction of multiple physiographic and atmospheric features, and there is a wide variability in these relationships from catchment to catchment. Though India has some of the world's most densely populated flood-prone catchments and characterizes diverse geographical settings, including various climates, landforms, and topography, a detailed assessment of flood characteristics at a national scale to understand flood climatology and the influencing hydrogeomorphic characteristics has been missing.

A data-intensive approach has emerged as the fourth paradigm of hydrology that would help find new relationships that are often hindered by conventional catchment or region-specific studies, and derive more robust conclusions on hydrological processes and relationships using a large sample of catchments (Addor et al., 2020; Kuntla et al., 2023; Peters-Lidard et al., 2017). For characterizing flood hydroclimatology in this context, firstly, we require a representative and comprehensive flood database with wide spatiotemporal representation. Globally, extensive efforts have been made to develop flood events databases for computational hydrology in the United States (Gourley et al., 2013; Huang et al., 2021) and Europe (Gaume et al., 2009). Similarly, several studies have attempted to characterize observed floods at larger scales in developed regions such as Europe (Marchi et al., 2010) and the United States (Costa, 1987; Saharia et al., 2017). In comparison, large-scale and large-sample studies are rare in developing countries such as India. The major impediment for such studies in India is the nonexistence of a comprehensive catalog of past flood events that occurred in the country, along with their physical characteristics such as flood level, discharge, duration, etc.

On the other hand, large-scale modeled streamflow has been widely used to address data gaps in hydrology but face significant limitations that hinder their reliability for flood characterization (Magotra et al., 2024; Salas et al., 2018; Sood & Smakhtin, 2015; Ward et al., 2015; Yamazaki et al., 2011). Hydrological models often struggle with capturing peak flows and extreme events as they are inherently subjected to many limitations due to uncertainties in

input data, coarse spatial resolution, and oversimplified parameterizations (Kupzig et al., 2024; Sood & Smakhtin, 2015; Ward et al., 2015). These biases are particularly pronounced in regions with complex hydrogeomorphological conditions or limited observational data for calibration. As a result, modeled runoff datasets may fail to accurately represent localized flood dynamics or provide reliable estimates of key flood parameters such as discharge and timing. This underscores the critical need for observed flood event data to ensure accuracy and robustness in hydrological analyses. Additionally, global flood databases like the Dartmouth Flood Observatory (DFO), and the Emergency Disasters Database (EM-DAT) primarily rely on surveys and reports, focusing on the impacts of only major flood events. However, they often lack consistent records of critical flood parameters such as water levels, time to peak, discharge, and duration, limiting their applicability in computational hydrology. This gap prevents them from capturing the full range of flood events, including variations in location, triggering mechanisms, severity, and catchment hydrogeomorphology. Similarly, available flood event datasets in India, such as the India Flood Inventory (IFI), do not include any information on flood discharges and associated catchment variables, which limits their usefulness in computational research and subject to many limitations as global databases (Saharia et al., 2021).

Nevertheless, many observational studies have investigated local catchment behavior in India through individual catchment or region-specific analyses. For instance, Dhar & Nandargi (2000) investigated flooding behavior in the Brahmaputra basin, while Kale et al. (1997) examined a few river basins in the Peninsula region. Bhatt & Ahmed (2014) have performed a thorough morphometric analysis to determine floods in the Upper Krishna basin. At the same time, Rana & Suryanarayana (2021) performed a sub-watershed analysis of the Vishwamitri river catchment to assess the flood-influencing characteristics and prioritize them for flood mitigation measures. Das & Scaringi (2021) investigated factors influencing the occurrence and magnitude of floods in the Mahi basin. Recently, a few simulation-based efforts have studied flooding at a regional or national scale in India, like finding the prominent drivers of flooding in the peninsula region of the country based on modeled streamflow (Nanditha & Mishra, 2022). The development of national-scale models such as the Indian Land Data Assimilation System (ILDAS), which integrates hydrologic-hydrodynamic modeling to estimate land surface states, channel discharge, and floodplain inundation have also started exploring large-scale characteristics of flooding in South Asia. (Magotra et al., 2024). However, catchment-specific and simulation-based studies cannot

replace large-scale and large-sample observational studies in order to derive general hydrological principles regarding flooding variables and their influencing catchment characteristics that could be robust and widely accepted (Addor et al., 2020; Kuntla et al., 2022). Large-sample observational studies are indispensable as they provide an actual representation of hydrological systems in different conditions and over various timescales, while simulated data are based on models that may not account for natural variability and complexity and have inherent uncertainties or limitations.

This study is the first such effort to develop a large-sample event-based flood database to carry out a comprehensive national-scale characterization study. A unique flood event database containing 7500 observed flood events is developed by combining the most spatially and temporally representative observed streamflow records and official flooding thresholds in 133 gauge stations. Furthermore, these observed flood events are augmented with an array of catchment characteristics including geomorphology, climatology, lithology, and soil, to investigate their relationship with flood magnitude using boxplots, spearman correlations, envelope curves, and interpretable machine learning methods. The manuscript is structured as follows: Section 2 describes the data sources and preprocessing methods, while Section 3 outlines the analytical approaches used to evaluate the first-order and multidimensional influence of catchment characteristics on flood magnitudes. Section 4 presents the spatial and temporal distribution of flood events across India. Sections 5 and 6 explore the roles of geomorphological and climatological characteristics, respectively. Section 7 compares the relative importance of these features using interpretable machine learning techniques. Section 8 establishes an empirical relationship between normalized flood magnitude and drainage area using envelope curves, and Section 9 investigates the influence of soil and lithology. Finally, Section 10 summarizes the key insights and provides concluding remarks.

2. Data

2.1. Observed flood event database

A first-of-its-kind database of flooding events in India has been developed by extensive processing of streamflow time series records from 1959 and official flooding thresholds defined by the Central Water Commission (CWC). The official flood thresholds for each gauge station were obtained from the CWC flood forecasting portal: https://ffs.india-water.gov.in/. The catchments in this dataset cover diverse climatological and physiological conditions in the country, which makes it a representative flood database of India. Floods are

classified from daily stream water level time series data based on the corresponding Warning level threshold defined for every gauge station by the CWC, as illustrated in Figure 1. The start and end of a flooding event is defined by when the water level crosses the CWC-defined Warning level at the respective gauge station. These thresholds are fixed by the CWC in consultation with the state governments and local authorities, based on a combination of historical records, field surveys, and ground experience (NDMA, 2008). The Warning level is typically set above the normal river water level and serves as an early alert to initiate preparedness measures. The danger level indicates a critical river stage at which floodwaters are likely to cause significant damage, triggering evacuation and relief operations (GFCC, 2004). For each flood event, the following information is reported in the final database: start and end date of the flood event, peak flood level and its date of occurrence, peak discharge and its date of occurrence, flood volume, event duration, time to peak, and recession time (time taken from peak time to end time).

Further, taking the geographical location of the gauging station as the outlet, catchments are delineated for all 133 stations. These shapefiles are subsequently used to compute the catchment-scale explanatory variables from other datasets.

2.2. Target variable: normalized flood magnitude

The peak discharge values of the observed 7500 floods are taken as flood magnitude for the corresponding flood events. Since larger catchments are expected to collect and carry more significant discharges, in this study, the flood magnitudes are normalized by catchment area to allow intercomparison of flood characteristics of various catchment sizes. This normalized discharge value, often known as unit peak discharge or specific discharge, has been widely used for studying high streamflow extremes and floods (Abdullah et al., 2019; Castellarin, 2007; Javier et al., 2007). For example, this index was used in describing floods and establishing a relationship between catchment areas and their peak discharges using envelope curves in Europe, the US, and Malaysia (Abdullah et al., 2019; Gaume et al., 2009; Marchi et al., 2010; Saharia et al., 2017). Lun et al. (2021) examined the process controls on spatial patterns of flood moments throughout European catchments, including the mean of unit peak discharge. Apart from these studies, Kuntla et al. (2022) have recently studied the characteristics of streamflow extremes utilizing a large sample of streamflow data and catchment characteristics belonging to more than 9000 catchments across the globe. At the

same time, the unit peak discharge indicates flood severity in the catchment, as it quantifies the flood volume in a catchment per unit area.

2.3. Explanatory variables: geomorphology

The geomorphology of a catchment has a significant impact on the characteristics of the streamflow (Ahn & Merwade, 2016; Saharia et al., 2017; Zhang et al., 2022). To explore how the catchment geomorphology influences normalized flood magnitude, a large set of catchments-scale geomorphology explanatory variables at a 1-km spatial scale were extracted from the Global Distributed Basin Characteristics (GDBC) database (Shen et al., 2017). The 1-km resolution of the GDBC is well-suited for our large-sample, nationwide analysis, where the average catchment size is 22,893 km². At this scale, any uncertainties at the grid level are minimized when aggregated at the catchment scale. Additionally, the dataset has been successfully applied in previous studies on hydrological extremes (Kuntla et al., 2022; Peluso et al., 2023), demonstrating its capability to capture relevant large-scale geomorphological patterns. Therefore, the GDBC dataset provides meaningful insights without significantly impacting on the robustness of our findings. The variables include catchment characteristics such as Strahler stream order, drainage area, catchment magnitude, all orders stream length, maximal flow length, down valley length, catchment relief, catchment length, catchment perimeter, and all orders streams mean length. To further enhance the analysis, we derived a secondary set of geomorphological variables using the primary products available in GDBC, such as the sinuosity index, form factor, relief ratio, elongation ratio, circularity ratio, lemniscate value, drainage texture, drainage density, compactness coefficient, wandering ratio, fitness ratio, channel frequency, drainage intensity, infiltration number, and ruggedness number. Descriptions of all these variables are available in the Supplementary material, Table S1.

2.4. Explanatory variables: climatology

Catchment-averaged climatological variables were derived from the 1-km resolution WorldClim datasets (Fick & Hijmans, 2017). Some of these variables include annual mean temperature, maximum temperature of warmest month, annual precipitation, precipitation of wettest month, and precipitation of driest month. These climatological variables are derived from the monthly rainfall and temperature averages for the years 1970–2000. The complete

list and the descriptions of all these variables can be found in Supplementary Material, Table S1.

2.5. Soil and lithology types

Soil properties, including texture, permeability, and lithology, influence a catchment's runoff by affecting infiltration capacity, permeability, runoff coefficients, topography, soil moisture content, and geological structures. To study soil and lithology concerning normalized flood magnitude, we derived the dominant lithology type from Hartmann & Moosdorf (2012) and soil type from Hengl et al. (2017) in each catchment. If one single type is present over more than 50 % of the catchment area, the catchment is classified as being dominated by that lithology (soil), respectively. Otherwise, 'No dominant class'.

3. Methods

Figure 2 presents the workflow of this study.

3.1. First-order relationships

Boxplots are employed to investigate the ranges of normalized flood magnitudes over the major basins of India, simultaneously exploring the ranges across various lithology and soil types. Spearman correlations are employed to examine first-order relationships between normalized flood magnitudes and the geomorphology and climatology of corresponding catchments. In addition, boxplots are used to illustrate the variability of these variables across different basins.

Envelope curves represent the upper limit of the given data and establish a relationship between the observed floods and their contributing catchment areas. This is the first study to develop envelope curves using a large sample approach in India. Envelope curves have long been used in flood hydrology studies to provide effective graphical summaries of extreme flood events and derive empirical equations for peak discharge estimations at ungauged stations for a specific region (Castellarin, 2007; Mimikou, 1984), country (Abdullah et al., 2019; Costa, 1987; Saharia et al., 2017), continent (Gaume et al., 2009; Marchi et al., 2010), or at the global scale (Herschy, 2002; Kuntla et al., 2022). Envelope curves are based on a power-law Equation (Eq. 1) and plotted on a log-log diagram.

$$Q = \alpha A^{\beta} \qquad \text{(Eq. 1)}$$

Here, Q $(m^3s^{-1}km^{-2})$ is the normalized flood magnitude, A (km^2) is the catchment area, α $(m^3s^{-1}km^{-2(1+\beta)})$ is the reduced discharge, and β as the scaling coefficient. The reduced discharge can serve as an indication of flood magnitude by reducing the dependence of the catchment area on analysis. Calculating the β value involves fitting a regression line between log(Q) and log(A) (Castellarin, 2007). This β value indicates the rate of change in normalized flood magnitude concerning alterations in the catchment area. A β value closer to zero signifies a smaller change in unit peak discharge with variations in the catchment area.

3.2. Random forest

Since the relationship between different catchment characteristics (explanatory variables) and normalized flood magnitude (target variable) is nonlinear and complex, this research investigates the connections between explanatory variables and target variable by building a random forest machine learning model (Breiman, 2001). Random forests have demonstrated its efficacy in hydrological studies, excelling at identifying nonlinear relationships and estimating uncertainties (Addor et al., 2018; Kuntla et al., 2022; Stein et al., 2021). In this study, a random forest model is trained on 75% of all the events and evaluated on the remaining 25%.

3.3. Model interpretability using SHAP

Alongside the random forest model, a game-theoretic interpretation method called Shapley Additive exPlanations (SHAP; S. Lundberg & Lee, 2017) has been used to interpret the intricate relationships between the target variable and explanatory variables. SHAP facilitates a detailed model analysis by identifying the most influential explanatory variables and explaining their impact on the model, with a reduced risk of multicollinearity (S. M. Lundberg et al., 2020). SHAP summary plots were generated to visualize and elucidate the importance, impact, and correlation of all individual explanatory variables with the normalized flood magnitude. Each data point in the plot represents a SHAP value associated with an explanatory variable and an instance. The mean absolute SHAP value across various instances for a given explanatory variable provides an overall assessment of the explanatory variable's importance concerning the normalized flood magnitude. The order of explanatory variables along the y-axis is determined by their importance, with higher (lower) absolute mean SHAP values on the right side indicating higher (lower) relative importance. Besides, the horizontal position of the points illustrates the magnitude of their influence on predictions.

Zero values on the x-axis denote no impact, and positive (negative) values to the right (left) indicate positive (negative) impact. On the other hand, the color of the points indicates whether the corresponding explanatory variable values for a given observation in the dataset are high (red) or low (blue). The combined observations of point distribution along the x-axis and their color convey the correlation of explanatory variables with the target variable.

Furthermore, grouped feature importance was calculated based on the SHAP values for a holistic understanding of the collective impact of geomorphology and climatology on the model (Au et al., 2022). This approach helps in understanding the combined influence of these groups on the normalized flood magnitude rather than individual explanatory variables and helps gain deeper insights into the underlying mechanisms.

3.3.1. Grouped feature importance analysis across multiple scales

By grouping geomorphological (G_{geo}) and climatological (G_{cli}) variables, we aggregate the SHAP values of the individual variables within each group and normalize by the total absolute contribution. This provides a measure of the combined influence of each group and reflects its relative importance in the predictions of the model. This grouping-based approach is applied at three scales – station ($s = i$), basin ($s = b$), and global ($s = global$) – to yield station-wise, basin-wise, and global perspective on the relative importance of geomorphological versus climatological factors.

For a given scale s and group G, the absolute group contribution is computed using Eq. 2:

$$S_{s,G} = \sum_{j \in G} |\psi_{sj}| \quad \text{(Eq. 2)}$$

Where ψ_{sj} represents the SHAP value of feature j at scale s.

The total absolute contribution at scale s is given by Eq. 3:

$$T_s = \sum_{j \in all\ features} |\psi_{sj}| = S_{s,\,G\text{geo}} + S_{s,\,G\text{cli}} \quad \text{(Eq. 3)}$$

The relative percentage influence of group G at scale s is then:

$$I_{s,G} = (S_{s,\,G} / T_s) \times 100 \quad \text{(Eq. 4)}$$

This framework enables the identification of the dominant driver (geomorphological or climatological) at each scale:

- Station-wise ($s = i$)*:* reveals spatial variability in controlling mechanisms at each gauge station.
- Basin-wise ($s = b$): aggregates station-level contributions ($S_{i,G}$) within each basin to highlight regional patterns.
- Global (s = *global*): Sums station-level contributions ($S_{i,G}$) across all N stations to provide an overall assessment for the entire study region.

The dominant group at each scale is identified as the one with the highest percentage influence ($I_{s,G}$), offering insights into spatial and hierarchical variability in flood magnitude drivers.

4. Spatial and temporal distribution of floods

Figure 3 presents the spatial distribution of the mean flood magnitude of all the flood events over the corresponding gauge stations normalized by their catchment area across the major basins of the country. This normalization allows a comparison between catchments of various spatial scales. The number of gauge stations in each basin is tabulated in Table 1. The normalized flood magnitude is high along the Western Ghats of the country in West-flowing rivers from Tapi to Kanyakumari Basin and a few subbasins in the extreme west of Krishna Basin. Generally, the West-flowing rivers from Tapi to Kanyakumari come under the direct influence of the southwest monsoon and receive heavy and assured rainfall between June and September. Most of the area receives an annual rainfall of more than 2700 mm, and almost 70 to 80% is within the southwest monsoon (Central Water Commission, 2018). At the same time, west Krishna receives maximum rainfall compared to other regions in the same basin, with huge proportions in the monsoon period. All these high normalized flood magnitude regions along the southwest coast of the country are characterized mainly by high annual precipitation, mountainous terrain, and relatively small sub-basins. This was also observed in mountainous terrain along the oceans over Europe and the West Coast of the United States (Marchi et al., 2010; Saharia et al., 2017).

Many lower normalized mean flood magnitude values are observed in the Cauvery Basin, including Arkavati, Shimsha Subbasins, and the main channel in the west. It is observed that all these catchments in Cauvery with low normalized mean flood magnitude predominantly have metamorphic lithology in their catchment. The greater resistance of metamorphic rocks to weathering contributes to the slower weathering and erosion rates in Cauvery, likely

contributing to the low normalized flood magnitude values in these catchments (Rajamani et al., 2009; Sharma & Rajamani, 2000). The Cauvery basin has a diverse spatial distribution of annual precipitation with an average rainfall of 500 mm to 1000 mm. Figure 4 confirms the low normalized flood magnitude values in the Cauvery basin, with a median of 0.02 among other basins in the country. It is followed by East-flowing rivers from Pennar to Kanyakumari basin and Ganga basin. Cauvery and East-flowing rivers from Pennar to Kanyakumari basin are less likely to experience short-duration, intense precipitation events, and the absence of a significant increase in extreme rainfall events in this regions (Chaubey et al., 2022) might have contributed to the floods not being more intense at these regions. Besides, the loss of the flood plain and the compromised meandering of the river in the Cauvery basin have reduced its ability to hold water, contributing to less intense floods (Schneider, 2017). On the other hand, the Ganga basin is the largest among all the basins in the country, with an expansive size of 8,61,452 sq. km. It comprises diverse topography ranging from high-elevated Himalayan mountains to low-lying Gangetic plains. The elevation ranges from less than 5 m to more than 6000 m. Figure 4 illustrates that the Central Ganga, including the lower and middle Yamuna subbasins, is observed to have relatively lower mean normalized flood magnitudes among other regions within the Ganga basin. There is evidence of decreasing average rainfall in these regions (Roxy et al., 2017).

Figure 5 presents the monthly distribution of flood events across different river basins, normalized by the total number of flood events in each basin. The numbers in the legend indicate the total number of recorded flood events in the respective basin. This normalization allows for a direct comparison of seasonal flood occurrence across basins, irrespective of differences in the absolute number of events.

Flood events in India are predominantly observed to be influenced by the southwest monsoon, with 86% of the total 7,500 flood events analyzed occurring during this period, underscoring its critical role in flood generation across India. The southwest monsoon (June - September) accounts for 80% of annual rainfall in the country. It hits the country in two branches: (i) the Arabian Sea branch, first impacting the Western Ghats, and (ii) the Bay of Bengal branch, bringing heavy rainfall to central and eastern India. This monsoonal pattern drives the flood seasonality observed in most river basins, with a peak in flood events occurring between July and September. In particular, West-flowing rivers from Tapi to Kanyakumari basin experience high flood frequencies in July, coinciding with the second

spell of monsoonal rainfall, as observed in Figure 5, when streams are already full, and catchments are saturated. At the same time, most other major basins, including the Ganga, Godavari, Krishna, Narmada, Subernarekha, and Tapi, exhibit a strong flood occurrence within the southwest monsoon period, reinforcing its dominant influence on flood generation. However, exceptions exist in the Cauvery basin and the East-flowing rivers between Pennar and Kanyakumari basin, which have distinct precipitation patterns due to their exposure to both the southwest and northeast monsoons.

The Cauvery basin and East-flowing rivers between Pennar and Kanyakumari basins have varied precipitation timings. The western side of these basins receives substantial rainfall from the southwest monsoon (June - September), and the coastal districts on the eastern side are significantly impacted by the northeast monsoon, which contributes to flooding in October, November, and December in these regions, as shown in Figure 5. For instance, the 2015 floods in Tamil Nadu, one of the deadliest and costliest floods in the history of India, resulted from 4 spells of extreme rainfall in November and December (NDMA, 2017; Vencatesan, 2021). Some of the inland regions of the East-flowing rivers between Pennar and Kanyakumari basin and the east of Cauvery basin also experience peak monthly rainfall in September and October. At the same time, the East-flowing rivers from Mahanadi to Pennar basin may also experience late monsoon flooding, with additional rainfall from the northeast monsoon influencing flood occurrences in late September and October. This highlights the complex interplay between monsoonal systems in shaping flood seasonality across different basins. The observed deviations in flood seasonality in certain basins emphasize the importance of considering regional climatic influences when assessing flood risks.

5. The relationship between flood magnitude and geomorphology

The geomorphology of the catchments affects streamflow generation by influencing processes such as infiltration, overland flow, and subsurface flow, which ultimately determine the timing, magnitude, and variability of streamflow. Spearman correlations in Figure 6 give the first-order relationship between normalized flood magnitude and various geomorphic and climate characteristics of the catchments. Figure 6 illustrates that elongation ratio, maximal flow length, number of the first-order and second-order streams, and their total length, stream order, and catchment perimeter are among the highly correlated geomorphologic variables with normalized flood magnitude. The elongation ratio of the catchments is positively

correlated with the normalized magnitude of the flood events observed over the corresponding catchments, which means an increase in the elongation ratio may increase the normalized flood magnitude. By definition, the elongation ratio is the proportion of the maximum length of the catchment to the diameter of a circle with an equivalent area (Schumm, 1956). Higher ratios suggest a circular shape, while a lower elongation ratio suggests an elongated catchment shape. Circular catchments (higher elongation ratio) produce higher peaks, leading to greater normalized flood magnitudes (Garzon et al., 2023).

As the Maximal flow length - the length along the longest watercourse from the mouth to the head of the channel (Mueller, 1968) - increases, the streamflow takes more time to reach the higher-order streams/outlet, leading to lower normalized flood magnitude. Similarly, streams of higher order, catchment perimeter, and length are associated with larger contributing catchment areas, resulting in increased streamflow travel time and, subsequently, lower normalized flood magnitude. For instance, as mentioned in Section 4, the catchments with high normalized flood magnitudes in the West-flowing rivers from Tapi to Kanyakumari are observed to be small catchments. At the same time, the number of first- and second-order streams in a catchment and their total length is proportional to the travel time of streamflow into the higher-order streams as they flow into and feed the higher-order streams. Hence, an increase in all these variables leads to a decrease in the normalized flood magnitude, i.e., they are negatively correlated.

Box plots are plotted for elongation ratio and maximal flow length for different basins to further explore these highly correlated geomorphological variables in the context of observations made in Section 4 from Figures 2 and 3. Figure 7(a) shows that the West-flowing rivers from Tapi to Kanyakumari basin have the highest median and quartile values of elongation ratio, which is positively correlated with the normalized flood magnitude, compared to other basins in the country. Besides, the same basin has the least median and quartile values of maximal flow length, which is negatively correlated to the normalized flood magnitude. These observations suggest that the West-flowing rivers from Tapi to Kanyakumari basin have highly favorable geomorphologic conditions for having high normalized flood magnitude relative to other basins in the country. At the same time, it is observed that East-flowing rivers from Pennar to Kanyakumari basin are geomorphologically favorable, and the Narmada basin is not favorable for high normalized flood magnitudes. However, despite unfavorable geomorphological conditions for high normalized flood

magnitude in the Narmada basin, the flood events observed over the same region have high normalized magnitude values based on Figure 4. Similarly, though the East-flowing rivers from Pennar to Kanyakumari basin possess favorable geomorphological conditions for high normalized flood magnitudes, the observed flood events exhibit relatively low normalized magnitudes (Figure 4). These observations suggest that while geomorphology plays a crucial role, its relationship with flood magnitude is complex and warrants further investigation. To delve deeper, we examined the influence of climatological variables on normalized flood magnitudes and compared the relative importance of geomorphological and climatic factors in the subsequent sections.

6. The relationship between flood magnitude and climatology

From Figure 6, It is worth mentioning that climatological variables exhibit a stronger correlation compared to other variables. It is observed that Precipitation of wettest quarter, annual Precipitation, and Precipitation of wettest month are the top three highly correlated variables among other climatological variables. Precipitation is the primary driver of floods across India. As anticipated, there is a positive correlation between the normalized flood magnitude and all these three highly correlated variables. A box plot of Precipitation of wettest quarter for various basins is shown in Figure 8. It is observed that the West-flowing rivers from Tapi to Kanyakumari basin has the highest median and quartile values, followed by the Narmada basin, and Subernarekha, Brahmani and Baitarni basin. In contrast, the lowest was observed in East-flowing rivers from Pennar to Kanyakumari basin, followed by Cauvery. These findings explain why the East-flowing rivers from Pennar to Kanyakumari basin has low normalized flood magnitude (Figure 4), though they have a few favorable geomorphological conditions for high values. A similar observation has been made in the case of the Narmada basin, with a few climatology being favorable and geomorphology being unfavorable for increased normalized flood magnitudes. On the other hand, catchments in West-flowing rivers from Tapi to Kanyakumari basin exhibit favorable geomorphology and climatology, leading to the highest observed normalized flood magnitude among others (Figure 4). This proves that a multitude of physiographic variables influences normalized flood magnitude or any other flood characteristic.

7. The relative importance of geomorphology and climatology

This section delves into a multidimensional analysis of catchment characteristics (explanatory variables) on normalized flood magnitude (target variable) using a random forest model and SHAP. The model predictions have an accuracy of 80.2 % with a mean absolute error of 0.11 when compared with their corresponding observation values for 25 % of the total observed events used for testing. Building on the earlier sections highlighting the intricate relationship between floods and catchment characteristics, this analysis helps us find the relative importance of different geomorphology and climatology explanatory variables combinedly and individually on normalized flood magnitude. The SHAP summary plot in Figure 9 (a) reveals that the precipitation of wettest month is the most influential explanatory variable on normalized flood magnitude compared to other explanatory variables with a mean absolute SHAP value of 0.279. It further reveals the positive and strong correlation of Precipitation of wettest month with the normalized flood magnitude, confirming that the higher Precipitation in the wettest month may lead to a higher normalized flood magnitude. Besides, Catchment magnitude, i.e., the number of first-order streams in the catchment boundary (Melton, 1957), the number of second-order streams, and the ratio between their total lengths, are also identified as the essential variables that impact normalized flood magnitude. The first and second-order streams are the outermost tributaries. At the same time, the mean length of all the first streams in a catchment is also important. As the length of the first-order streams increases, the streamflow takes more time to reach the higher-order streams/outlet, leading to lower normalized flood magnitudes at the outlet, as confirmed by the SHAP summary plot. Figure 9(a) depicts that Precipitation in the coldest quarter has also been an influential variable on normalized flood magnitude. The same variable would be important over the floods observed in the Cauvery basin and East-flowing rivers from Pennar to Kanyakumari basin, as these two basins are affected by northeast monsoon from October to December, as discussed in Section 4. At the same time, catchment relief, which refers to the elevation difference between the highest point and the mouth of the catchment (Costa, 1987), can significantly influence flood magnitude. For instance, steeper slopes in the catchment lead to faster precipitation runoff, causing more rapid flows with high velocities in the streams and potentially leading to higher flood peaks.

Figure 9(b) illustrates the relative importance of geomorphological versus climatological variables on normalized flood magnitude at the station level, alongside a box-and-whisker plot comparing the flood magnitudes of stations dominated by each group. In this analysis, 112 catchments are observed to be climatology-dominated while 21 are geomorphology-dominated. The inset box plot reveals that stations dominated by

geomorphology-based influence tend to have a higher median flood magnitude (0.61) compared to those dominated by climatology (0.21), and the geomorphology-dominated stations also show a narrower spread. This suggests that the influence of geomorphology on normalized flood magnitude is relatively stable, and responses are more consistent, whereas the larger spread and extreme values in climatology-dominated stations suggest that climatic factors lead to greater uncertainty or variability in flood magnitudes across stations. This may imply that climatic influences on floods are more unpredictable and that climatological features introduce more uncertainty in modeling flood magnitudes.

When grouped SHAP contributions are aggregated across all stations in the country, climatological variables account for approximately 61% of the total influence, with geomorphological variables contributing around 39%. Further aggregation by basin reveals a similar trend, with climatology generally ranging between 55% and 60% and geomorphology between 35% and 40%. However, in the West-flowing rivers from Tapi to Kanyakumari basin, a more pronounced climatological dominance (70.3%) emerges, underscoring the spatial heterogeneity of flood controls. These findings highlight that geomorphological factors can lead to higher but more consistent flood magnitudes, whereas climatological factors often result in greater variability. Identifying the dominant driver at both station and basin scales is therefore essential for tailoring flood management and mitigation strategies.

8. The empirical relationship between flood magnitude and drainage area

This section presents an empirical approach to estimate normalized flood magnitude for ungauged basins based on their contributing drainage area using envelope curves. Figure 10 illustrates the upper bound of observed normalized flood magnitudes across Indian basins using envelope curves, represented by black dots, which are derived from the database developed in this study.. It is accompanied by envelope curves of maximum floods reported in Global (Kuntla et al., 2022), USA (Saharia et al., 2017), and Europe (Marchi et al., 2010). The reduced discharge (α) value for India is observed to be 875.9, while the scaling coefficient (β) is observed to be -0.43. These values, along with Eq. 1, help in first-order approximations of the normalized flood magnitude at ungauged stations based on their catchment area. Besides, comparing these numbers with the values of Global (α=4130.6, β=-0.58) (Kuntla et al., 2022), USA (α=108, β=-0.47) (Saharia et al., 2017), Europe (α=97, β=-0.40) (Marchi et al., 2010), dictates that the flood magnitudes in India is much higher than USA, and Europe, but not the highest globally. On the other hand, the rate of change in flood

magnitudes with unit catchment area in India is comparable with Europe, while the rate of change is lower than in the USA and the globe.

This empirical relationship based on envelope curve ($Q = 875.9A^{-0.43}$) can serve as a useful tool for rapid flood magnitude estimation in ungauged regions in India and for benchmarking extreme events. However, it is important to note that envelope curves represent an upper bound and do not account for event-to-event variability or catchment-specific features beyond area.

9. The relationship between flood magnitude and soil and lithology

Although the machine learning model in this study focused on geomorphological and climatological variables due to their quantitative nature and compatibility with machine learning algorithms, we include a exploratory analysis of lithological and soil characteristics in this section to provide a broader understanding of factors influencing normalized flood magnitude. Soil has been identified as an essential variable influencing flood characteristics in a catchment (Ahn & Merwade, 2016; Basri et al., 2022; Kuntla, 2021). Different soil types have varying properties like infiltration capacity, saturation and percolation, and runoff potential, leading to differences in their interaction with water. Boxplots of normalized flood magnitude for different soil types have been plotted (Figure 11(a)) to examine which soil types produce floods with high and low normalized flood magnitudes in India. Catchments dominated by Leptosols in their boundary (more than 50 %) are observed to have high median and quartile values of flood magnitude normalized by catchment area. Leptosols is manifested by a very shallow profile depth dominated by sand, and gravel in texture and are associated with increased susceptibility to erosion depending on climate and topography (Ebabu et al., 2023). Therefore, they can result in more runoff due to their free-draining nature and inability to hold water. On the other hand, Lixisols and Luvisols, which are classified as Hydrologic Soil Group A, are characterized by high infiltration rates and low runoff potential (Ouedraogo et al., 2017).

The lithology of an area also influences flood characteristics as the geological composition of an area, including the types of rocks present, affects how the water interacts with the landscapes (Gaume & Borga, 2008; He et al., 2021). Some of the essential characteristics of rocks that influence the flood characteristics are: The permeability of rocks determines water infiltration and storage; Aquifer storage capacity influences streamflow

regulation, affecting flood duration and magnitude; Rock types also impact channel stability, erosion patterns, and landform configuration, influencing flood dynamics. Additionally, lithology plays a role in shaping topography and landforms, affecting the overall flow of water. Figure 11(b) dictates that floods in catchments with acid plutonic and metamorphic rocks have relatively high and low median normalized magnitude values, respectively. Acid plutonic rocks are igneous rocks formed from magma that has cooled below the surface of the Earth. They are characterized by their high silica content and are typically light in color. Examples of acid-plutonic rocks include granite. Acid plutonic rocks generally have low permeability, meaning water is less likely to move through them quickly. This can lead to increased surface runoff and potentially contribute to flooding in areas where these rocks are prevalent. On the other hand, the hydrologic runoff response of Metamorphic rocks can vary depending on the specific type of metamorphic rock and its characteristics, including the mineral composition of the parent rock, temperature, pressure, and the presence of fluids, such as water, which in turn influence the porosity, permeability, and structural features of the rock, subsequently, hydrologic runoff response (Earle, 2019).

10. Conclusions

This study performs a spatio-temporally comprehensive characterization study based on a unique event-based flooding database over India developed by combining streamflow time series records and operational definition of flooding thresholds. The database is further enhanced with a diverse set of physiographic variables to analyze their relationship with flood magnitude. By incorporating a machine learning based approach with interpretability methods, this study offers robust insights into the dominant characteristics and factors influencing flood hydroclimatology. Some of the key insights from the spatial and temporal analysis of flood events are:

- The normalized flood magnitude is high along the Western Ghats of the country in West-flowing rivers from Tapi to Kanyakumari Basin with a median value of 0.79, while lower values are observed in the Cauvery Basin (median = 0.02).
- The higher values along the Western Ghats suggest this region is a hotspot for severe floods, as the volume of floods is high per unit area of the catchment.
- Southwest monsoon (June – September) triggers floods across most basins, accounting for 86% of the 7500 flood events studied. However, the northeast monsoon (October to December) dominates the flood occurrence in Cauvery and East-flowing rivers from Pennar to Kanyakumari Basins.

On the other hand, our investigations into the relationships between flood magnitude and catchment characteristics reveal first-order relationships and intricate connections among them.

- The exploration of the correlations between normalized flood magnitude and catchment characteristics has yielded that climatological variables are highly correlated compared to geomorphology, with precipitation of wettest quarter, annual precipitation, and precipitation of wettest month being the top three highly correlated climatological variables, among others. On the other hand, the elongation ratio, maximal flow length, number of the first-order and second-order streams, and their total length, stream order, and catchment perimeter are among the highly correlated geomorphologic explanatory variables.
- The catchments dominated by Leptosols soil in their boundary have high median normalized flood magnitudes (median = 1.45). Meanwhile, catchments with lixisols (median =0.02) and luvisols (median =0.03) are characterized by lower values due to their low runoff potential. At the same time, the lithology analysis dictates that floods in catchments dominated by Acid plutonic rocks have relatively high normalized magnitude values (median = 0.31), among others.
- Developed envelope curves provide an empirical formula that is a first-order approximation of normalized flood magnitudes at ungauged stations in the country as a function of a catchment area ($Q = 875.9A^{-0.43}$). The magnitude of the maximum floods in India are higher than those reported in the USA and Europe.
- SHAP interpretation enabled us to unravel that the precipitation of wettest month is the most influential explanatory variable to model normalized flood magnitude compared to other catchment characteristics. It is followed by the number of first-order and second-order streams in the catchment boundary and the ratio between their total lengths.
- Grouped feature importance analysis indicates that climatological variables exert a substantially greater influence (61%) on the normalized flood magnitude compared to geomorphological variables (39%), underscoring the dominant role of climatic factors in driving flood behavior across the nation. Furthermore, the station-wise grouped analysis suggests that dominant geomorphology-based influence can lead to higher but more consistent flood magnitudes, whereas climatological factors often result in greater variability.

This study has certain limitations, including the non-inclusion of data from key basins such as Brahmaputra, Indus, and Mahanadi due to data unavailability. Additionally, the study does not capture flash flood characteristics due to a lack of sub-daily data. The data for the Brahmaputra and Indus basins is classified and not openly accessible as these are transboundary rivers. Although streamflow data for several gauge stations in the Mahanadi basin is available, the absence of official flood thresholds, as defined in Section 2.1, prevents the delineation and inclusion of floods from this basin in our study. These will guide future improvements to enhance the database's scope and applicability. Nevertheless, this study marks one of the first large-sample hydrological studies in India, leveraging a standardized flood event database to bridge critical data gaps and enable systematic, nationwide flood characterization. Furthermore, the developed database provides a valuable resource for future hydrological studies, flood risk assessment, policy-making, and operational flood management, setting the stage for more data-driven approaches in Indian flood science. Also, the robust insights gained from this large-sample study provide a strong foundation for pushing data-centric science in the prediction of floods at ungauged locations.

Acknowledgments

This research was conducted in the HydroSense lab (https://hydrosense.iitd.ac.in/) of IIT Delhi and the authors acknowledge the IIT Delhi High Performance Computing facility for providing computational and storage resources. Dr. Manabendra Saharia gratefully acknowledges financial support for this work through grants from ISRO Space Technology Cell (STC0374/RP04139); Ministry of Earth Sciences Monsoon Mission III (RP04574); Ministry of Earth Sciences (RP04741); and DST IC-IMPACTS (RP04558). The authors gratefully acknowledge the Central Water Commission (CWC), National Water Informatics Centre (NWIC), and the Ministry of Jal Shakti (MoJS) for providing the streamflow datasets used in this study.

Compliance with Ethical Standards

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author Contributions

Sai Kiran Kuntla: Conceptualization, Methodology, Formal analysis, Data Curation, Writing - Original Draft

Manabendra Saharia: Conceptualization, Writing - Review & Editing

Data Availability

The data that support the findings of this study were derived from resources available in the public domain:

- Streamflow data from India Water Resources Information System (WRIS), Central Water Commission: https://indiawris.gov.in/
- Global Distributed Basin Characteristics (GDBC): https://figshare.com/s/6cd00491b850bad716d7
- Worldclim: https://worldclim.org/data/worldclim21.html.
- Lithology: https://doi.pangaea.de/10.1594/PANGAEA.788537
- Soil: https://doi.org/10.1371/journal.pone.0169748

References

Abdullah, J., Muhammad, N. S., Muhammad, S. A., & Julien, P. Y. (2019). Envelope curves for the specific discharge of extreme floods in Malaysia. *Journal of Hydro-Environment Research*, *25*, 1–11. https://doi.org/10.1016/j.jher.2019.05.002

Addor, N., Do, H. X., Alvarez-Garreton, C., Coxon, G., Fowler, K., & Mendoza, P. A. (2020). Large-sample hydrology: Recent progress, guidelines for new datasets and grand challenges. *Hydrological Sciences Journal*, *65*(5), 712–725. https://doi.org/10.1080/02626667.2019.1683182

Addor, N., Nearing, G., Prieto, C., Newman, A. J., Vine, N. L., & 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

Ahn, K.-H., & Merwade, V. (2016). Role of Watershed Geomorphic Characteristics on Flooding in Indiana, United States. *Journal of Hydrologic Engineering*, *21*(2), 05015021. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001289

Au, Q., Herbinger, J., Stachl, C., Bischl, B., & Casalicchio, G. (2022). Grouped feature importance and combined features effect plot. *Data Mining and Knowledge Discovery*, *36*(4), 1401–1450. https://doi.org/10.1007/s10618-022-00840-5

Basri, H., Syakur, S., Azmeri, A., & Fatimah, E. (2022). Floods and their problems: Land uses and soil types perspectives. *IOP Conference Series: Earth and Environmental Science*, *951*(1), 012111. https://doi.org/10.1088/1755-1315/951/1/012111

Bhatt, S., & Ahmed, S. A. (2014). Morphometric analysis to determine floods in the Upper Krishna basin using Cartosat DEM. *Geocarto International*, *29*(8), 878–894. https://doi.org/10.1080/10106049.2013.868042

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Castellarin, A. (2007). Probabilistic envelope curves for design flood estimation at ungauged sites. *Water Resources Research*, *43*(4). https://doi.org/10.1029/2005WR004384

Central Water Commission, M. of jal shakti, Government of India. (2018). *Hydro Meteorological Water Year Book*. https://cwc.gov.in/water-year-book-sediment-year-book-water-quality-year-book-meteorological-year-book

Chaubey, P. K., Mall, R. K., Jaiswal, R., & Payra, S. (2022). Spatio-Temporal Changes in Extreme Rainfall Events Over Different Indian River Basins. *Earth and Space Science*, *9*(3), e2021EA001930. https://doi.org/10.1029/2021EA001930

Costa, J. E. (1987). Hydraulics and basin morphometry of the largest flash floods in the conterminous United States. *Journal of Hydrology*, *93*(3), 313–338. https://doi.org/10.1016/0022-1694(87)90102-8

Das, S., & Scaringi, G. (2021). River flooding in a changing climate: Rainfall-discharge trends, controlling factors, and susceptibility mapping for the Mahi catchment, Western India. *Natural Hazards*, *109*(3), 2439–2459. https://doi.org/10.1007/s11069-021-04927-y

Dhar, O. N., & Nandargi, S. (2000). A study of floods in the Brahmaputra basin in India. *International Journal of Climatology*, *20*(7), 771–781. https://doi.org/10.1002/1097-0088(20000615)20:7<771::AID-JOC518>3.0.CO;2-Z

Earle, S. (2019). *Physical Geology* (2nd ed.). BCcampus. https://opentextbc.ca/physicalgeology2ed/

Ebabu, K., Taye, G., Tsunekawa, A., Haregeweyn, N., Adgo, E., Tsubo, M., Fenta, A. A., Meshesha, D. T., Sultan, D., Aklog, D., Admasu, T., van Wesemael, B., & Poesen, J. (2023). Land use, management and climate effects on runoff and soil loss responses in

the highlands of Ethiopia. *Journal of Environmental Management*, *326*, 116707. https://doi.org/10.1016/j.jenvman.2022.116707

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Garzon, L. F. L., Johnson, M. F., Mount, N., & Gomez, H. (2023). Exploring the effects of catchment morphometry on overland flow response to extreme rainfall using a 2D hydraulic-hydrological model (IBER). *Journal of Hydrology*, *627*, 130405. https://doi.org/10.1016/j.jhydrol.2023.130405

Gaume, E., Bain, V., Bernardara, P., Newinger, O., Barbuc, M., Bateman, A., Blaškovičová, L., Blöschl, G., Borga, M., Dumitrescu, A., Daliakopoulos, I., Garcia, J., Irimescu, A., Kohnova, S., Koutroulis, A., Marchi, L., Matreata, S., Medina, V., Preciso, E., … Viglione, A. (2009). A compilation of data on European flash floods. *Journal of Hydrology*, *367*(1), 70–78. https://doi.org/10.1016/j.jhydrol.2008.12.028

Gaume, E., & Borga, M. (2008). Post-flood field investigations in upland catchments after major flash floods: Proposal of a methodology and illustrations. *Journal of Flood Risk Management*, *1*(4), 175–189. https://doi.org/10.1111/j.1753-318X.2008.00023.x

GFCC, Govt. of I. (2004). *Flood managment guidelines of GFCC 2004*. Ganga Flood Control Commission. https://gfcc.gov.in/sites/default/files/Flood%20Management%20Guidelines%20of%20GFCC%202004.pdf

Gourley, J. J., Hong, Y., Flamig, Z. L., Arthur, A., Clark, R., Calianno, M., Ruin, I., Ortel, T., Wieczorek, M. E., Kirstetter, P.-E., Clark, E., & Krajewski, W. F. (2013). A Unified Flash Flood Database across the United States. *Bulletin of the American Meteorological Society*, *94*(6), 799–805. https://doi.org/10.1175/BAMS-D-12-00198.1

Hartmann, J., & Moosdorf, N. (2012). Global Lithological Map Database v1.0 (gridded to 0.5° spatial resolution) [Dataset]. In *Supplement to: Hartmann, Jens; Moosdorf, Nils (2012): The new global lithological map database GLiM: A representation of rock properties at the Earth surface. Geochemistry, Geophysics, Geosystems, 13, Q12004, https://doi.org/10.1029/2012GC004370*. PANGAEA. https://doi.org/10.1594/PANGAEA.788537

He, J., Liu, X., Zhu, X., Jiang, T., He, H., Zhou, L., Liu, Q., Zhu, Y., & Liu, L. (2021). Water-flooding characteristics of lithologic reservoir in Ordos basin. *Scientific Reports*, *11*(1), Article 1. https://doi.org/10.1038/s41598-021-82035-4

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Herschy, R. W. (2002). The world's maximum observed floods. *Flow Measurement and Instrumentation*, *13*(5), 231–235. https://doi.org/10.1016/S0955-5986(02)00054-7

Huang, Z., Wu, H., Adler, R. F., Schumann, G., Gourley, J. J., Kettner, A., & Nanding, N. (2021). Multisourced Flood Inventories over the Contiguous United States for Actual and Natural Conditions. *Bulletin of the American Meteorological Society*, *102*(6), 1133–1149. https://doi.org/10.1175/BAMS-D-20-0001.1

Javier, J. R. N., Smith, J. A., England, J., Baeck, M. L., Steiner, M., & Ntelekos, A. A. (2007). Climatology of extreme rainfall and flooding from orographic thunderstorm systems in the upper Arkansas River basin. *Water Resources Research*, *43*(10). https://doi.org/10.1029/2006WR005093

Kale, V. S., Hire, P., & Baker, V. R. (1997). Flood Hydrology and Geomorphology of Monsoon-dominated Rivers: The Indian Peninsula. *Water International*, *22*(4), 259–265. https://doi.org/10.1080/02508069708686717

Kuntla, S. K. (2021). An era of Sentinels in flood management: Potential of Sentinel-1, -2, and -3 satellites for effective flood management. *Open Geosciences*, *13*(1), 1616–1642. https://doi.org/10.1515/geo-2020-0325

Kuntla, S. K., Saharia, M., & Kirstetter, P. (2022). Global-scale characterization of streamflow extremes. *Journal of Hydrology*, *615*, 128668. https://doi.org/10.1016/j.jhydrol.2022.128668

Kuntla, S. K., Saharia, M., Prakash, S., & Villarini, G. (2023). Precipitation inequality exacerbates streamflow inequality, but dams moderate it. *Science of The Total Environment*, 169098. https://doi.org/10.1016/j.scitotenv.2023.169098

Kupzig, J., Kupzig, N., & Flörke, M. (2024). Regionalization in global hydrological models and its impact on runoff simulations: A case study using WaterGAP3 (v 1.0.0). *Geoscientific Model Development*, *17*(17), 6819–6846. https://doi.org/10.5194/gmd-17-6819-2024

Lun, D., Viglione, A., Bertola, M., Komma, J., Parajka, J., Valent, P., & Blöschl, G. (2021). Characteristics and process controls of statistical flood moments in Europe – a data-based analysis. *Hydrology and Earth System Sciences*, *25*(10), 5535–5560. https://doi.org/10.5194/hess-25-5535-2021

Lundberg, S., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *arXiv:1705.07874 [Cs, Stat]*. http://arxiv.org/abs/1705.07874

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), Article 1. https://doi.org/10.1038/s42256-019-0138-9

Magotra, B., Prakash, V., Saharia, M., Getirana, A., Kumar, S., Pradhan, R., Dhanya, C. T., Rajagopalan, B., Singh, R. P., Pandey, A., & Mohapatra, M. (2024). Towards an Indian land data assimilation system (ILDAS): A coupled hydrologic-hydraulic system for water balance assessments. *Journal of Hydrology*, *629*, 130604. https://doi.org/10.1016/j.jhydrol.2023.130604

Marchi, L., Borga, M., Preciso, E., & Gaume, E. (2010). Characterisation of selected extreme flash floods in Europe and implications for flood risk management. *Journal of Hydrology*, *394*(1), 118–133. https://doi.org/10.1016/j.jhydrol.2010.07.017

Melton, M. A. (with Columbia University Libraries). (1957). *An analysis of the relations among elements of climate, surface properties, and geomorphology*. New York : Department of Geology, Columbia University. http://archive.org/details/analysisofrelati00melt

Mimikou, M. (1984). Envelope curves for extreme flood events in northwestern and western Greece. *Journal of Hydrology*, *67*(1), 55–66. https://doi.org/10.1016/0022-1694(84)90232-4

Mueller, J. E. (1968). An Introduction to the Hydraulic and Topographic Sinuosity Indexes. *Annals of the Association of American Geographers*, *58*(2), 371–385. https://doi.org/10.1111/j.1467-8306.1968.tb00650.x

Nanditha, J. S., & Mishra, V. (2022). Multiday Precipitation Is a Prominent Driver of Floods in Indian River Basins. *Water Resources Research*, *58*(7), e2022WR032723. https://doi.org/10.1029/2022WR032723

NDMA, Govt. of I. (2008). *National disaster management guidelines: Management of floods*. National Disaster Management Authority. https://nidm.gov.in/PDF/guidelines/floods.pdf

NDMA, Govt. of I. (2017). *Tamil Nadu Floods: Lessons learnt and best practices*. National Disaster Management Authority (NDMA). https://ndma.gov.in/sites/default/files/PDF/Reports/TAMIL-NADU-FLOODS-english.pdf

Newman, R., & Noy, I. (2023). The global costs of extreme weather that are attributable to climate change. *Nature Communications*, *14*(1), Article 1. https://doi.org/10.1038/s41467-023-41888-1

Ouedraogo, J., Ouedraogo, E., Traore, M., Youl, S., & Nacro, H. B. (2017). Interaction between the management of the soil fertility and macrofauna reduces runoff on a lixisol in the north-sudanian zone of Burkina Faso. *Experimental Agriculture*, *53*(1), 12–26. https://doi.org/10.1017/S0014479716000016

Peluso, L. M., Mateus, L., Penha, J., Súarez, Y., & Lemes, P. (2023). Climate change may reduce suitable habitat for freshwater fish in a tropical watershed. *Climatic Change*, *176*(4), 44. https://doi.org/10.1007/s10584-023-03526-z

Peters-Lidard, C. D., Clark, M., Samaniego, L., Verhoest, N. E. C., van Emmerik, T., Uijlenhoet, R., Achieng, K., Franz, T. E., & Woods, R. (2017). Scaling, similarity, and the fourth paradigm for hydrology. *Hydrology and Earth System Sciences*, *21*(7), 3701–3713. https://doi.org/10.5194/hess-21-3701-2017

Rajamani, V., Tripathi, J. K., & Malviya, V. P. (2009). Weathering of lower crustal rocks in the Kaveri river catchment, southern India: Implications to sediment geochemistry. *Chemical Geology*, *265*(3), 410–419. https://doi.org/10.1016/j.chemgeo.2009.05.007

Rana, V. K., & Suryanarayana, T. M. V. (2021). Estimation of flood influencing characteristics of watershed and their impact on flooding in data-scarce region. *Annals of GIS*, *27*(4), 397–418. https://doi.org/10.1080/19475683.2021.1960603

Roxy, M. K., Ghosh, S., Pathak, A., Athulya, R., Mujumdar, M., Murtugudde, R., Terray, P., & Rajeevan, M. (2017). A threefold rise in widespread extreme rain events over central India. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/s41467-017-00744-9

Saharia, M., Jain, A., Baishya, R. R., Haobam, S., Sreejith, O. P., Pai, D. S., & Rafieeinasab, A. (2021). India flood inventory: Creation of a multi-source national geospatial database to facilitate comprehensive flood research. *Natural Hazards*, *108*(1), 619–633. https://doi.org/10.1007/s11069-021-04698-6

Saharia, M., Kirstetter, P.-E., Vergara, H., Gourley, J. J., & Hong, Y. (2017). Characterization of floods in the United States. *Journal of Hydrology*, *548*, 524–535. https://doi.org/10.1016/j.jhydrol.2017.03.010

Salas, F. R., Somos-Valenzuela, M. A., Dugger, A., Maidment, D. R., Gochis, D. J., David, C. H., Yu, W., Ding, D., Clark, E. P., & Noman, N. (2018). Towards Real-Time Continental Scale Streamflow Simulation in Continuous and Discrete Space. *JAWRA Journal of the American Water Resources Association*, *54*(1), 7–27. https://doi.org/10.1111/1752-1688.12586

Schneider, K. (2017, April 19). *Water Scarcity Causes Cauvery Delta Anguish*. Circle of Blue. https://www.circleofblue.org/2017/world/water-scarcity-causes-cauvery-delta-anguish/

Schumm, S. A. (1956). Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *GSA Bulletin*, *67*(5), 597–646. https://doi.org/10.1130/0016-7606(1956)67[597:EODSAS]2.0.CO;2

Sharma, A., & Rajamani, V. (2000). Weathering of gneissic rocks in the upper reaches of Cauvery river, south India: Implications to neotectonics of the region. *Chemical Geology*, *166*(3), 203–223. https://doi.org/10.1016/S0009-2541(99)00222-3

Shen, X., Anagnostou, E. N., Mei, Y., & Hong, Y. (2017). A global distributed basin morphometric dataset. *Scientific Data*, *4*(1), Article 1. https://doi.org/10.1038/sdata.2016.124

Sood, A., & Smakhtin, V. (2015). Global hydrological models: A review. *Hydrological Sciences Journal*, *60*(4), 549–565. https://doi.org/10.1080/02626667.2014.950580

Stein, L., Clark, M. P., Knoben, W. J. M., Pianosi, F., & Woods, R. A. (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), e2020WR028300. https://doi.org/10.1029/2020WR028300

Vencatesan, A. (2021, July 1). *From Rains to Floods: A Case of Chennai in 2015*. Environment & Society Portal. https://doi.org/10.5282/rcc/9323

Ward, P. J., Jongman, B., Salamon, P., Simpson, A., Bates, P., De Groeve, T., Muis, S., de Perez, E. C., Rudari, R., Trigg, M. A., & Winsemius, H. C. (2015). Usefulness and limitations of global flood risk models. *Nature Climate Change*, *5*(8), Article 8. https://doi.org/10.1038/nclimate2742

Yamazaki, D., Kanae, S., Kim, H., & Oki, T. (2011). A physically based description of floodplain inundation dynamics in a global river routing model. *Water Resources Research*, *47*(4). https://doi.org/10.1029/2010WR009726

Zhang, J., Dai, Q., Nanding, N., & Han, D. (2022). Exploring the effect of catchment morphology on streamflow characteristics with virtual experiments. *Journal of Hydrology*, *608*, 127606. https://doi.org/10.1016/j.jhydrol.2022.127606

Table 1. Number of gauge stations in various basins.

Basin	No. of gauge stations	No. of flood events	Gauge stations/Flood events
Ganga	47	3594	74.47
Subernarekha, Brahmani and Baitarni	6	202	33.67
East-flowing rivers from Mahanadi to Pennar	4	178	44.5
East-flowing rivers from Pennar to Kanyakumari	4	54	13.5
Cauvery	10	766	76.6
West-flowing rivers from Tapi to Kanyakumari	25	1100	44
Krishna	24	632	26.33
Godavari	6	448	74.67
Tapi	4	426	106.5
Narmada	3	100	33.33

Figure captions

Figure 1. Graphical illustration of the classification of a flood event from streamflow data based on flooding thresholds by the Central Water Commission.

Figure 2. Overview of the workflow of this study.

Figure 3. Spatial distribution of gauge stations where floods are observed and analyzed in this study. The points are color-coded by the mean flood magnitude of all the events recorded over the corresponding station normalized by its catchment area.

Figure 4. Box-and-whisker plot of normalized flood magnitude for different basins. The box spans the interquartile range (i.e., 25th and 75^{th} percentiles). The whiskers are the two vertical lines outside the box extended until the extremes, 5^{th} and 95^{th} percentiles. A horizontal line inside the box marks the median, displaying its value above the same line.

Figure 5. Monthly distribution of flood events normalized by the total number of flood events in different basins.

Figure 6. Spearman-correlation coefficients between the normalized flood magnitude of all flood events and multiple catchment characteristics. The numbers range from 1 to -1. Positive (negative) numbers reflect positive (negative) correlations, with 1 (-1) representing the highest positive (negative) correlation. A value of 0 indicates that there is no correlation.

Figure 7. Box-and-whisker plot of (a) Elongation ratio and (b) Maximal flow length over different basins. *Please refer to the caption of Figure 4 for the description of the box plots.*

Figure 8. Box-and-whisker plot of precipitation of wettest quarter over different basins. *Please refer to the caption of Figure 4 for the description of the box plots.*

Figure 9. (a) SHAP summary plot with the top 10 important explanatory variables for the random forest model. The explanatory variables are arranged in descending order of importance along the y-axis. Mean absolute SHAP values for the associated explanatory variables are indicated by the value to the right of the explanatory variable's name along the y-axis. The horizontal placement of the dots represents the strength of its influence on the model. Zero values on the x-axis indicate no impact, and positive (negative) values towards the right (left) side indicate positive (negative) impact. The color of the dots represents the value of the relevant explanatory variables for that observation in the dataset, whether it is high (red) or low (blue). **(b)** Station-wise grouped feature importance of geomorphological variables and climatological variables on normalized flood magnitude and a Box-and-Whisker plot of normalized flood magnitude between corresponding geomorphological dominated and climatological dominated gauge stations. *Please refer to the caption of Figure 4 for the description of the box plots.*

Figure 10. Envelope curves for observed floods in India. The envelope curves for Global, USA, and Europe are retrieved from Kuntla et al. (2022), Saharia et al. (2017), and Marchi et al. (2010), respectively.

Figure 11. Box-and-whisker plot of normalized flood magnitude for different (a) Soil and (b) Lithology types. *Please refer to the caption of Figure 4 for the description of the box plots.*

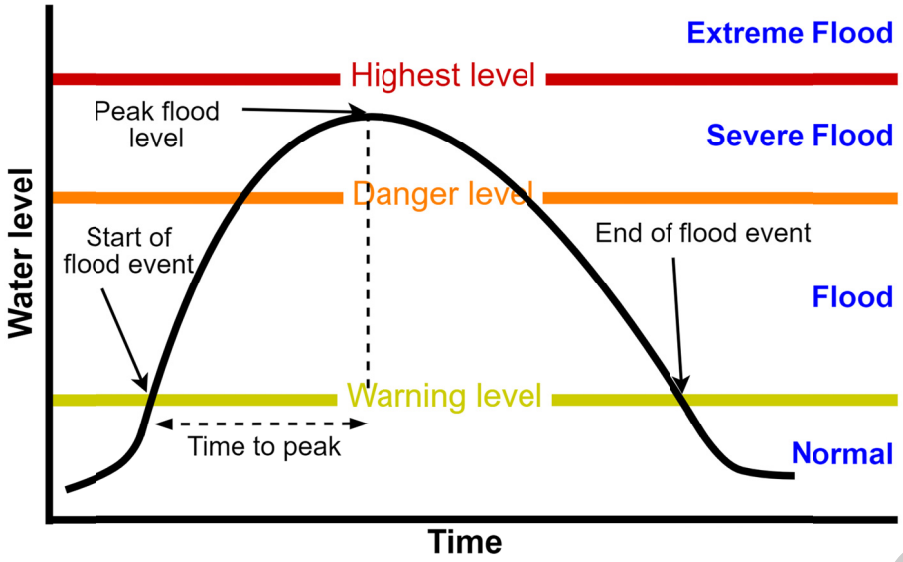


Figure 1

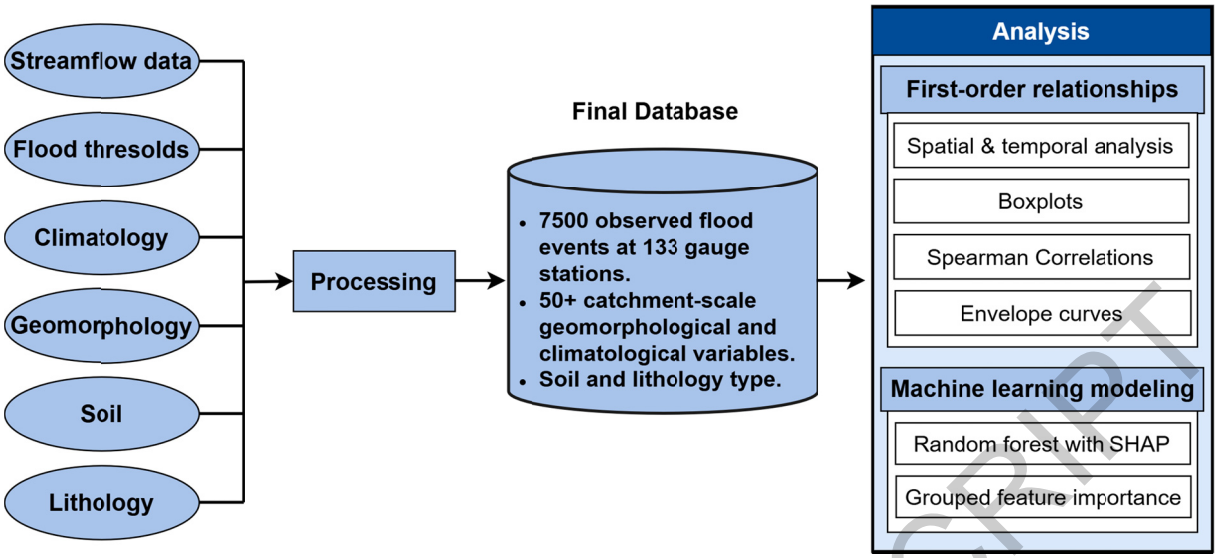
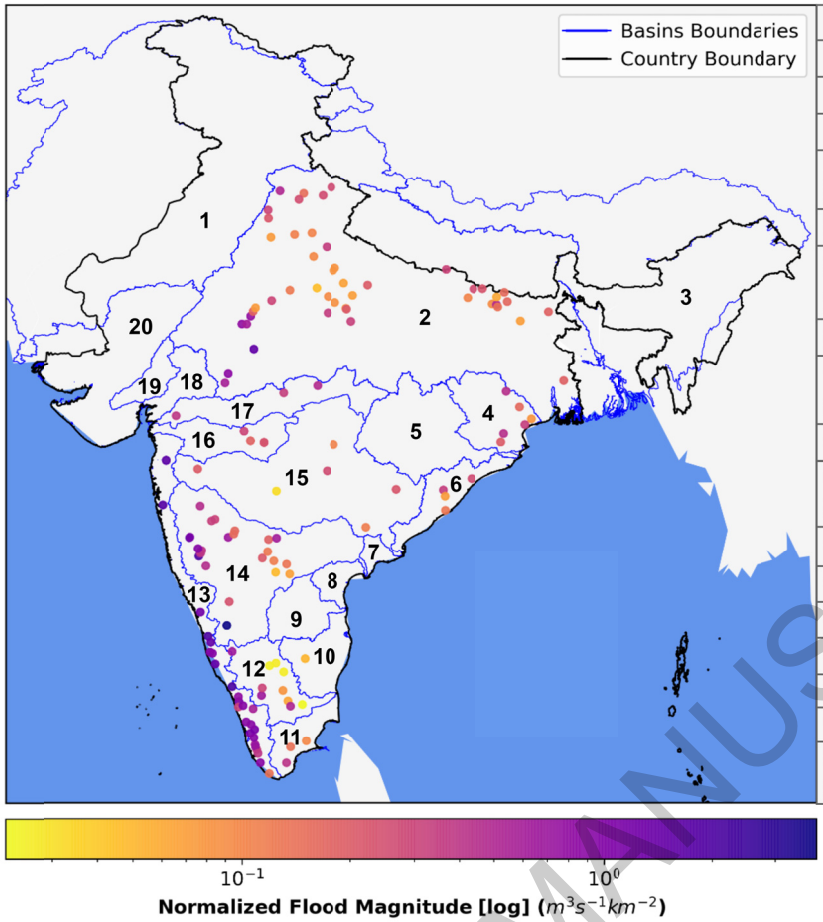


Figure 2



Basin
1. Indus
2. Ganga
3. Brahmaputra
4. Subernarekha, Brahmani and Baitarni
5. Mahanadi
6, 7, 8. East-flowing rivers from Mahanadi to Pennar
9. Pennar
10, 11. East-flowing rivers from Pennar to Kanyakumari
12. Cauvery
13. West-flowing rivers from Tapi to Kanyakumari
14. Krishna
15. Godavari
16. Tapi
17. Narmada
18. Mahi
19. Sabarmati
20. West-flowing rivers of Kutch and Saurashtra including Luni

Figure 3

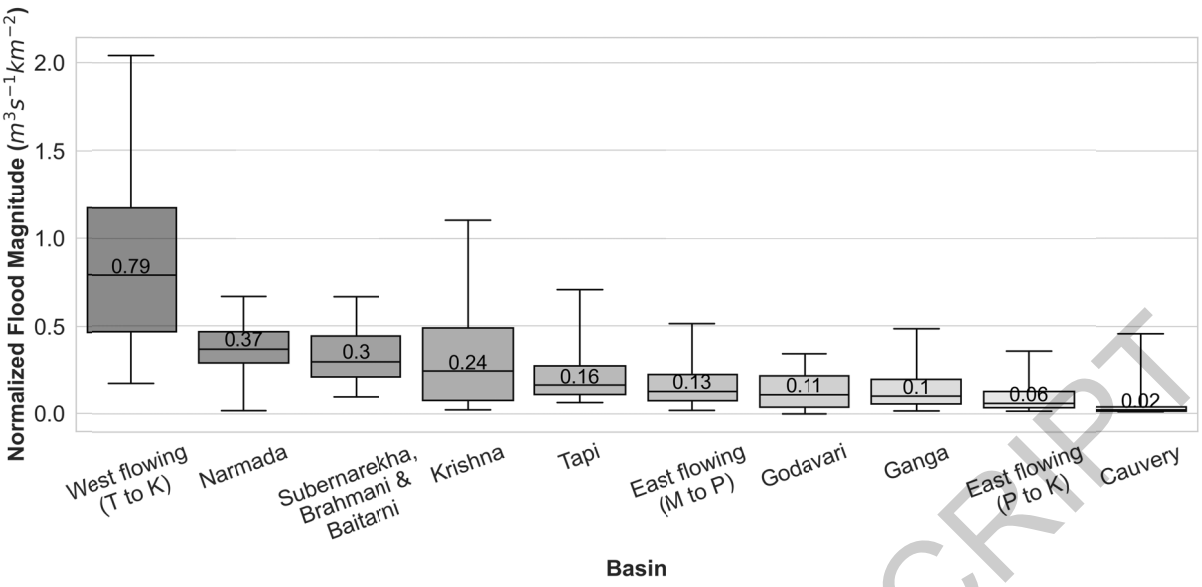


Figure 4

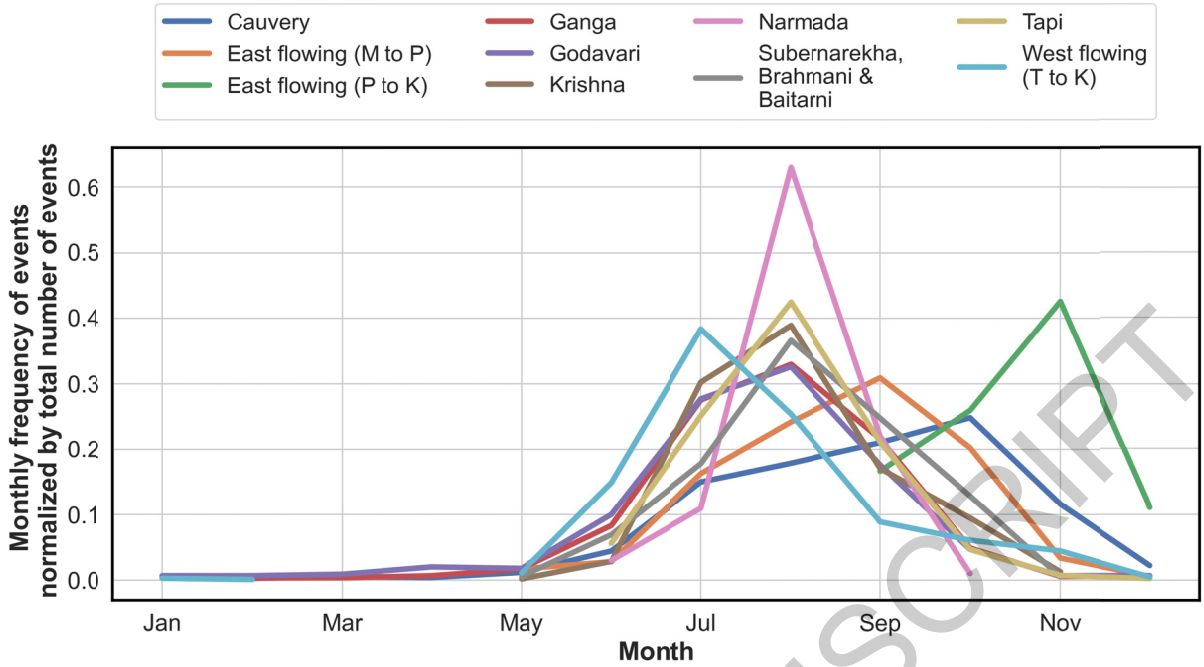


Figure 5

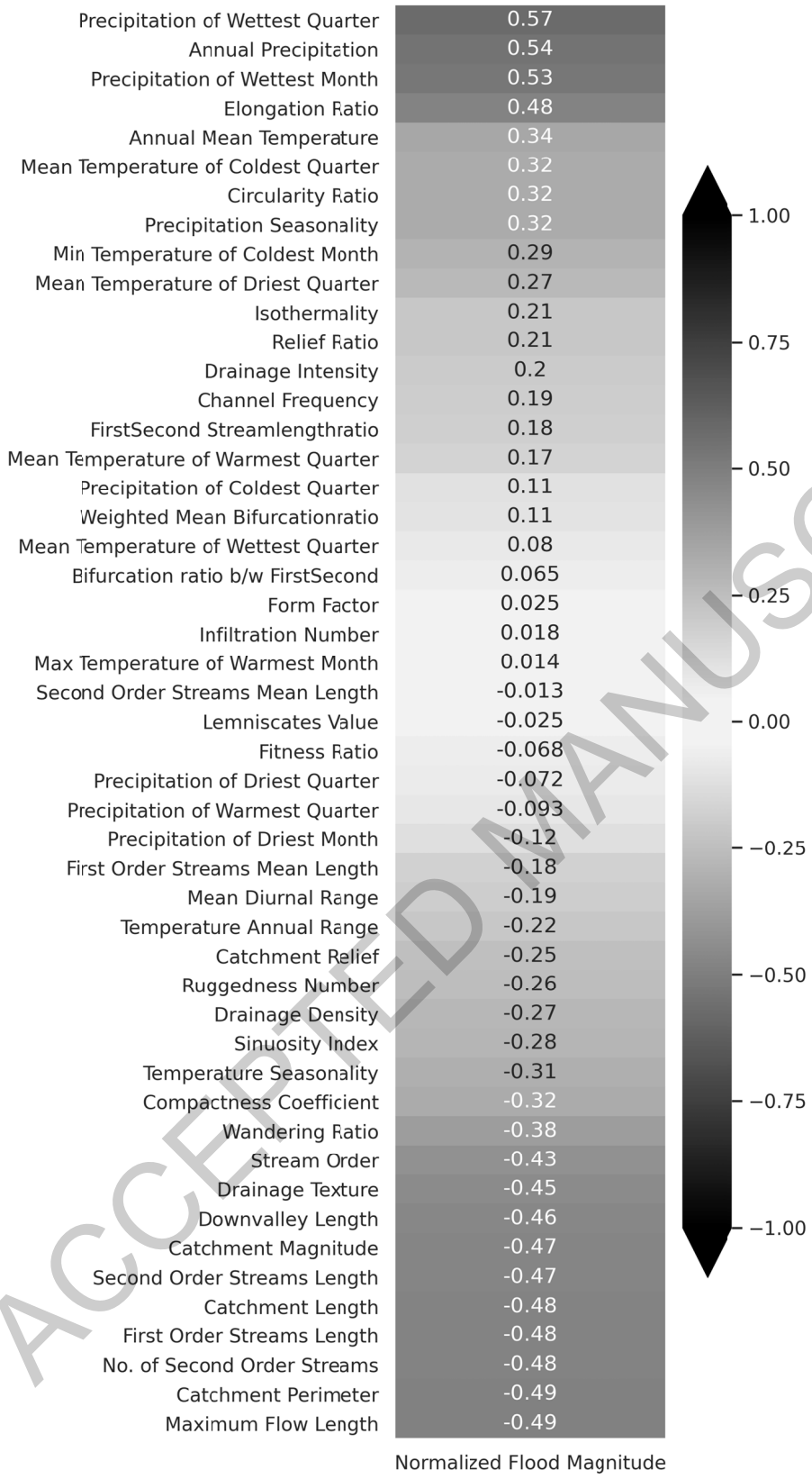


Figure 6

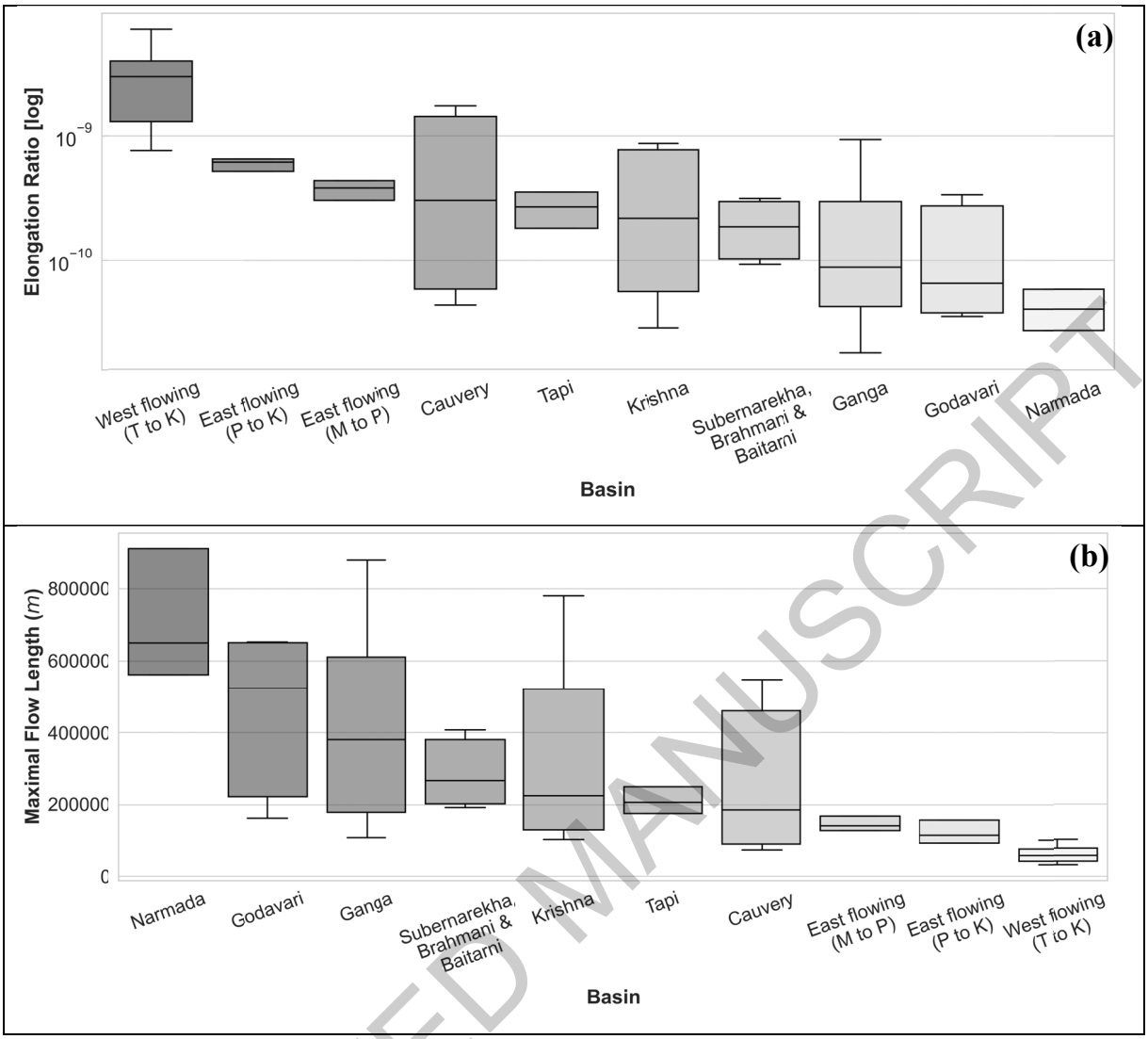


Figure 7

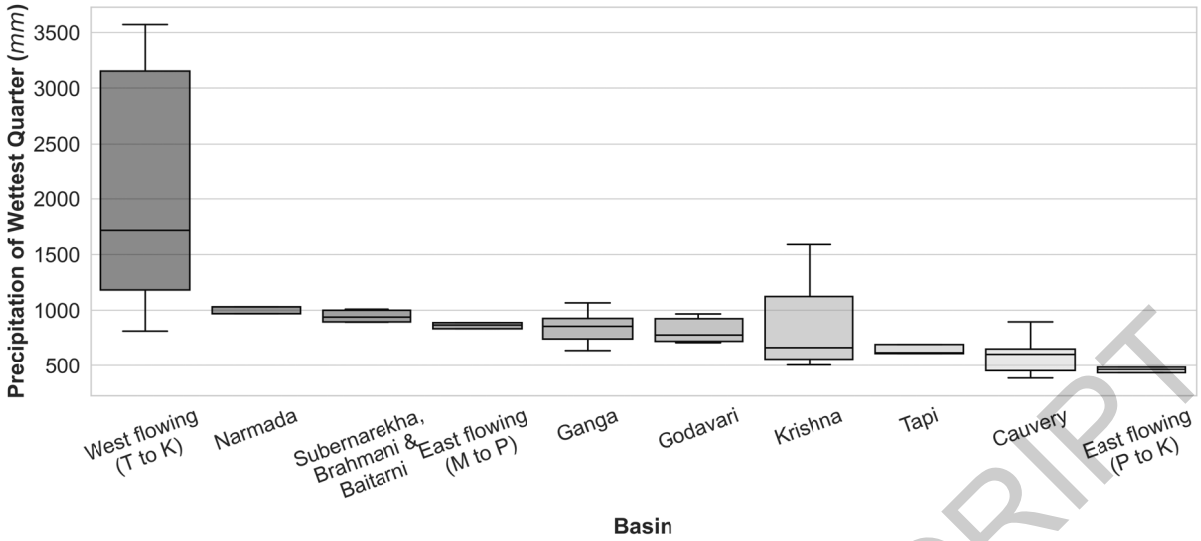
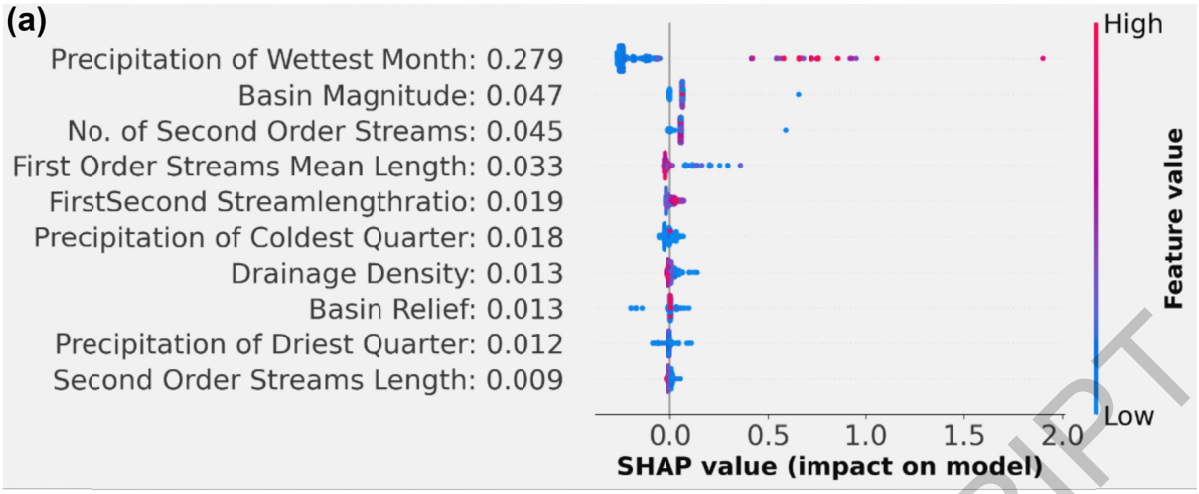


Figure 8



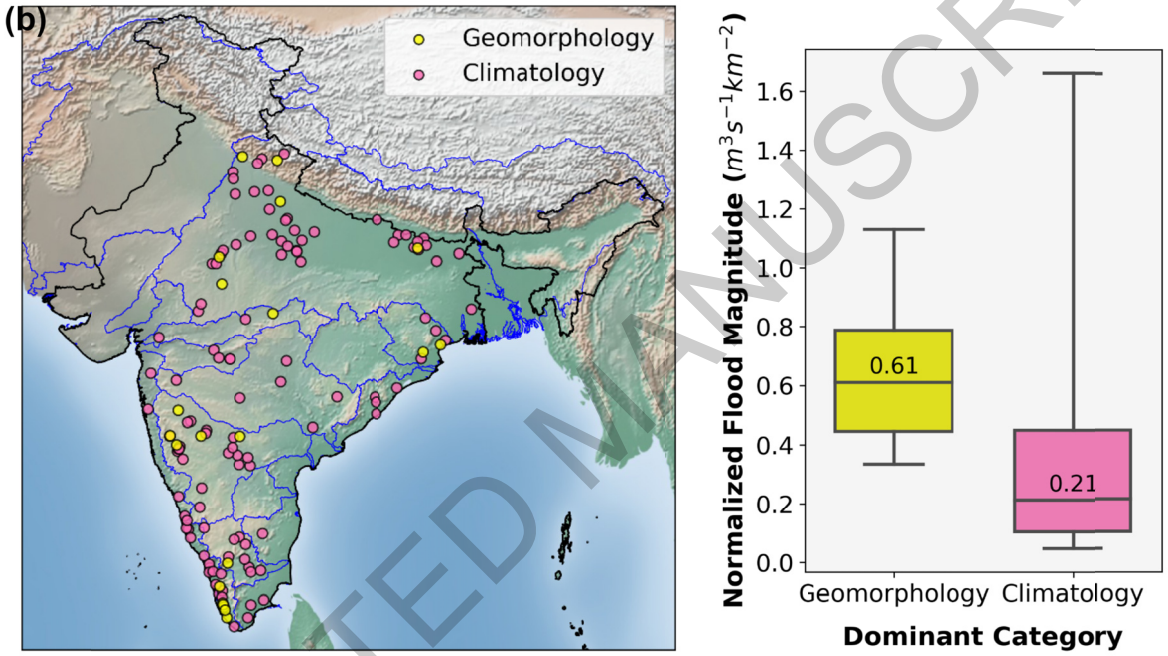


Figure 9

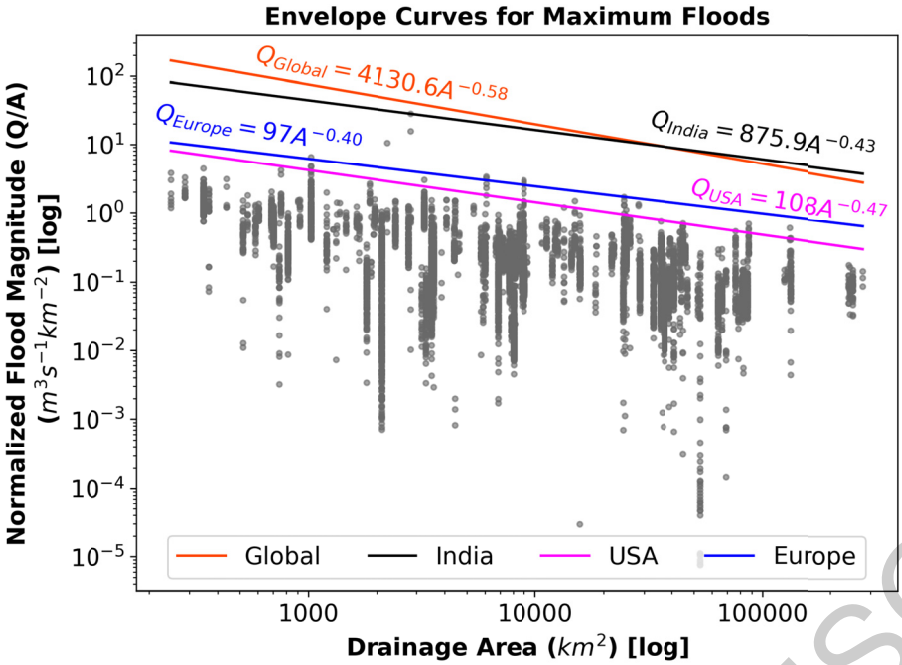


Figure 10

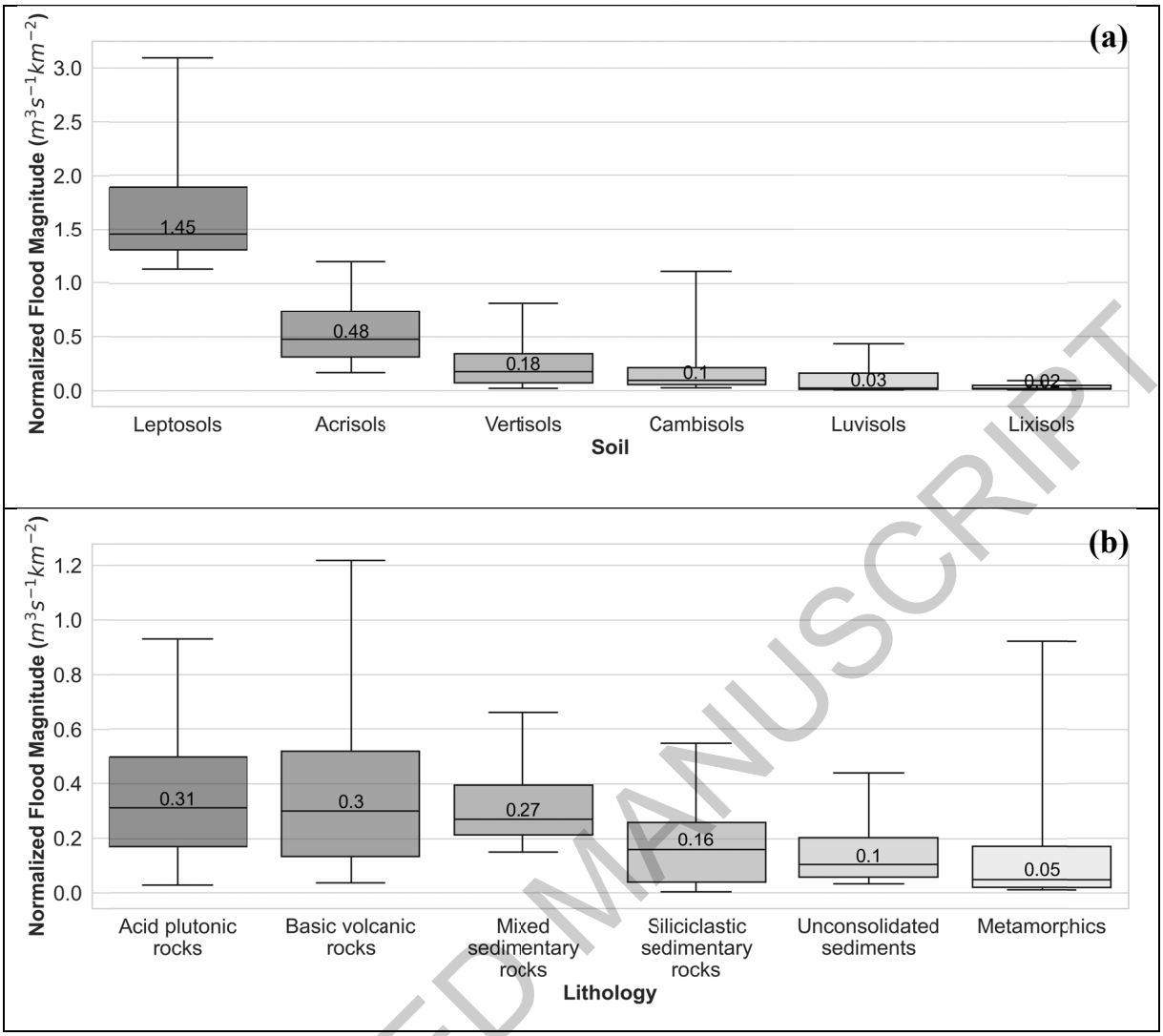


Figure 11

Supporting Information for

Large-sample Characterization of Flooding Events in India

Sai Kiran Kuntla[1] and Manabendra Saharia[1]

[1]Department of Civil Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi 110016, India

Contents of this file

Table S1: List of all the catchment characteristics considered in this study.

Table S1: List of all the catchment characteristics considered in this study.

Catchment Characteristics	Description	Reference
Geomorphology		
Stream order (SO)	Strahler stream order, numerical measure of river's branching complexity	(Strahler, 1952)
Drainage area (DA)	The surface area of the catchment	
Catchment magnitude (CM)	The number of first order streams	(Melton, 1957)
First order streams length (1SL)	The total length of first order streams	(Horton, 1945)
Maximal flow length (MFL)	the length along the longest watercourse from the mouth to the head of the channel	(Mueller, 1968)
Down valley length (DVL)	The straight distance from the river cell of interest to the basin mouth	(Mueller, 1968)
Catchment relief (CR)	The elevation difference between the highest point on the drainage divide and the mouth	(Costa, 1987)
Catchment length (CL)	The maximal length of the line from a basin mouth to a point on the perimeter equidistant from the basin mouth in either direction around the perimeter	(Gregory & Walling, 1968)
Catchment perimeter (CP)	The outer boundary of the watershed that enclosed its area	(Schumm, 1956)
First order streams mean length (1SML)	Mean length of first order streams. 1SML = 1SL/CM	(Strahler, 1964)
Sinuosity index (SI)	SI = MFL/DVL	(Wolman & Miller, 1960)
Form factor (FF)	$FF = DA/CL^2$	(Horton, 1945)
Relief ratio (RR)	RR = BR/CL	(Schumm, 1956)
Elongation ratio (ER)	$ER = 2/(CL \times (DA/\pi)^{0.5})$	(Schumm, 1956)
Circularity ratio (CR)	$CR = 4\ \pi DA/CP^2$	(Miller & Summerson, 1960)
Lemniscate value (LV)	$LV = CL^2/DA$	(Chorley, 1957)
Drainage texture (DT)	DT = Total number of streams of all order/CP	(Horton, 1945)
Drainage density (DD)	DD = Total length of streams of all orders/DA	(Horton, 1945)
Compactness coefficient (CC)	$CC = 0.2841\ (CP/DA^{0.5})$	(Gravelius, 1914)
Wandering ratio (WR)	MFL/CL	(Smart & Surkan, 1967)
Fitness ratio (FR)	MFL/CP	(Melton, 1957)
Channel frequency (CF)	CF = Total number of streams of all	(Horton, 1932)

	orders/DA	
Drainage intensity (DI)	CF/DD	(Faniran, 1968)
Infiltration number (IN)	CF x DD	(Faniran, 1968)
Ruggedness number (RN)	BR x DD	(Strahler, 1964)
Climatological		
Annual mean temperature (Bio1)	The annual mean temperature	(Fick & Hijmans, 2017)
Mean diurnal range (Bio2)	The mean of the monthly temperature ranges (monthly maximum minus monthly minimum)	
Isothermality (Bio3)	Isothermality quantifies how large the day-to-night temperatures oscillate relative to the summer-to-winter (annual) oscillations. Bio3= (Bio2/Bio7)x100	
Temperature seasonality (Bio4)	The amount of temperature variation over a given year (or averaged years) based on the standard deviation (variation) of monthly temperature averages.	
Maximum temperature of warmest month (Bio5)	The maximum monthly temperature occurrence over a given year (time-series) or averaged span of years (normal).	
Minimum temperature of coldest month (Bio6)	The minimum monthly temperature occurrence over a given year (time-series) or averaged span of years (normal).	
Temperature annual range (Bio7)	A measure of temperature variation over a given period. Bio7=Bio5-Bio6	
Mean temperature of wettest Quarter (Bio8)	This quarterly index approximates mean temperatures that prevail during the wettest season.	
Mean temperature of driest quarter (Bio9)	This quarterly index approximates mean temperatures that prevail during the driest quarter.	
Mean temperature of warmest quarter (Bio10)	This quarterly index approximates mean temperatures that prevail during the warmest quarter.	
Mean temperature of coldest quarter (Bio11)	This quarterly index approximates mean temperatures that prevail during the coldest quarter.	
Annual precipitation (Bio12)	This is the sum of all total monthly precipitation values.	
Precipitation of wettest month (Bio13)	This index identifies the total precipitation that prevails during the	

	wettest month.	
Precipitation of driest month (Bio14)	This index identifies the total precipitation that prevails during the driest month.	
Precipitation seasonality (Bio15)	This is a measure of the variation in monthly precipitation totals over the course of the year. This index is the ratio of the standard deviation of the monthly total precipitation to the mean monthly total precipitation (also known as the coefficient of variation) and is expressed as a percentage.	
Precipitation of wettest quarter (Bio16)	This quarterly index approximates total precipitation that prevails during the wettest quarter.	
Precipitation of driest quarter (Bio17)	This quarterly index approximates total precipitation that prevails during the driest quarter.	
Precipitation of warmest quarter (Bio18)	This quarterly index approximates total precipitation that prevails during the warmest quarter.	
Precipitation of coldest quarter (Bio19)	This quarterly index approximates total precipitation that prevails during the coldest quarter.	
Other catchment characteristics		
Lithology	catchment lithology if one single lithology type is present over more than 50% of the catchment area. Otherwise, 'No dominant class'.	(Hartmann & Moosdorf, 2012)
Soil	catchment soil class (WRB) if one single soil class is present over more than 50% of the catchment area. Otherwise, 'No dominant class'.	(Hengl et al., 2017)

References

Chorley, R. J. (1957). Illustrating the Laws of Morphometry. *Geological Magazine*, *94*(2), 140–150. Scopus. https://doi.org/10.1017/S0016756800068412

Costa, J. E. (1987). Hydraulics and basin morphometry of the largest flash floods in the conterminous United States. *Journal of Hydrology*, *93*(3), 313–338. https://doi.org/10.1016/0022-1694(87)90102-8

Faniran, A. (1968). The index of drainage intensity—A provisional new drainage factor. *Aust J Sci*, *31*(9), 328–330. Scopus.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: New 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Gravelius, H. (1914). *Flusskunde*. Goschen'sche Verlagshandlung.

Gregory, K. J., & Walling, D. E. (1968). The Variation of Drainage Density Within a Catchment. *International Association of Scientific Hydrology. Bulletin*, *13*(2), 61–68. https://doi.org/10.1080/02626666809493583

Hartmann, J., & Moosdorf, N. (2012). Global Lithological Map Database v1.0 (gridded to 0.5° spatial resolution). In *Supplement to: Hartmann, Jens; Moosdorf, Nils (2012): The new global lithological map database GLiM: A representation of rock properties at the Earth surface. Geochemistry, Geophysics, Geosystems, 13, Q12004, https://doi.org/10.1029/2012GC004370*. PANGAEA. https://doi.org/10.1594/PANGAEA.788537

Hengl, T., Jesus, J. M. de, Heuvelink, G. B. M., Gonzalez, M. R., Kilibarda, M., Blagotić, A., Shangguan, W., Wright, M. N., Geng, X., Bauer-Marschallinger, B., Guevara, M. A., Vargas, R., MacMillan, R. A., Batjes, N. H., Leenaars, J. G. B., Ribeiro, E., Wheeler, I., Mantel, S., & Kempen, B. (2017). SoilGrids250m: Global gridded soil information based on machine learning. *PLOS ONE*, *12*(2), e0169748. https://doi.org/10.1371/journal.pone.0169748

Horton, R. E. (1932). Drainage-basin characteristics. *Transactions, American Geophysical Union*, *13*, 350–361. https://doi.org/10.1029/TR013i001p00350

Horton, R. E. (1945). Erosional development of streams and their drainage basins; Hydrophysical approach to quantitative morphology. *GSA Bulletin*, *56*(3), 275–370. https://doi.org/10.1130/0016-7606(1945)56[275:EDOSAT]2.0.CO;2

Melton, M. A. (1957). *An analysis of the relations among elements of climate, surface properties, and geomorphology*. New York : Department of Geology, Columbia University. http://archive.org/details/analysisofrelati00melt

Miller, O. M., & Summerson, C. H. (1960). Slope-Zone Maps. *Geographical Review*, *50*(2), 194–202. https://doi.org/10.2307/211507

Mueller, J. E. (1968). An Introduction to the Hydraulic and Topographic Sinuosity Indexes. *Annals of the Association of American Geographers*, *58*(2), 371–385. https://doi.org/10.1111/j.1467-8306.1968.tb00650.x

Schumm, S. A. (1956). Evolution of drainage systems and slopes in badlands at Perth Amboy, New Jersey. *GSA Bulletin*, *67*(5), 597–646. https://doi.org/10.1130/0016-7606(1956)67[597:EODSAS]2.0.CO;2

Smart, J. S., & Surkan, A. J. (1967). The relation between mainstream length and area in drainage basins. *Water Resources Research*, *3*(4), 963–974. Scopus. https://doi.org/10.1029/WR003i004p00963

Strahler, A. N. (1952). Dynamic Basis of Geomorphology. *GSA Bulletin*, *63*(9), 923–938. https://doi.org/10.1130/0016-7606(1952)63[923:DBOG]2.0.CO;2

Strahler, A. N. (1964). *Quantitative Geomorphology of Drainage Basins and Channel Networks*. McGraw Hill.

Wolman, M. G., & Miller, J. P. (1960). Magnitude and Frequency of Forces in Geomorphic Processes. *The Journal of Geology*, *68*(1), 54–74. https://doi.org/10.1086/626637