

EMBRACING LARGE-SAMPLE HYDROLOGY AND MACHINE LEARNING FOR PREDICTION OF FLOODS

KUNTLA SAI KIRAN



**DEPARTMENT OF CIVIL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY DELHI**

AUGUST 2024

© Indian Institute of Technology Delhi (IITD), New Delhi, 2024

EMBRACING LARGE-SAMPLE HYDROLOGY AND MACHINE LEARNING FOR PREDICTION OF FLOODS

by

KUNTLA SAI KIRAN

Department of Civil Engineering

submitted

In fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

to the



INDIAN INSTITUTE OF TECHNOLOGY DELHI

AUGUST 2024

This dissertation is dedicated to my family.

To my parents, Venkatesham Kuntla and Suvarna Kuntla, for their unconditional love, sacrifices,
and unwavering belief in me. I will forever be in their debt.

And to my sister, Supriya Kuntla, for her love and affection.

CERTIFICATE

This is to certify that the dissertation entitled “**EMBRACING LARGE-SAMPLE HYDROLOGY AND MACHINE LEARNING FOR PREDICTION OF FLOODS**”, being submitted by **Mr. Kuntla Sai Kiran**, to the **Indian Institute of Technology Delhi** for the award of the degree of ‘**Doctor of Philosophy**’ in the Department of Civil Engineering is a record bonafide research work carried out by him under my supervision and guidance. The dissertation work, in my opinion, has reached the requisite standards fulfilling the requirement for the Degree of **Doctor of Philosophy**.

The material contained in the dissertation is original and has not been submitted in part or full to any other University or Institute for the award of any other degree or diploma.

Prof. Manabendra Saharia

Department of Civil Engineering

Indian Institute of Technology Delhi

New Delhi-110016, India

ACKNOWLEDGEMENTS

Completing this Ph.D. dissertation has been a journey filled with challenges and triumphs, and it would not have been possible without the support, guidance, and encouragement of many individuals. I am deeply grateful to all those who have helped me along the way.

First and foremost, I would like to express my sincere gratitude to my advisor, Dr. Manabendra Saharia, for his unwavering support, invaluable guidance, and insightful feedback throughout my doctoral journey. His patience, expertise, and encouragement have been instrumental in shaping this dissertation and my development as a researcher.

I would also like to extend my heartfelt thanks to my research committee members, Prof. Vasant Matsagar, Prof. B. R. Chahar, and Dr. Sandeep Sukumaran, for their time, valuable suggestions, and steadfast support.

Special thanks go to all my colleagues from the Water Resources Engineering section for creating a stimulating and supportive environment. I am especially indebted to my colleagues from HydroSense Lab for their friendship, collaboration, and encouragement during this journey.

My gratitude extends to all my friends from the Girnar hostel of IIT Delhi, who have always been beside me during this journey. I will cherish the beautiful moments I spent with you guys throughout my life.

I would also like to thank Dr. Archana M. Nair (IIT Guwahati) and Dr. Panchagnula Manjusree (ISRO, Hyderabad) for their invaluable guidance and support, and for introducing me to the world of research.

Acknowledging everyone who has contributed to my learning journey over the years is a daunting task. By any chance, if you come across this acknowledgment, please understand that I am deeply appreciative of the opportunities, mentorship, and friendships you have provided. Your contributions have been instrumental in making this achievement possible.



(Kuntla Sai Kiran)

ABSTRACT

Floods are causing increasing harm worldwide, with regional differences, and India is expected to experience higher-than-average losses in life, infrastructure, and the economy. A comprehensive understanding of flooding behavior based on observational data is vital for various aspects of flood management. Yet, catchment characterization studies are generally limited to case studies or smaller regions. The heterogeneity of catchment characteristics, analysis periods, chosen hydrological indices, and methods lead to fragmented and sometimes contrasting scientific findings that can't be applied at large scales. On the other hand, with ever-increasing geospatial datasets, increased accessibility to streamflow observations, and advancements in computational resources, there is an opportunity for large-sample hydrological studies for a comprehensive characterization of floods globally and nationally. This dissertation addresses three key science questions embracing large-sample hydrology: (1) How do catchment characteristics influence high streamflow extremes across the globe?; (2) Is there any change in the nature of flooding in India over the last few decades?; (3) How do catchment characteristics influence the floods in India, and to what extent can these relationships help extrapolate the hydrological signatures over ungauged catchments?

The first empirical investigation in this research focuses on characterizing streamflow extremes across 9,700 catchments worldwide, using multiple physiographic explanatory variables derived from various global archives. Along with the spatial and temporal variations of high streamflow extremes, their correlation with catchment characteristics, such as geomorphology, meteorology, climatology, landcover, lithology, etc., were examined to identify the first-order relationships and relatively most dominant factors in varying climate classes. The relationships derived help us to understand the hydrological processes in various climates across the globe,

which play a crucial role in flood modeling and assessing the nature of extremes, consequently implicating risk management methods. At the same time, using a globally representative observational streamflow dataset and Gini coefficients, this study investigates how streamflow inequality varies spatially and temporally and its relationship with different underlying catchment characteristics. It is found that higher dam storage capacity in the catchment corresponds to more evenly distributed streamflow availability throughout the year by reducing high stream flows. Thus, by identifying the underlying socioeconomic factors leading to disparities in streamflow inequalities across the globe, international organizations and local governments can identify locations that need immediate attention.

Besides, one of the primary reasons hindering large-sample hydrological studies in developing countries such as India is the limited hydrometric records and the unavailability of a holistic database describing the physical characteristics of past flood events, such as flood level, discharge, duration, etc. Hence, this dissertation addresses this critical issue by developing a comprehensive event-based flood database called INDOFLOODS. This is the first-of-its-kind flood events database for India, with a total of 8342 flood events observed between 1959 and 2020. The database is augmented with many geomorphological, climatological, event-scale precipitation, land cover, soil, lithology, and anthropogenic characteristics derived at the catchment scale to cater to various applications advancing flood research and management in the country. Subsequently, this database has been used to characterize normalized flood magnitudes over India comprehensively. Envelope curves reveal that the maximum magnitude of floods in India is much higher than those reported for the US and Europe. The investigation into multidimensional relationships using random forests in conjunction with the SHAP model gave insights into the combined and individual importance of catchment characteristics. The grouped

feature importance suggests that climatology contributes more (66.7%) to the model than geomorphology (33.3%). Thus, the insights of this study play an important role in improved understanding of flood mechanisms, large-scale modeling, and prediction at ungauged locations in the country.

On the other hand, using the most representative streamflow records in India to date, we identified a decreasing trend in flood magnitude across 76% of the catchments and regional variation in the change of the timing of floods. Besides, we investigated if any continental-scale climate signals lead to these trends in the country. In addition, our investigation delves into the potential correlations between alterations in flooding and catchment characteristics. Our findings reveal that there is a dampening of flood discharges within basins situated in arid climates and a substantial influence of catchment size on changes in floods. Understanding these dynamics in magnitude and timing and associated reasons leading to these changes is essential as they have an impact on disaster management and water resources planning. Furthermore, with the aim of finding the flooding characteristics at ungauged locations, two important hydrological signatures, the mean annual peak discharge and flood severity, are extrapolated to ungauged locations from the known values at gauged locations as a function of multiple catchment characteristics. Flood events from INDOFLOODS are used to derive the flood severity at gauged locations based on a newly proposed formula. Subsequently, a quantile random forest machine learning model has been used to model and predict flooding variables and their uncertainties across the country.

By embracing data-driven science, this research provides extensive insights into flooding behavior and the influence of catchment characteristics both globally and nationally. The public release of all developed datasets enhances the potential for further knowledge accumulation and advances flood research and management in the country.

सार

बाढ़ दुनिया भर में क्षेत्रीय अंतरों के साथ बढ़ती हुई क्षति का कारण बन रही है, और भारत में जीवन, बुनियादी ढांचे और अर्थव्यवस्था में औसत से अधिक नुकसान होने की उम्मीद है। बाढ़ प्रबंधन के विभिन्न पहलुओं के लिए अवलोकन संबंधी आंकड़ों के आधार पर बाढ़ व्यवहार की व्यापक समझ महत्वपूर्ण है। फिर भी, जलग्रहण विशेषता अध्ययन आम तौर पर केस स्टडीज या छोटे क्षेत्रों तक ही सीमित होते हैं। जलग्रहण विशेषताओं, विश्लेषण अवधि, चुने हुए जल विज्ञान सूचकांक और विधियों की विविधता के कारण खंडित और कभी-कभी विपरीत वैज्ञानिक निष्कर्ष निकलते हैं जिन्हें बड़े पैमाने पर लागू नहीं किया जा सकता है। दूसरी ओर, लगातार बढ़ते भू-स्थानिक डेटासेट, स्ट्रीमफ्लो अवलोकनों तक बढ़ती पहुंच और कम्प्यूटेशनल संसाधनों में प्रगति के साथ, वैश्विक और राष्ट्रीय स्तर पर बाढ़ के व्यापक लक्षण वर्णन के लिए बड़े-नमूने वाले जल विज्ञान अध्ययनों का अवसर है। (3) जलग्रहण क्षेत्र की विशेषताएँ भारत में बाढ़ को किस प्रकार प्रभावित करती हैं, तथा ये संबंध किस हद तक गैर-मापन किए गए जलग्रहण क्षेत्रों पर जल विज्ञान संबंधी विशेषताओं को जानने में मदद कर सकते हैं?

इस शोध में पहली अनुभवजन्य जांच दुनिया भर के 9,700 जलग्रहण क्षेत्रों में धारा प्रवाह की चरम सीमाओं को चिह्नित करने पर केंद्रित है, जिसमें विभिन्न वैश्विक अभिलेखों से प्राप्त कई भौतिक व्याख्यात्मक चर का उपयोग किया गया है। उच्च धारा प्रवाह चरम सीमाओं के स्थानिक और लौकिक विविधताओं के साथ-साथ, जलग्रहण विशेषताओं, जैसे कि भू-आकृति विज्ञान, मौसम विज्ञान, जलवायु विज्ञान, भू-आवरण, लिथोलॉजी, आदि के साथ उनके सहसंबंध की जांच की गई, ताकि विभिन्न जलवायु वर्गों में प्रथम-क्रम संबंधों और अपेक्षाकृत सबसे प्रमुख कारकों की पहचान की जा सके। प्राप्त संबंध हमें दुनिया भर में विभिन्न जलवायु में हाइड्रोलॉजिकल प्रक्रियाओं को समझने में मदद करते हैं, जो बाढ़ मॉडलिंग और चरम सीमाओं की प्रकृति का आकलन करने में महत्वपूर्ण भूमिका निभाते हैं, जिसके परिणामस्वरूप जोखिम प्रबंधन विधियों को

शामिल किया जाता है। साथ ही, वैश्विक रूप से प्रतिनिधि अवलोकन धारा प्रवाह डेटासेट और गिनी गुणांक का उपयोग करते हुए, यह अध्ययन जांच करता है कि कैसे धारा प्रवाह असमानता स्थानिक और लौकिक रूप से भिन्न होती है और विभिन्न अंतर्निहित जलग्रहण विशेषताओं के साथ इसका संबंध है। यह पाया गया है कि जलग्रहण क्षेत्र में उच्च बांध भंडारण क्षमता उच्च धारा प्रवाह को कम करके पूरे वर्ष अधिक समान रूप से वितरित धारा प्रवाह उपलब्धता के अनुरूप है। इस प्रकार, विश्व भर में जल प्रवाह असमानताओं के लिए जिम्मेदार अंतर्निहित सामाजिक-आर्थिक कारकों की पहचान करके, अंतर्राष्ट्रीय संगठन और स्थानीय सरकारें उन स्थानों की पहचान कर सकती हैं जिन पर तत्काल ध्यान देने की आवश्यकता है।

इसके अलावा, भारत जैसे विकासशील देशों में बड़े-नमूने वाले हाइड्रोलॉजिकल अध्ययनों में बाधा डालने वाले प्राथमिक कारणों में से एक सीमित हाइड्रोमेट्रिक रिकॉर्ड और पिछले बाढ़ की घटनाओं की भौतिक विशेषताओं, जैसे बाढ़ का स्तर, निर्वहन, अवधि आदि का वर्णन करने वाले समग्र डेटाबेस की अनुपलब्धता है। इसलिए, यह शोध प्रबंध INDOFLOODS नामक एक व्यापक घटना-आधारित बाढ़ डेटाबेस विकसित करके इस महत्वपूर्ण मुद्दे को संबोधित करता है। यह भारत के लिए अपनी तरह का पहला बाढ़ घटना डेटाबेस है, जिसमें 1959 और 2020 के बीच कुल 8342 बाढ़ की घटनाएँ देखी गई हैं। देश में बाढ़ अनुसंधान और प्रबंधन को आगे बढ़ाने वाले विभिन्न अनुप्रयोगों को पूरा करने के लिए जलग्रहण पैमाने पर प्राप्त कई भू-आकृति विज्ञान, जलवायु विज्ञान, घटना-पैमाने पर वर्षा, भूमि आवरण, मिट्टी, लिथोलॉजी और मानवजनित विशेषताओं के साथ डेटाबेस को संवर्धित किया गया है। इसके बाद, इस डेटाबेस का उपयोग भारत में सामान्यीकृत बाढ़ परिमाणों को व्यापक रूप से चिह्नित करने के लिए किया गया है। लिफाफा वक्रों से पता चलता है कि भारत में बाढ़ की अधिकतम परिमाण अमेरिका और यूरोप के लिए रिपोर्ट की गई तुलना में बहुत अधिक है। SHAP मॉडल के साथ यादृच्छिक वनों का उपयोग करके बहुआयामी संबंधों की जांच ने जलग्रहण विशेषताओं के संयुक्त और व्यक्तिगत महत्व के बारे में जानकारी दी। समूहीकृत विशेषता महत्व

से पता चलता है कि भूआकृति विज्ञान (33.3%) की तुलना में जलवायु विज्ञान मॉडल में अधिक (66.7%) योगदान देता है। इस प्रकार, इस अध्ययन की अंतर्दृष्टि बाढ़ तंत्र, बड़े पैमाने पर मॉडलिंग और देश में बिना मापे गए स्थानों पर भविष्यवाणी की बेहतर समझ में महत्वपूर्ण भूमिका निभाती है।

दूसरी ओर, भारत में आज तक के सबसे प्रतिनिधि स्ट्रीमफ्लो रिकॉर्ड का उपयोग करते हुए, हमने 76% जलग्रहण क्षेत्रों में बाढ़ की तीव्रता में कमी की प्रवृत्ति और बाढ़ के समय में परिवर्तन में क्षेत्रीय भिन्नता की पहचान की। इसके अलावा, हमने जांच की कि क्या कोई महाद्वीपीय-स्तरीय जलवायु संकेत देश में इन प्रवृत्तियों को जन्म देते हैं। इसके अलावा, हमारी जांच बाढ़ और जलग्रहण विशेषताओं में परिवर्तन के बीच संभावित सहसंबंधों की गहराई से जांच करती है। हमारे निष्कर्षों से पता चलता है कि शुष्क जलवायु में स्थित बेसिनों के भीतर बाढ़ के निर्वहन में कमी आई है और बाढ़ में परिवर्तन पर जलग्रहण आकार का पर्याप्त प्रभाव पड़ा है। परिमाण और समय में इन गतिशीलताओं और इन परिवर्तनों के लिए संबंधित कारणों को समझना आवश्यक है क्योंकि वे आपदा प्रबंधन और जल संसाधन नियोजन को दृढ़ता से प्रभावित कर सकते हैं। इसके अलावा, बिना मापे गए स्थानों पर बाढ़ की विशेषताओं का पता लगाने के उद्देश्य से, दो महत्वपूर्ण हाइड्रोलॉजिकल हस्ताक्षर, औसत वार्षिक शिखर निर्वहन और बाढ़ की गंभीरता, कई जलग्रहण विशेषताओं के एक फ़ंक्शन के रूप में मापे गए स्थानों पर ज्ञात मूल्यों से बिना मापे गए स्थानों पर एक्सट्रपलेशन किए जाते हैं। इंडोफ़्लूड्स से बाढ़ की घटनाओं का उपयोग नए प्रस्तावित सूत्र के आधार पर मापे गए स्थानों पर बाढ़ की गंभीरता को निकालने के लिए किया जाता है। इसके बाद, देश भर में बाढ़ के कारकों और उनकी अनिश्चितताओं का मॉडल तैयार करने और भविष्यवाणी करने के लिए क्वांटाइल रैंडम फॉरेस्ट मशीन लर्निंग मॉडल का उपयोग किया गया है।

डेटा-संचालित विज्ञान को अपनाकर, यह शोध वैश्विक और राष्ट्रीय स्तर पर बाढ़ के व्यवहार और जलग्रहण विशेषताओं के प्रभाव के बारे में व्यापक जानकारी प्रदान करता है। सभी विकसित डेटासेट को

सार्वजनिक रूप से जारी करने से आगे ज्ञान संचय की संभावना बढ़ती है और देश में बाढ़ अनुसंधान और प्रबंधन में प्रगति होती है।

शोध प्रबंध का शीर्षक: बाढ़ की भविष्यवाणी के लिए बड़े-नमूने वाले जल विज्ञान और मशीन लर्निंग को अपनाना

छात्र का नाम: कुंतला साई किरण

प्रवेश संख्या: 2019CEZ8628

TABLE OF CONTENTS

CERTIFICATE.....	i
ACKNOWLEDGEMENTS	ii
ABSTRACT.....	iv
LIST OF FIGURES	xvii
LIST OF TABLES	xxvi
LIST OF ABBREVIATIONS	xxix
CHAPTER 1	1
INTRODUCTION.....	1
1.1 PROBLEM STATEMENT	2
1.2 SCIENCE QUESTIONS.....	5
1.3 OBJECTIVES	5
1.4 STRUCTURE OF THE DISSERTATION.....	6
CHAPTER 2.....	8
A COMPREHENSIVE CHARACTERIZATION OF HIGH STREAMFLOW EXTREMES ACROSS THE GLOBE.....	8
2.1 INTRODUCTION	8
2.2 DATA USED AND METHODS	10
2.2.1 DATA	10
2.2.2 STUDY AREA	14

2.2.3 METHODS	16
2.3 SPATIAL AND TEMPORAL DISTRIBUTION OF HSE EVENTS	19
2.4 EXPLORING THE FIRST-ORDER RELATIONSHIP BETWEEN HSE EVENTS AND VARIOUS GEOMORPHOLOGICAL CHARACTERISTICS OF CATCHMENTS	25
2.5 EXPLORING THE FIRST-ORDER RELATIONSHIP BETWEEN HSE EVENTS AND VARIOUS CLIMATOLOGICAL CHARACTERISTICS OF CATCHMENTS	29
2.6 EXPLORING THE RELATIONSHIP BETWEEN HSE EVENTS AND CATCHMENT CHARACTERISTICS	31
2.7 RELATIVE IMPORTANCE OF CATCHMENT CHARACTERISTICS	36
2.8 CONCLUSIONS.....	42
CHAPTER 3.....	46
DAMS MODERATE STREAMFLOW INEQUALITY, WHILE PRECIPITATION INEQUALITY EXACERBATES IT	46
3.1 INTRODUCTION	46
3.2 DATA USED AND METHODS	48
3.2.1 DATA	48
3.2.2 STUDY AREA	50
3.2.3 METHODS	51
3.3 RESULTS	53
3.3.1 SPATIAL DISTRIBUTION AND THE TRENDS IN GINI COEFFICIENTS OF STREAMFLOW	53

3.3.2 EXPLORING THE RELATIONSHIP BETWEEN STREAMFLOW INEQUALITY AND CATCHMENT CHARACTERISTICS	57
3.4 CONCLUSIONS.....	60
CHAPTER 4.....	64
A COMPREHENSIVE DATABASE FOR FLOOD EVENTS IN INDIA ENHANCED WITH HYDROMETEOROLOGICAL AND CATCHMENT ATTRIBUTES.....	64
4.1 INTRODUCTION	64
4.2 DATA SOURCES	66
4.3 DATA STRUCTURE AND PROCESSING	68
4.3.1 METADATA	68
4.3.2 FLOOD-EVENTS.....	70
4.3.3 CATCHMENT SHAPEFILES	72
4.3.4 CATCHMENT CHARACTERISTICS	74
4.4 PRILIMINARY OBSERVATIONS.....	80
4.5 SUMMARY	85
CHAPTER 5.....	87
CHANGING MAGNITUDE AND TIMING OF RIVERINE FLOODS IN INDIA	87
5.1 INTRODUCTION	87
5.2 DATA USED AND METHODS	90
5.2.1 DATA	90

5.2.2 METHODS	93
5.3 RESULTS AND DISCUSSIONS	95
5.3.1 REGIONAL PATTERNS OF CHANGE IN THE MAGNITUDE AND TIMING OF FLOODS	95
5.3.2 EVOLUTION OF FLOODS AND THEIR DRIVERS	99
5.3.3 CHANGES IN 100-YEAR FLOOD DISCHARGE	103
5.3.4 RELATIONSHIP BETWEEN FLOODS AND CATCHMENT CHARACTERISTICS	104
5.4 SUMMARY	106
CHAPTER 6.....	108
LARGE-SAMPLE CHARACTERIZATION OF FLOODING EVENTS IN INDIA.....	108
6.1 INTRODUCTION	108
6.2 DATA USED AND METHODS	110
6.2.1 DATA	110
6.2.2 METHODS	112
6.3 SPATIAL AND TEMPORAL DISTRIBUTION OF FLOODS	115
6.4 THE RELATIONSHIP BETWEEN FLOOD MAGNITUDE AND GEOMORPHOLOGY	120
6.5 THE RELATIONSHIP BETWEEN FLOOD MAGNITUDE AND CLIMATOLOGY .	125
6.6 THE RELATIONSHIP BETWEEN FLOOD MAGNITUDE AND SOIL AND LITHOLOGY	126

6.7 RELATIVE IMPORTANCE OF GEOMORPHOLOGY AND CLIMATOLOGY	128
6.8 CONCLUSIONS.....	131
CHAPTER 7.....	133
MAPPING THE MEAN ANNUAL PEAK DISCHARGE AND FLOOD SEVERITY OVER UNGAUGED LOCATIONS ACROSS INDIA	133
7.1 INTRODUCTION	133
7.2 DATA USED AND METHODS	135
7.2.1 DATA	136
7.2.2 MODELING AND EVALUATION.....	139
7.3 ASSOCIATION OF PREDICTANDS WITH BASIN GEOMORPHOLOGY AND CLIMATOLOGY	141
7.3.1 MEAN ANNUAL PEAK DISCHARGE.....	142
7.3.2 FLOOD SEVERITY	145
7.4 SPATIAL DISTRIBUTION OF MEAN ANNUAL PEAK DISCHARGE AND FLOOD SEVERITY	150
7.4.1 MEAN ANNUAL PEAK DISCHARGE.....	150
7.4.2 FLOOD SEVERITY	153
7.5 CONCLUSIONS.....	158
CHAPTER 8.....	160
SUMMARY AND FUTURE WORK.....	160

8.1 SUMMARY	160
8.2 LIMITATIONS AND FUTURE WORK	162
REFERENCES.....	165
APPENDIX.....	202
JOURNAL PUBLICATIONS.....	206
CURRICULUM VITAE.....	207

LIST OF FIGURES

- Figure 2. 1: The geographic location of all the gauge stations and their associated climatic classes used in this study. The yellow box projects a cluster of gauge stations whose catchments are featured with Polar climate, overlayed on a 30-year (1970-2000) average of August month precipitation. 15
- Figure 2. 2: Spatial distribution of unit peak discharge for the highest streamflow events recorded over the corresponding gauge stations across the globe. 22
- Figure 2. 3: Box-and-Whisker plot of unit peak discharge for different climatic classes. The box spans the interquartile range in the plot, i.e., the upper and lower ends of the box correspond to the first quartile and third quartiles, respectively. The whiskers are the two vertical lines outside the box extended until the observations' extremes. The median is marked by a horizontal line inside the box, and the mean is plotted with a white circle. The black circles beyond the extremes are considered outliers at a distance of more than 1.5 times the interquartile range. The red cross marks correspond to the top ten catchments in terms of high unit peak discharge values. . 23
- Figure 2. 4: Monthly distribution of high streamflow extreme (HSE) events normalized by the total number of events in different climate classes. 25
- Figure 2. 5: Spearman-correlation coefficients between unit peak discharge of all the high streamflow events and multiple catchment characteristics. The values range from 1 to -1. The positive values represent a positive correlation, with 1 being the highest/perfect positive correlation, while the negative values represent a negative

correlation, with -1 being the highest/perfect negative correlation. A value equal to 0 represents no correlation.	27
Figure 2. 6: Unit peak discharge vs. drainage area along with their envelope curves for all the high streamflow extreme (HSE) events across the globe. The envelope curves for continental United States and Europe are taken from Marchi et al. (2010) and Saharia et al. (2017).	29
Figure 2. 7: Box-and-Whisker plot of (a) annual precipitation and (b) precipitation of wettest month for different climatic classes. The red cross marks correspond to the top ten catchments in terms of high unit peak discharge values. <i>Please refer to the caption of Figure 2.3 for the description of the box plots.</i>	31
Figure 2. 8: SHAP summary plots for different climatic classes. The order of the features along the y-axis are based on their importance in descending order. The value on the right side of the features name along the y-axis denotes mean absolute SHAP values for the corresponding features. The horizontal location of the dots depicts the intensity of its influence on predictions. Zero values on the x-axis indicate no impact, and positive (negative) values towards the rights (left) side indicate positive (negative) impact. The color of the dots represents the values of the corresponding features if it is high (red) or low (blue) for that observation in the dataset.	41
Figure 3. 1: The spatial location of all the gauge stations and their associated climatic classes used in this study.	50

Figure 3. 2: Spatial distribution of the mean Gini coefficient for streamflow over the corresponding stream gauge stations between 1990 and 2013. A value of 0 indicates a uniform distribution of flows throughout the year, whereas a value of 1 indicates that all the flows occur on a single day. The boxplots in the inset plot summarize the mean streamflow Gini over different climates. The box spans the interquartile range (i.e., 25th and 75th percentiles), with the bar in the middle representing the median. The whiskers are the two vertical lines outside the box extended until the 5th and 95th percentiles. 54

Figure 3. 3: Spatial distribution of the mean Gini coefficient for precipitation over the catchments of corresponding stream gauge stations between 1990 and 2013. *Please refer to the caption of figure 3.2 for the description of Gini coefficient values and box plots....* 55

Figure 3. 4: Trends in streamflow Gini coefficient (streamflow inequality) between 1990 and 2013 based on the Modified Mann-Kendall trend test at a 10% significance level. The inset text boxes display the percentage of stations with statistically significant decreasing and increasing trends in each continent. 56

Figure 3. 5: Trends in precipitation Gini coefficient between 1990 and 2013 based on the Modified Mann Kendall trend test at 10% significance level. The inset text boxes display the percentage of stations with statistically significant decreasing and increasing trends in each continent. 57

Figure 3. 6: A straight line fit between streamflow Gini and their corresponding precipitation Gini for different climates. 58

Figure 3. 7: Scatter plot between (a) Streamflow Gini and Population count in the corresponding catchments across the globe. The points are color-coded by the Human Development Index (HDI, i.e., more developed areas have higher HDI scores and vice versa). (b) Streamflow Gini vs. Total storage volume in million cubic meters of all the dams in the respective catchment. (c) Box plot of streamflow Gini pre- and post-dam construction. The box spans the interquartile range (i.e., 25th and 75th percentiles), with the bar in the middle representing the median and its corresponding value as text within the box. The whiskers are the two horizontal lines outside the box extended until the 5th and 95th percentiles. 59

Figure 4. 1: The geographical location of the gauge stations where floods are observed and whose data is available in the INDOFLOODS database. The base layer of the plot showcases the Köppen-Geiger climate (Kottek et al., 2006), major river networks, and basins in the country. The inset figure on the lower right corner illustrates how floods are extracted from the time series of water level data. 67

Figure 4. 2: Dependence of flood events characteristics (a) peak discharge, and (b) duration on geomorphology of the catchments. 81

Figure 4. 3: Box plots showing the distribution of daily precipitation before flood events. The box spans the interquartile range (i.e., 25th and 75th percentiles), with the bar in the middle representing the median. The whiskers are the two vertical lines outside the box extended until the 5th and 95th percentiles..... 83

Figure 4. 4: Peak specific discharge as a function of the catchment area for all the 8342 flood events, along with the envelope curves for extreme floods. The envelope curves for

observed extreme floods over the Global and continental United States are taken from Kuntla et al. (2022) and Saharia et al. (2017a), respectively.	84
Figure 5. 1: Annual and seasonal trends in (a) flood magnitude and (b) timing between 1970 and 2010 based on the Theil-Sen slope estimator.....	99
Figure 5. 2: Long-term temporal evolutions of flood magnitude along with its drivers for hotspots with clear regional patterns. The data represents observed floods (blue), peak 5-day precipitation (orange), peak monthly soil moisture (green), and peak monthly temperature (red). Solid lines show the median, and the shaded region shows the variability within the hotspots (25 th and 75 th percentile). All data were subjected to a 10-year weighted moving average filter.	101
Figure 5. 3: Long-term temporal evolutions of flood timing along with their drivers for hotspots with a clear regional pattern in flood magnitude (Hotspot M1, and M2). The data represents observed floods (blue), peak 5-day precipitation (orange), peak monthly soil moisture (green), and peak monthly temperature (red). Solid lines show the median, and the shaded region shows the variability within the hotspots (25 th and 75 th percentile). All data were subjected to a 10-year weighted moving average filter.	102
Figure 5. 4: Long-term temporal evolutions of flood timing along with its drivers for hotspots with clear regional patterns. <i>Please refer to the caption of Figure 5.2 for the description of the plots.</i>	103

Figure 5. 5: Estimated (a) percentage change in the magnitude of 1970 100-year flood discharge in 2010. and (b) return period in 2010 for the 1970 100-year flood discharge. Grey-colored points represent stationary stations.	104
Figure 5. 6: Percentage of catchments with decreasing and increasing trends in flood magnitude in each climate over different seasons.	105
Figure 5. 7: Drainage area vs. percentage change in flood magnitude per decade based on the Theil-Sen slope estimator over all the catchments. Points are color-coded based on the mean elevation of the catchment. The background is color-coded based on the size of catchments: small (less than 1,000 sq.km.), medium (1,000 – 10,000 sq.km.), and large (more than 10,000 sq.km.)	106
Figure 6. 1: 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.	116
Figure 6. 2: Box-and-Whisker plot of normalized flood magnitude for different basins. The box spans the interquartile range (i.e., 25 th 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.....	118
Figure 6. 3: Monthly distribution of flood events normalized by the total number of flood events in different basins. The numbers in the legend correspond to the total number of flood events observed in respective basins.....	119

- Figure 6. 4: 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..... 122
- Figure 6. 5: Box-and-Whisker plot of (a) Elongation ratio and (b) Maximal flow length over different basins. *Please refer to the caption of Figure 6.2 for the description of the box plots.* 124
- Figure 6. 6: Box-and-Whisker plot of Precipitation of wettest quarter over different basins. *Please refer to the caption of Figure 6.2 for the description of the box plots.* 126
- Figure 6. 7: Box-and-Whisker plot of normalized flood magnitude for different (a) Soil and (b) Lithology types. *Please refer to the caption of Figure 6.2 for the description of the box plots.* 127
- Figure 6. 8: (a) Group feature importance of geomorphological variables and climatological variables on normalized flood magnitude. (b) 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).	130
Figure 7. 1: Methodology for predicting mean annual peak discharge (APD) and flood severity over ungauged locations across the country.	135
Figure 7. 2: SHAP summary plot of the top 15 most influencing explanatory variables across the spectrum. 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) show positive (negative) impact. The color of the dots represents the value of the relevant explanatory variables for that observation in the dataset, whether high (red) or low (blue).	143
Figure 7. 3: A correlation matrix between the mean annual peak discharge and catchment characteristics.....	144
Figure 7. 4: SHAP summary plot of the top 15 most influencing explanatory variables across the spectrum. <i>Please refer to the caption of Figure 7.2 for the description of the SHAP summary plot.</i>	146
Figure 7. 5: correlation matrix between the flood severity and catchment characteristics.....	148

Figure 7. 6: Spatial distribution of (a) observed mean annual peak discharge, (b) predicted annual peak discharge across all the streams in the country, and (c) their uncertainty ranges.

..... 151

Figure 7. 7: Spatial distribution of (a) observed flood severity over gauge stations, (b) predicted flood severity across the country, and (c) their uncertainty ranges. 155

LIST OF TABLES

Table 2. 1: Number of gauge stations in various climate regions.....	15
Table 2. 2: Climate formula of Koppen-Geiger for the main climates (Kottek et al., 2006). Where, <i>Tmax</i> and <i>Tmn</i> are the monthly mean temperatures of the warmest and coldest months, respectively. <i>Pann</i> is the accumulated annual precipitation. <i>Pth</i> is a dryness threshold in mm, which depends on $\{Tann\}$, the absolute measure of the annual mean temperature in °C, and on the annual cycle of precipitation: $Pth = 2\{Tann\}$ if at least 2/3 of the annual precipitation occurs in winter; $Pth = 2\{Tann\}+28$ if at least 2/3 of the annual precipitation occurs in summer; $Pth = 2\{Tann\}+14$ otherwise... 16	16
Table 2. 3: Random Forest model predictions statistics	36
Table 3. 1: Climate formula of Koppen-Geiger for the main climates (Kottek et al., 2006). Where, <i>Tmax</i> and <i>Tmin</i> are the monthly mean temperatures of the warmest and coldest months, respectively. <i>Pann</i> is the accumulated annual precipitation. <i>Pth</i> is a dryness threshold in mm, which depends on $\{Tann\}$, the absolute measure of the annual mean temperature in °C, and on the annual cycle of precipitation: $Pth = 2\{Tann\}$ if at least 2/3 of the annual precipitation occurs in winter; $Pth = 2\{Tann\}+28$ if at least 2/3 of the annual precipitation occurs in summer; $Pth = 2\{Tann\}+14$ otherwise.....	49
Table 3. 2: Number of gauge stations in various climate regions.....	51
Table 4. 1: Information provided in the metadata of the database.....	68
Table 4. 2: Definition of flood-event variables contributing to the database.	70

Table 4. 3: List of catchment-scale variables provided in the INDOFLOODS.....	74
Table 5. 1: Climate formula of Koppen-Geiger for the main climates (Kottek et al., 2006). Where, <i>T_{max}</i> and <i>T_{min}</i> are the monthly mean temperatures of the warmest and coldest months, respectively. <i>P_{ann}</i> is the accumulated annual precipitation. <i>P_{th}</i> is a dryness threshold in mm, which depends on { <i>T_{ann}</i> }, the absolute measure of the annual mean temperature in °C, and on the annual cycle of precipitation: <i>P_{th}</i> = 2{ <i>T_{ann}</i> } if at least 2/3 of the annual precipitation occurs in winter; <i>P_{th}</i> = 2{ <i>T_{ann}</i> }+28 if at least 2/3 of the annual precipitation occurs in summer; <i>P_{th}</i> = 2{ <i>T_{ann}</i> }+14 otherwise.....	92
Table 5. 2: Spearman’s rank correlation coefficient (r) between hotspot medians of the annual series of flood discharge and timing with their drivers.....	101
Table 5. 3: Spearman’s rank correlation coefficient (r) between the median of flood timing and their drivers over hotspots of catchments showing a significant trend in flood magnitude.....	102
Table 6. 1: Number of gauge stations in various basins.	116
Table 7. 1: Catchment characteristics used for training the model and predicting the mean annual peak discharge across the country.....	144
Table 7. 2: Catchment characteristics used for training the model and predicting the flood severity across the country.	148
Table 7. 3: Cross validation (k=5) model evaluation based on the observed and predicted values over the test data.	152

Table 7. 4: Cross validation (k=5) model evaluation based on the observed and predicted values over the test data.	156
Table A 1: List of the catchment characteristics.....	202

LIST OF ABBREVIATIONS

APD: Annual Peak Discharges

AIC: Akaike Information Criterion

ANOVA: An analysis of variance

CCI-LC: Climate Change Initiative Land Cover

CPC: Climate Prediction Center

CRU: Climatic Research Unit

CWC: Central Water Commission

DEM: Digital Elevation Model

DFO: Dartmouth Flood Observatory

DMI: Dipole Mode Index

ECMWF: European Centre for Medium-Range Weather Forecasts

EM-DAT: Emergency Disasters Database

ERA5: ECMWF Reanalysis 5th Generation

FDR: Flow Direction

GDBC: Global Distributed Basin Characteristics

GEV: Generalized Extreme Value

GPCC: Global Precipitation Climatology Centre

GSIM: Global Streamflow Indices and Metadata

HDI: Human Development Index

HSE: High Streamflow Extremes

ILDAS: Indian Land Data Assimilation System

IMD: India Meteorological Department

IDW: Inverse Distance Weighting

INDOFLOODS: Indian Observational Flood Events Database

LSH: Large-sample Hydrological

MK: Mann-Kendall

MSWEP: Multi-Source Weighted-Ensemble Precipitation

NH: Northern Hemisphere

PSL: Physical Science Laboratory

QRF: Quantile Random Forest

RF: Random Forests

SHAP: Shapley Additive exPlanations

SH: Southern Hemisphere

SRTM: Shuttle Radar Topography Mission

TFPW: Trend Free Pre-Whitening

UPD: Unit Peak Discharge

USA: United States of America

USD: United States Dollar