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

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 1ocations 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 साईं किरण

प्रवेश संख्या: 2019CEZ8628

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

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