A PHYSICALLY INFORMED DEEP LEARNING FRAMEWORK FOR SCALABLE STREAMFLOW MODELING IN INDIA

BHANU MAGOTRA

DEPARTMENT OF CIVIL ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2025

A PHYSICALLY INFORMED DEEP LEARNING FRAMEWORK FOR SCALABLE STREAMFLOW MODELING IN INDIA

by

BHANU MAGOTRA

Department of Civil Engineering

submitted

In fulfilment of the requirement for the degree of

DOCTOR OF PHILOSOPHY

to the

INDIAN INSTITUTE OF TECHNOLOGY DELHI

JULY 2025

This dissertation is dedicated to my family.

To my parents, Darshan Kumar and Shakti Devi, for their unconditional love, sacrifices, and unwavering belief in me; to my brother, Safal Magotra, for his constant support.

And to my spouse, Vidushi Koul, for her invaluable love, understanding, and encouragement.

CERTIFICATE

This is to certify that the dissertation entitled **"A PHYSICALLY INFORMED DEEP LEARNING FRAMEWORK FOR SCALABLE STREAMFLOW MODELING IN INDIA"**, being submitted by **Mr. Bhanu Magotra**, 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 of 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

Prof. Dhanya C.T.
Department of Civil Engineering
Indian Institute of Technology Delhi
New Delhi- 110016, India

ACKNOWLEDGEMENTS

Completing this Ph.D. dissertation has been a journey marked by both challenges and triumphs, and it would not have been possible without the support, guidance, and affirmation of numerous individuals. I am profoundly grateful for the assistance provided by all those who have accompanied me on this path.

Foremost, I extend my sincere gratitude to my advisors, Prof. Manabendra Saharia and Prof. Dhanya C.T., for their unwavering support, invaluable guidance, and insightful feedback throughout my doctoral studies. Their patience, expertise, and encouragement have been instrumental in shaping this dissertation and fostering my development as a researcher.

I also express my heartfelt thanks to my research committee members, Prof. Gazala Habib, Prof. A.K. Keshari, and Prof. B. K. Panigrahi, for their time, valuable suggestions, and steadfast support, which were crucial to this work.

I am particularly indebted to my colleagues from the HydroSense Lab; their friendship, collaboration, and continued encouragement have been invaluable during this period. Special recognition is due to all my colleagues within the HydroX Lab and Water Resources Engineering section for cultivating a stimulating and supportive environment.

I would like to express my sincere gratitude to my masters' supervisor, Prof. H. Ramesh, who introduced me to the world of research and provided valuable guidance and encouragement to pursue this work.

Acknowledging all who have influenced my learning journey presents a considerable challenge. To anyone who encounters this acknowledgment, please accept my deep appreciation for the opportunities, mentorship, and friendships you have offered. Your contributions have been indispensable in achieving this milestone.

(Bhanu Magotra)

ABSTRACT

Streamflow plays a central role in both providing water for human and economic needs and making a critical contribution to ecological health, making its accurate prediction vital for holistic water management. From ensuring agricultural productivity and industrial development to providing reliable drinking water and preserving ecological balance, the multifaceted demands on this vital resource underscore the need for sophisticated streamflow modeling techniques that balance scalability with local relevance. Addressing this challenge is pivotal for enhancing water resources management strategies and ensuring water security across the nation. This dissertation is oriented towards developing and assessing such models, with a particular emphasis on integrating advanced deep learning techniques with traditional hydrological modeling approaches to predict streamflow more effectively.

The first aim of this research focused on developing an Indian Land Data Assimilation System (ILDAS), motivated by the need for a high-resolution, modular, and open-source hydrological modelling framework over the Indian Subcontinent. ILDAS integrates a coupled land surface model with a hydrodynamic model, enhancing its ability to estimate various hydrological parameters crucial for managing water resources effectively. A detailed reanalysis product spanning 1981 to 2021 is generated by integrating data from four distinct meteorological forcing datasets, selected for their reliability and specificity, to account for various atmospheric and climatic influences. Additionally, all major components of the terrestrial water balance, such as surface runoff, soil moisture, terrestrial water storage anomalies, evapotranspiration, and streamflow, were validated against a diverse array of satellite and in situ observations. This rigorous validation process not only ensures the accuracy of ILDAS outputs but also underscores the system's potential to become a vital tool for water resources management, adapted to the unique challenges faced by India. By having the capabilities of both terrestrial and atmospheric data assimilation, ILDAS represents a

significant advancement in the generation of hydrologically robust and locally relevant information for managing water resources sustainably and mitigating the impacts of extreme weather events.

The streamflow evaluation from ILDAS revealed that at an uncalibrated stage, the predictions are not suitable for operational purposes. Additionally, calibrating a large-scale hydrological model on multiple river basins is computationally prohibitive. Therefore, this research explored an alternative data-driven approach by training state-of-the-art deep learning models: Long Short-Term Memory (LSTM) networks to predict daily streamflow at 220 gauge locations across India. We assessed the efficacy of incorporating meteorological information, catchment-specific geophysical parameters and dynamic land surface states as supplementary inputs to LSTM models. Our results demonstrated that incorporating catchment geophysics into the models led to a substantial improvement in performance over the base LSTM model, However, when it came to the inclusion of dynamic land surface states, the impact on model performance did not align with the expectations, indicating that such information did not contribute significantly to the overall predictive capability of the models.

Building on these findings, the dissertation introduces a hybrid solution aimed at optimizing streamflow prediction by combining ensemble outputs from large-scale hydrological models of ILDAS with advanced deep learning post-processor based on LSTM networks. We proposed a two-stage correction and combination process that involved the residual correction of ensemble members followed by their optimal integration with the moving average of observed streamflow data. Designed to enhance local relevance of streamflow predictions, this approach is validated across the same 220 gauge locations, demonstrating marked improvements in accuracy and reliability. This methodology, sensitive to local hydrological contexts, presents a significant advancement in the area of generating accurate streamflow estimates relevant to various water resources management practices.

Overall, this dissertation seeks to understand the potential of hybrid models to overcome complex challenges in streamflow modeling and provide the hydrological community with a scalable framework that can integrate physical information with deep learning models. The strategies explored in this dissertation demonstrate that hybrid models, which merge physical hydrological principles with deep learning algorithms, yield superior performance, proving especially beneficial for streamflow modeling in complex environments like India.

सार

जल प्रबंधन में स्ट्रीमफ्लो (नदी प्रवाह) की सटीक भविष्यवाणी महत्वपूर्ण है क्योंकि यह मानवीय और आर्थिक ज़रूरतों के लिए पानी उपलब्ध कराने के साथ-साथ पारिस्थितिक स्वास्थ्य में भी अहम योगदान देता है। कृषि उत्पादकता और औद्योगिक विकास सुनिश्चित करने से लेकर विश्वसनीय पेयजल प्रदान करने और पारिस्थितिक संतुलन बनाए रखने तक, इस महत्वपूर्ण संसाधन की बहुआयामी मांगें परिष्कृत स्ट्रीमफ्लो मॉडलिंग तकनीकों की आवश्यकता को रेखांकित करती हैं। इन तकनीकों में मापनीयता और स्थानीय प्रासंगिकता के बीच संतुलन होना चाहिए। इस चुनौती का समाधान जल संसाधन प्रबंधन रणनीतियों को बढ़ाने और देश भर में जल सुरक्षा सुनिश्चित करने के लिए महत्वपूर्ण है। यह थीसिस ऐसे मॉडल विकसित करने और उनका आकलन करने पर केंद्रित है, जिसमें स्ट्रीमफ्लो की अधिक प्रभावी ढंग से भविष्यवाणी करने के लिए उन्नत डीप लर्निंग तकनीकों को पारंपरिक हाइड्रोलॉजिकल मॉडलिंग दृष्टिकोणों के साथ एकीकृत करने पर विशेष जोर दिया गया है।

इस शोध का पहला उद्देश्य भारतीय उपमहाद्वीप पर उच्च-रिज़ॉल्यूशन, मॉड्यूलर और ओपन-सोर्स हाइड्रोलॉजिकल मॉडलिंग ढांचे की आवश्यकता से प्रेरित होकर भारतीय भूमि डेटा एसिमिलेशन सिस्टम (ILDAS) विकसित करना था। ILDAS एक युग्मित भूमि सतह मॉडल को हाइड्रोडायनामिक मॉडल के साथ एकीकृत करता है, जो जल संसाधनों के प्रभावी ढंग से प्रबंधन के लिए महत्वपूर्ण विभिन्न हाइड्रोलॉजिकल मापदंडों का अनुमान लगाने की इसकी क्षमता को बढ़ाता है। 1981 से 2021 तक का विस्तृत पुनर्विश्लेषण उत्पाद चार अलग-अलग मौसम संबंधी बल डेटासेट से डेटा को एकीकृत करके तैयार किया गया है, जिन्हें विभिन्न वायुमंडलीय और जलवायु प्रभावों को ध्यान में रखते हुए उनकी विश्वसनीयता और विशिष्टता के लिए चुना गया है। इसके अतिरिक्त, स्थलीय जल संतुलन के सभी प्रमुख घटकों, जैसे सतही अपवाह, मिट्टी की नमी, स्थलीय जल भंडारण विसंगतियाँ, वाष्पोत्सर्जन और जलप्रवाह को उपग्रह और इन-सीटू अवलोकनों की एक विविध सरणी के विरुद्ध मान्य किया गया था। यह कठोर

सत्यापन प्रक्रिया न केवल ILDAS आउटपुट की सटीकता सुनिश्चित करती है, बल्कि भारत के सामने आने वाली अनूठी चुनौतियों के अनुकूल जल संसाधन प्रबंधन के लिए एक महत्वपूर्ण उपकरण बनने की प्रणाली की क्षमता को भी रेखांकित करती है। स्थलीय और वायुमंडलीय दोनों डेटा को आत्मसात करने की क्षमता रखने से, ILDAS जल संसाधनों को स्थायी रूप से प्रबंधित करने और चरम मौसम की घटनाओं के प्रभावों को कम करने के लिए हाइड्रोलॉजिकल रूप से मजबूत और स्थानीय रूप से प्रासंगिक जानकारी के उत्पादन में एक महत्वपूर्ण प्रगति का प्रतिनिधित्व करता है। आईएलडीएएस के प्रवाह मूल्यांकन से पता चला है कि बिना कैलिब्रेट किए गए चरण में, पूर्वानुमान परिचालन उद्देश्यों के लिए उपयुक्त नहीं हैं। इसके अतिरिक्त, कई नदी घाटियों पर बड़े पैमाने पर हाइड्रोलॉजिकल मॉडल को कैलिब्रेट करना कम्प्यूटेशनल रूप से निषेधात्मक है। इसलिए, इस शोध ने अत्याधुनिक डीप लर्निंग मॉडल को प्रशिक्षित करके एक वैकल्पिक डेटा-संचालित दृष्टिकोण की खोज की: भारत भर में 220 गेज स्थानों पर दैनिक प्रवाह की भविष्यवाणी करने के लिए लॉन्ग शॉर्ट-टर्म मेमोरी (एलएसटीएम) नेटवर्क। हमने एलएसटीएम मॉडल में पूरक इनपुट के रूप में मौसम संबंधी जानकारी, जलग्रहण-विशिष्ट भूभौतिकीय मापदंडों और गतिशील भूमि सतह की स्थिति को शामिल करने की प्रभावकारिता का आकलन किया। हमारे परिणामों ने प्रदर्शित किया कि मॉडल में जलग्रहण भूभौतिकी को शामिल करने से आधार एलएसटीएम मॉडल की तुलना में प्रदर्शन में पर्याप्त सुधार हुआ, इन निष्कर्षों के आधार पर, थीसिस एक हाइब्रिड समाधान प्रस्तुत करती है जिसका उद्देश्य एलएसटीएम नेटवर्क पर आधारित उन्नत डीप लर्निंग पोस्ट-प्रोसेसर के साथ ILDAS के बड़े पैमाने के हाइड्रोलॉजिकल मॉडल से एनसेंबल आउटपुट को जोड़कर स्ट्रीमफ्लो भविष्यवाणी को अनुकूलित करना है। हमने दो-चरणीय सुधार और संयोजन प्रक्रिया का प्रस्ताव दिया जिसमें एनसेंबल सदस्यों के अवशिष्ट सुधार के बाद उनके देखे गए स्ट्रीमफ्लो डेटा के मूविंग एवरेज के साथ इष्टतम एकीकरण शामिल था। स्ट्रीमफ्लो भविष्यवाणियों की स्थानीय प्रासंगिकता को बढ़ाने के लिए डिज़ाइन किया गया, इस दृष्टिकोण को समान 220 गेज स्थानों पर मान्य किया गया है, जो पूर्वानुमान सटीकता और विश्वसनीयता में उल्लेखनीय सुधार प्रदर्शित करता है। स्थानीय

हाइड्रोलॉजिकल संदर्भों के प्रति संवेदनशील यह पद्धति बाढ़ जोखिम प्रबंधन के क्षेत्र में एक महत्वपूर्ण प्रगति प्रस्तुत करती है, साथ ही व्यापक जल संसाधन प्रबंधन अभ्यास भी। इस थीसिस के दौरान किए गए शोध के संश्लेषण से संकेत मिलता है कि मशीन लर्निंग फ्रेमवर्क में भौतिक अंतर्दृष्टि को शामिल करने से बाढ़ की भविष्यवाणी क्षमताओं में काफी वृद्धि होती है। अध्ययन न केवल बाढ़ जोखिम मूल्यांकन और प्रबंधन में निहित जटिल चुनौतियों पर काबू पाने में हाइब्रिड मॉडल की क्षमता के बारे में हमारी समझ को आगे बढ़ाते हैं बल्कि हाइड्रोलॉजिकल समुदाय को विभिन्न मॉडलिंग दृष्टिकोणों को एकीकृत करने के लिए एक व्यावहारिक और लचीला ढांचा भी प्रदान करते हैं। इस थीसिस में प्रस्तुत मान्य प्रगति अधिक लचीली, अनुकूली और स्थानीय रूप से प्रासंगिक बाढ़ जोखिम प्रबंधन रणनीतियों के विकास के लिए आधार प्रदान करती है। ये रणनीतियाँ जलवायु परिवर्तन और जल विज्ञान चक्र में मानवजनित परिवर्तनों से उत्पन्न होने वाले बढ़ते खतरों से निपटने के लिए अनिवार्य हैं, जो महत्व को रेखांकित करती हैं।

शोध का शीर्षक: भारत में स्केलेबल स्ट्रीमफ्लो मॉडलिंग के लिए भौतिक रूप से सूचित गहन शिक्षण ढाँचा

छात्र का नाम: भानु मगोत्रा

प्रवेश संख्या: 2019CEZ8621

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ADAM	Adaptive Moment Estimation
ANN	Artificial Neural Networks
BMA	Bayesian Model Averaging
CaLDAS	Canadian Land Data Assimilation System
CAMELS	Catchment Attributes and Meteorology for Large-sample Studies
CCD	Cold Cloud Duration
CHIRPS	Climate Hazards Center Infrared Precipitation with Station data
CLSM	Catchment Land Surface Model
CONUS	Conterminous United States
CPU	Central Processing Unit
CRNP	Cosmic Ray Neutron Probe
CWC	Central Water Commission
DEM	Digital Elevation Model
DL	Deep Learning
ECMWF	European Centre for Medium-Range Weather Forecasts
ELDAS	European Land Data Assimilation System
ERA	ECMWF Reanalysis
ESA-CCI	European Space Agency - Climate Change Initiative
ET	Evapotranspiration
FDC	Flow Duration Curve
FHV	High Flow Volume
FLV	Low Flow Volume
FMS	Mid-Section Slope
GEOS	Goddard Earth Observing System

GLDAS	Global Land Data Assimilation System
GLMPP	General Linear Model Post-Processor
GMAO	Global Modeling and Assimilation Office
GPCC	Global Precipitation Climatology Centre
GPU	Graphics Processing Unit
GRACE	Gravity Recovery and Climate Experiment
GRUN	Global Runoff Reconstruction
GSIM	Global Streamflow Indices and Metadata
GSWP	Global Soil Wetness Project
HPC	High Performance Computing
HyMAP	Hydrological Modeling and Analysis Platform
ICON	Indian Cosmic Ray Network
ILDAS	Indian Land Data Assimilation System
IMD	Indian Meteorological Department
IMDAA	Indian Monsoon Data Assimilation and Analysis
IQR	Inter Quartile Range
IS	In-Sample
IWRIS	India Water Resources Information
JJAS	June July August September
KGE	Kling-Gupta Efficiency
KGESS	Kling-Gupta Efficiency Skill Score
LAI	Leaf Area Index
LDAS	Land Data Assimilation System
LDT	Land Data Toolkit
LISF	Land Information System Framework

LPRM	Land Parameter Retrieval Model
LSM	Land Surface Model
LSTM	Long Short-Term Memory
LSTM-CA	Long Short-Term Memory with Catchment Attributes
LSTM-PBM	Long Short-Term Memory with Physically Based Model
LU/LC	Land Use/ Land Cover
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDH	Madahalli
MERIT	Multi-Error-Removed Improved-Terrain
MERRA	Modern-Era Retrospective analysis for Research and Applications
MODIS	Moderate Resolution Imaging Spectroradiometer
MP	Multi Parameterization
NASA	National Aeronautics and Space Administration
NLDAS	North American Land Data Assimilation System
NMAE	Normalized Mean Absolute Error
NRSC	National Remote Sensing Centre
NSE	Nash Sutcliffe Efficiency
OS	Out-of-Sample
RMSE	Root Mean Squared Error
RNN	Recurrent Neural Networks
RRMSE	Relative Root Mean Squared Error
RTX	Ray Tracing Texel eXtreme
SALDAS	South American Land Data Assimilation System
SGR	Singanallur

SIMGM	Simple Groundwater Model
TOPMODEL	Topography-based Hydrological Model
TWSA	Terrestrial Water Storage Anomalies
RMSE	Root Mean Squared Error
USD	United States Dollars
VRAM	Video Random Access Memory