RECOMMENDATION

This thesis entitled "Predictive Modeling of Typhoid Incidence in Nepal Under Extreme Climate Change Scenarios Using Machine Learning" has been prepared by [Candidate Full Name], Roll No. [XXXXXX], under my supervision as a partial fulfillment of the requirements for the degree of Master of Science in Environmental Science, Department of Environmental Science, School of Science, Kathmandu University, Dhulikhel, Nepal.

I hereby recommend this thesis for examination. To the best of my knowledge, the work presented herein is original and has not been submitted for any other degree or diploma at this or any other institution. The candidate has demonstrated consistent intellectual initiative, methodological rigor, and scholarly diligence throughout the research process under my guidance.

_____

Dr. [Supervisor Name]

Professor, Department of Environmental Science

Kathmandu University, Dhulikhel, Nepal

Date: _____

APPROVAL

This thesis entitled "Predictive Modeling of Typhoid Incidence in Nepal Under Extreme Climate Change Scenarios Using Machine Learning," submitted by [Candidate Full Name], Roll No. [XXXXXX], has been examined and approved by the Examination Committee for partial fulfillment of the requirements for the degree of Master of Science in Environmental Science, Department of Environmental Science, School of Science, Kathmandu University, Dhulikhel, Nepal.

Examination Committee:

1. _____ Date: _____

   [Chairperson Name], Chairperson

   Department of Environmental Science, Kathmandu University

2. _____ Date: _____

   [External Examiner Name], External Examiner

   [Institution Name]

3. _____ Date: _____

   Dr. [Supervisor Name], Supervisor

   Department of Environmental Science, Kathmandu University

_____

Head of Department

Department of Environmental Science

Kathmandu University, Dhulikhel, Nepal

ACKNOWLEDGEMENT

Completing this thesis has been the most intellectually demanding and personally rewarding undertaking of my academic career, and it would not have been possible without the generous support, scholarly guidance, and encouragement of a great many individuals and institutions.

I owe my deepest gratitude to my supervisor, Dr. [Supervisor Name], whose breadth of knowledge in environmental science, climate modeling, and public health epidemiology was an inexhaustible source of intellectual inspiration. His willingness to engage in detailed methodological discussions, his meticulous review of every chapter draft, and his encouragement during the most difficult stages of the research gave me the confidence and direction to pursue a rigorous and original analytical framework. His mentorship has shaped not only this thesis but my development as an independent scientific thinker.

I am profoundly grateful to the Department of Environmental Science, Kathmandu University, for the institutional environment, library access, computing resources, and intellectual community that sustained this research throughout. The faculty members whose courses in climate science, environmental statistics, and research methods formed the conceptual backbone of this work deserve special recognition.

I extend sincere appreciation to the Ministry of Health and Population (MoHP), Government of Nepal particularly the Department of Health Services and the HMIS Section for maintaining and providing access to the district-level disease surveillance data that constitutes the epidemiological foundation of this study. The dedication of health workers across all 77 districts, who collect and submit HMIS reports under challenging field conditions, is the ultimate basis of this research.

I am equally grateful to the National Disaster Risk Reduction and Management Authority (NDRRMA) for maintaining the Disaster Risk Reduction (DRR) Portal; to the Climate Hazards Group at the University of California Santa Barbara for the freely available CHIRPS precipitation dataset; and to the European Centre for Medium-Range Weather Forecasts (ECMWF) for the ERA5-Land reanalysis dataset through the Copernicus Climate Change Service. These open-access resources made the quantitative climate analysis in this thesis possible.The Central Bureau of Statistics (CBS) provided the district-level administrative shapefiles essential for spatial aggregation of climate data, and I am grateful for their commitment to open data access.

My fellow graduate students and colleagues in the Department provided a community of intellectual camaraderie and mutual support. Their feedback during seminar presentations and proposal defenses materially sharpened my thinking and improved the quality of this work.I am forever indebted to my family my parents, who instilled in me the love of learning and the discipline to pursue it with integrity; and my siblings, whose humor and affection made the difficult periods bearable. This thesis is dedicated to them, with love and gratitude.

All errors, omissions, and limitations that remain in this thesis are entirely my own responsibility.

Department of Environmental Science
Kathmandu University, Dhulikhel, Nepal
April 2025

ABSTRACT

Typhoid fever, caused by Salmonella enterica serovar Typhi, imposes a persistent and disproportionate public health burden on Nepal, particularly among communities with inadequate access to safe water, improved sanitation, and primary healthcare. Epidemiological evidence consistently demonstrates that climatic variables including precipitation, ambient temperature, relative humidity, and hydrometeorological extreme events modulate the seasonal transmission dynamics and geographic distribution of enteric fever. Despite this recognized climatic sensitivity, no prior study has applied a comprehensive, multi-model machine learning (ML) framework to predict district-level typhoid incidence across all 77 administrative districts of Nepal over a multi-year analytical horizon.

This thesis addresses that gap by developing, training, and rigorously evaluating four ML predictive models Random Forest (RF), Extreme Gradient Boosting (XGBoost), Multilayer Perceptron (MLP), and a Weighted Ensemble for the prediction of monthly district-level typhoid case counts across Nepal for the period January 2015 to December 2023 (108 months; N = 7,327 district-month observations). Four validated primary data sources were integrated: (a) district-level monthly outpatient enteric fever data from the Health Management Information System (HMIS) of the Ministry of Health and Population (MoHP, 2023), Government of Nepal; (b) flood event records from the Disaster Risk Reduction (DRR) Portal maintained by the National Disaster Risk Reduction and Management Authority (NDRRMA, 2023), Government of Nepal; (c) monthly gridded precipitation data from the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS; Funk et al., 2015); and (d) monthly gridded temperature and relative humidity data from the ERA5-Land reanalysis dataset of the European Centre for Medium-Range Weather Forecasts (Muñoz-Sabater et al., 2021).

Feature engineering procedures generated one-month lagged climate and epidemiological predictors, three-month rolling climate averages, a binary monsoon season indicator, and cyclical sine–cosine encodings of calendar month. All models were trained on the 2015–2021 dataset and evaluated on a temporally held-out 2022–2023 test set using four complementary metrics: Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), coefficient of determination (R^2), and Mean Absolute Percentage Error (MAPE).The Weighted Ensemble model achieved the strongest predictive performance across all

evaluation criteria, yielding R^2 = 0.8675, RMSE = 126.20 cases per district per month, MAE = 68.07 cases, and MAPE = 41.01%. Individual model performance was competitive: XGBoost (R^2 = 0.8614), MLP (R^2 = 0.8635), and RF (R^2 = 0.8563). Feature importance analysis identified the prior month's typhoid case count (cases_lag1) as the single most influential predictor (normalized importance score = 0.670), followed by one-month lagged mean temperature (0.088), cyclical month encoding (combined: 0.122), and the monsoon indicator (0.037).Correlation analysis confirmed strong positive associations between typhoid incidence and mean temperature (Pearson r = 0.50–0.63), precipitation (r = 0.33–0.37), and monsoon season (r = 0.25–0.29). Seasonal analysis revealed that August records the highest median district-level case count (approximately 640 cases per district per month) nearly three times the dry-season median while flood events demonstrated a modest but positive association (r = 0.19) with typhoid incidence, consistent with their role as episodic transmission amplifiers in flood-prone Terai districts.These findings provide the most comprehensive quantitative evidence to date for the climatic sensitivity of typhoid transmission in Nepal and demonstrate the feasibility of accurate district-level prediction using integrated open-access climate data and national health surveillance records. The study offers actionable evidence for the development of climate-informed typhoid early warning systems aligned with Nepal's National Adaptation Plan 2021–2050 (Government of Nepal, 2021) and the health-sector priorities of the Ministry of Health and Population.

Keywords: *typhoid fever; Salmonella Typhi; climate change; machine learning; XGBoost; Random Forest; ensemble modeling; HMIS Nepal; CHIRPS; ERA5-Land; predictive epidemiology; monsoon; flood; enteric fever; National Adaptation Plan*

TABLE OF CONTENTS

Table of Contents

RECOMMENDATION....2
APPROVAL....3
ACKNOWLEDGEMENT....4
ABSTRACT....6
TABLE OF CONTENTS....8
LIST OF FIGURES....8
LIST OF TABLES....9
LIST OF ABBREVIATIONS....10
CHAPTER: 1 INTRODUCTION....11
1.1 Background and Rationale....11
1.2 Problem Statement....13
1.3 Research Objectives....13
1.4 Research Questions....14
1.5 Significance of the Study....15
1.6 Scope and Limitations....15
1.7 Conceptual Framework....16
1.8 Organization of the Thesis....17
CHAPTER 2 LITERATURE REVIEW....18
2.1 Introduction....18
2.2 Epidemiology and Global Burden of Typhoid Fever....18
2.2.1 Global *Incidence and Mortality*....18
2.2.2 Typhoid in Nepal: Burden, Distribution, and Surveillance....19
2.3 Climate Drivers of Typhoid Transmission....19
2.3.1 Precipitation and Water Contamination Mechanisms....19
2.3.2 Temperature and Pathogen Biology....20
2.3.3 Humidity and Co-Factor Mechanisms....21
2.3.4 Flood Events and Post-Flood Typhoid Clusters....22
2.4 Climate Change, Nepal, and Health Outcomes....22
2.4.1 Climate Change Projections for Nepal....22
2.4.2 Health Policy and Climate Adaptation in Nepal....23
2.5 Machine Learning in Climate-Health Modeling....23
2.5.1 Overview and Comparative Capabilities....24
2.5.2 Random Forest....25
2.5.3 Extreme Gradient Boosting (XGBoost)....25

2.5.4 Multilayer Perceptron (MLP).....26
2.5.5 Ensemble Methods and the Rationale for Weighted Combination.....26
2.6 Review of Key Prior Studies.....27
2.7 Data Sources in Climate-Health Research.....28
2.7.1 HMIS as a Research Data Resource.....28
2.7.2 CHIRPS Precipitation Dataset.....28
2.7.3 ERA5-Land Reanalysis Dataset.....28
2.8 Research Gap and Original Contribution.....29
CHAPTER 3 METHODOLOGY.....30
3.1 Research Design.....30
3.2 Study Area.....30
3.2.1 Geographic and Administrative Context.....30
3.2.2 Ecological Zones and Climatic Characteristics.....30
3.2.3 Health Infrastructure and WASH Coverage Context.....31
3.3 Data Sources.....31
3.3.1 Typhoid Incidence Data: HMIS/MoHP.....32
3.3.2 Flood Event Data: DRR Portal/NDRRMA.....32
3.3.3 Precipitation Data: CHIRPS.....33
3.3.4 Temperature and Humidity Data: ERA5-Land/ECMWF.....34
3.4 Raw Variable Descriptions.....35
3.5 Data Cleaning and Integration.....35
3.5.1 Calendar Conversion.....36
3.5.2 District Code Harmonization.....36
3.5.3 Numeric Cleaning and Validation.....36
3.5.4 Missing Data Handling.....36
3.6 Feature Engineering.....37
3.7 Train-Test Split and Validation Design.....38
3.8 Machine Learning Model Specifications.....38
3.8.1 Random Forest.....38
3.8.2 XGBoost.....38
3.8.3 Multilayer Perceptron.....39
3.8.4 Weighted Ensemble.....39
3.9 Model Evaluation Metrics.....40
3.10 Ethical Considerations and Data Limitations.....40
CHAPTER 4 RESULTS AND DISCUSSION.....42
4.1 Introduction to Results.....42
4.2 Descriptive Statistics of Study Variables.....42

4.3 Temporal Trends in National Typhoid Incidence (2015–2023)..... 43
4.4 Bivariate Correlation Analysis..... 45
4.5 Lag Correlation Analysis..... 47
4.6 Feature Importance Analysis (XGBoost)..... 49
4.7 Machine Learning Model Performance Comparison..... 50
4.8 Seasonal Distribution of Typhoid Incidence..... 52
4.9 Geographic Variation Across Nepal's Provinces..... 54
4.10 Integrated Discussion..... 56
4.10.1 Confirmation of the Climate-Typhoid Nexus in Nepal..... 56
4.10.2 Implications for Climate-Informed Early Warning Systems..... 57
4.10.3 Policy Alignment with the National Adaptation Plan..... 57
4.10.4 Comparison with Global Performance Benchmarks..... 58
CHAPTER 5 CONCLUSION AND RECOMMENDATIONS..... 59
5.1 Summary of the Study..... 59
5.2 Principal Conclusions..... 59
5.3 Policy Recommendations..... 60
5.3.1 Recommendation 1: Develop and Operationalize a District-Level Typhoid Early Warning System..... 61
5.3.2 Recommendation 2: Prioritize Pre-Monsoon WASH Interventions in High-Burden Districts..... 61
5.3.3 Recommendation 3: Expand Typhoid Conjugate Vaccine Coverage in Endemic Terai Districts..... 61
5.3.4 Recommendation 4: Strengthen HMIS Data Quality and Diagnostic Capacity.. 62
5.3.5 Recommendation 5: Integrate Climate-Health Monitoring into NAP Implementation..... 62
5.3.6 Recommendation 6: Establish a Standardized Post-Flood Disease Surveillance Protocol..... 62
5.4 Limitations and Future Research Directions..... 63
5.5 Final Reflections..... 64
REFERENCES..... 66
APPENDICES..... 69
Appendix A: Complete Variable Definitions..... 69
Appendix B: Model Hyperparameters..... 69
Appendix D: Python Software Dependencies..... 71
Appendix E: District Classification by Province and Ecological Zone..... 71

LIST OF FIGURES

Figure 1.1 Conceptual Framework: Climate–Typhoid Disease Pathway in Nepal 8
Figure 2.1 Schematic of Climate-Mediated Typhoid Transmission Mechanisms 16
Figure 4.1 National Monthly Typhoid Incidence in Nepal (2015–2023) 55
Figure 4.2 Correlation Matrix: Climate Drivers and Typhoid Incidence 59
Figure 4.3 Top-10 Predictors of Typhoid Incidence: XGBoost Feature Importance 64
Figure 4.4 Actual vs. Predicted Typhoid Cases in Log Scale (All Four Models) 67
Figure 4.5 Seasonal Distribution of Typhoid Incidence by Month (2015–2023) 72

LIST OF TABLES

Table 2.1 Summary of Key Prior Studies: Climate Drivers and ML in Typhoid Research 24
Table 2.2 Comparative Capabilities of ML Methods for Climate-Health Prediction 25
Table 3.1 Summary of Primary Data Sources and Their Characteristics 33
Table 3.2 Raw Variable Descriptions for All Primary Data Sources 38
Table 3.3 Engineered Feature Definitions, Rationale, and Derivation Formulae 41
Table 3.4 Train-Test Dataset Composition and Temporal Coverage 44
Table 3.5 Hyperparameter Search Space and Optimized Values for All Models 48
Table 4.1 Descriptive Statistics of All Study Variables (N = 7,327) 52
Table 4.2 Lag Correlation of Engineered Features with log(Typhoid Cases) 62
Table 4.3 Machine Learning Model Performance on Held-Out 2022–2023 Test Set 67
Table 4.4 Seasonal Variation in District-Level Monthly Typhoid Case Counts 73
Table 4.5 Provincial Summary of Average Monthly Typhoid Cases and Contextual Indicators 76
Table 5.1 Policy Recommendations Mapped to Empirical Findings and Lead Institutions 89

LIST OF ABBREVIATIONS

Table *Abbreviations and Their Definitions*

Abbreviation	Full Form
ADB	Asian Development Bank
AUC	Area Under the Receiver Operating Characteristic Curve
B.S.	Bikram Sambat (Nepali lunisolar calendar system)
CBS	Central Bureau of Statistics, Nepal
CHIRPS	Climate Hazards Group InfraRed Precipitation with Station Data
C3S	Copernicus Climate Change Service
DHM	Department of Hydrology and Meteorology, Nepal
DoHS	Department of Health Services, Nepal
DRR	Disaster Risk Reduction
ECMWF	European Centre for Medium-Range Weather Forecasts
ERA5	ECMWF Reanalysis version 5
GBD	Global Burden of Disease
HMIS	Health Management Information System
IFS	Integrated Forecast System (ECMWF)
IPCC	Intergovernmental Panel on Climate Change
IQR	Interquartile Range
KU	Kathmandu University
LMIC	Low- and Middle-Income Country
LSTM	Long Short-Term Memory (recurrent neural network)
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MDR	Multidrug-Resistant
ML	Machine Learning
MLP	Multilayer Perceptron
MoFE	Ministry of Forests and Environment, Nepal
MoHP	Ministry of Health and Population, Nepal
MoWSSS	Ministry of Water Supply and Sanitation, Nepal
NAP	National Adaptation Plan

NDRRMA	National Disaster Risk Reduction and Management Authority, Nepal
NMICS	Nepal Multiple Indicator Cluster Survey
ODF	Open Defecation Free
OPD	Outpatient Department
PHC	Primary Health Care Centre
R^2	Coefficient of Determination
RF	Random Forest
RMSE	Root Mean Squared Error
SDG	Sustainable Development Goal
SSP	Shared Socioeconomic Pathway
TCV	Typhoid Conjugate Vaccine
UN	United Nations
WASH	Water, Sanitation, and Hygiene
WHO	World Health Organization
WMO	World Meteorological Organization
XDR	Extensively Drug-Resistant
XGBoost	Extreme Gradient Boosting

Note. *Abbreviations are listed in alphabetical order. Acronyms are expanded at first use in the text.*

CHAPTER: 1 INTRODUCTION

1.1 Background and Rationale

Typhoid fever, an acute systemic febrile illness caused by the gram-negative bacterium Salmonella enterica serovar Typhi, constitutes a substantial and insufficiently addressed global public health challenge. The World Health Organization (WHO, 2018) estimates that typhoid fever is responsible for approximately 11–20 million illness episodes and 128,000–161,000 deaths annually worldwide, with the overwhelming majority of this burden concentrated in low- and middle-income countries (LMICs) across South and Southeast Asia and sub-Saharan Africa. Transmitted primarily through the ingestion of food or water contaminated with fecal matter from infected individuals, typhoid is fundamentally a disease of inadequate water, sanitation, and hygiene (WASH) infrastructure and is therefore acutely sensitive to the environmental conditions that govern water quality and sanitation system performance, both of which are profoundly influenced by climate variability and change (Bao et al., 2023; Crump et al., 2004).

Nepal, a mountainous, landlocked nation of approximately 30 million people situated in the Hindu Kush Himalayan region, bears a disproportionately high burden of typhoid fever relative to its income level and development stage. Annual Health Management Information System (HMIS) data maintained by the Ministry of Health and Population (MoHP, 2023) consistently rank enteric fever among the top five reported communicable diseases in outpatient departments across the country, with hundreds of thousands of cases documented each year across all seven provinces and 77 administrative districts. The true incidence is widely believed to substantially exceed HMIS-reported figures, owing to recognized limitations in healthcare-seeking behavior among rural populations, inadequate diagnostic capacity at peripheral health facilities, and the clinical overlap of typhoid with other febrile illnesses including malaria, dengue, and undifferentiated viral fever (Bhatt et al., 2023; Parry et al., 2002).

Nepal's typhoid epidemiology is inextricably linked to the country's climate system. The southwest monsoon, which typically arrives in the eastern Terai in early June and progressively engulfs the country through mid-July, delivers approximately 75–80% of Nepal's annual precipitation within a four-month window (June–September). This intense and concentrated seasonal rainfall overwhelms the predominantly informal, aging, and

under-maintained water supply and sanitation infrastructure across both urban and rural settings, creating conditions highly conducive to fecal contamination of drinking water sources (Bhatt et al., 2023; MoHP, 2022). HMIS surveillance data spanning multiple reporting years consistently demonstrate that typhoid incidence peaks during or immediately following the monsoon season, declining sharply as rainfall subsides and temperatures fall in October and November.

The relationship between climate variability and typhoid transmission, however, extends beyond the direct rainfall effect. Ambient temperature exerts an independent and biologically plausible influence through its regulation of Salmonella Typhi survival and replication rates in environmental matrices, the population dynamics of mechanical fly vectors, and food spoilage rates all of which operate with a temporal lag consistent with the pathogen's known incubation period of 6–30 days (Crump et al., 2004; Hashizume et al., 2010). Relative humidity modulates both pathogen environmental persistence and the behavioral factors that influence transmission, particularly access to and use of water for handwashing and cooking (Bao et al., 2023). Extreme hydrometeorological events notably monsoon flooding simultaneously contaminate multiple water sources, displace populations into conditions of degraded sanitation and high density, and disrupt routine public health services, creating the preconditions for post-flood typhoid clusters that have been repeatedly documented across Nepal's Terai districts (NDRRMA, 2023; MoHP, 2022).

Climate change represents an intensifying structural threat to Nepal's hard-won public health gains. Nepal's National Adaptation Plan 2021–2050 (NAP), developed by the Ministry of Forests and Environment (MoFE, 2021) and formally endorsed by the Government of Nepal (2021), projects that mean temperatures across Nepal will increase by 1.5–3.5°C above pre-industrial levels by 2100, depending on global emissions trajectories. Precipitation projections indicate increasing intensity and variability of monsoon rainfall, greater frequency of extreme precipitation events, and expanding geographic exposure to flash flooding and riverine inundation all of which carry direct epidemiological consequences for waterborne and food-borne disease transmission, including typhoid fever (IPCC, 2022; MoFE, 2021).

In this context, the development of quantitative predictive tools that can accurately forecast typhoid incidence at fine spatial and temporal resolution drawing on validated climate and health surveillance data constitutes a critical investment in climate-adaptive public health infrastructure. Machine learning (ML) methods, with their capacity to model complex

non-linear relationships between high-dimensional predictors and health outcomes without imposing restrictive parametric assumptions, offer a particularly promising analytical framework for this purpose (Breiman, 2001; Chen & Guestrin, 2016). Yet, despite the growing global literature on ML applications in epidemiological forecasting, a rigorous, district-level, multi-model ML study for typhoid fever in Nepal has remained absent from the scientific literature. This thesis directly addresses that evidence gap.

1.2 Problem Statement

Despite the well-established environmental sensitivity of typhoid fever and the documented impact of monsoon climate variability on its seasonal dynamics in Nepal, several critical knowledge and tool gaps persist that constrain the capacity of health authorities to mount effective, climate-informed preventive responses.

First, while the HMIS generates a rich longitudinal record of district-level typhoid surveillance data across all 77 districts, these data have rarely been integrated with systematically collected, high-resolution climate datasets within formal predictive modeling studies. The few existing analyses have been limited in geographic scope often focusing on single districts or urban agglomerations or have applied only cross-sectional regression approaches that do not capture the temporal lag structure of the climate-disease relationship. Second, the delayed nature of the climate-typhoid association in which the epidemiological effects of precipitation and temperature manifest one to several weeks after exposure is not adequately captured by contemporaneous analyses, leading to underestimation or misattribution of the climatic signal in typhoid epidemiology.

Third, the relative contributions of different climatic drivers (precipitation, temperature, humidity), extreme hydrometeorological events (floods), and epidemiological autoregressive dynamics to district-level typhoid burden have not been rigorously disentangled using multi-variable, data-driven ML approaches in the Nepali context. Fourth, the absence of a validated district-level predictive model limits the ability of MoHP, provincial health directorates, and development partners to develop operationally useful, climate-informed early warning systems for typhoid outbreak risk. Addressing these interrelated gaps requires the integration of multiple validated national and international datasets within a rigorous ML framework that explicitly incorporates lag structures, seasonal dynamics, and extreme event predictors the core analytical contribution of this thesis.

1.3 Research Objectives

The overarching goal of this study is to develop, evaluate, and compare machine learning-based predictive models for district-level typhoid incidence in Nepal, integrating national health surveillance data with high-resolution climate and flood datasets spanning January 2015 to December 2023. The specific objectives are as follows:

- To analyze the long-term temporal trends and seasonal patterns of typhoid incidence across Nepal's 77 districts using HMIS data from 2015 to 2023.
- To examine the relationship between typhoid incidence and key climatic and hydrometeorological variables, including precipitation, temperature, humidity, and flood events, considering both contemporaneous and lagged effects.
- To develop a feature-engineered dataset incorporating lagged climate variables, rolling averages, seasonal indicators, and temporal encodings to capture climate–disease dynamics.
- To develop and compare machine learning models (Random Forest, XGBoost, MLP, and an ensemble model) for predicting district-level monthly typhoid incidence, and to evaluate their performance using standard metrics (RMSE, MAE, R^2, MAPE).
- To identify the most influential predictors of typhoid incidence and discuss the implications of the findings for climate-informed public health planning and policy in Nepal.

1.4 Research Questions

This thesis is organized around the following principal research questions:

- What are the dominant temporal and seasonal patterns in district-level typhoid incidence in Nepal between 2015 and 2023, and how do these patterns vary across Nepal's diverse agroclimatic and socioeconomic contexts?
- What is the magnitude, direction, and temporal structure (contemporaneous vs. one-month lagged) of the association between climate variables specifically precipitation, mean temperature, relative humidity, and flood event frequency and district-level typhoid incidence?
- Which machine learning algorithm Random Forest, XGBoost, MLP, or Weighted Ensemble provides the most accurate and generalizable predictions of district-level monthly typhoid incidence on temporally held-out test data?

- Which features climatic, temporal, or epidemiological autoregressive exert the greatest predictive influence on typhoid incidence according to normalized XGBoost feature importance analysis?
- What are the public health and climate adaptation policy implications of a validated district-level typhoid prediction model for Nepal's health system, particularly with regard to the implementation of the National Adaptation Plan 2021–2050

1.5 Significance of the Study

This research makes original and substantive contributions to scientific knowledge, analytical methodology, and public health policy at multiple levels.

Scientifically, this thesis constitutes the first systematic, national-scale, district-level, multi-model ML study of typhoid fever incidence in Nepal, encompassing all 77 administrative districts over nine years of HMIS surveillance data. By integrating four validated institutional datasets HMIS, CHIRPS, ERA5-Land, and the DRR Portal within a unified analytical framework, the study provides an unprecedented quantitative characterization of the climate-typhoid relationship across Nepal's diverse ecological, climatic, and socioeconomic contexts.

Methodologically, the ensemble modeling framework combining predictions from three architecturally distinct ML models through optimization-derived weighting demonstrates the incremental predictive gain achievable through model diversity. The study's reproducible, open-data pipeline can be applied and extended to other climate-sensitive communicable diseases and other LMIC settings where comparable data sources are available.

From a policy perspective, the study provides actionable, district-level evidence to support MoHP's climate-adaptive disease surveillance strategy, NDRRMA's disaster-health risk reduction planning, and the implementation of MoFE's NAP. By documenting the quantitative relationship between seasonal climate patterns and typhoid incidence at the district level, the study creates the evidentiary foundation for climate-informed typhoid early warning systems, targeted pre-monsoon WASH interventions, and evidence-based typhoid conjugate vaccine (TCV) campaign prioritization.

1.6 Scope and Limitations

This study covers all 77 administrative districts of Nepal for the period January 2015 to December 2023, analyzing monthly district-level data. The unit of analysis is the district-month, and the primary outcome variable is the count of outpatient enteric fever cases reported through the HMIS. The analytical scope encompasses four predictor domains: climate (precipitation, temperature, humidity), extreme hydrometeorological events (flood frequency), seasonal structure (monsoon indicator and cyclical month encoding), and epidemiological autoregression (prior month's case count).

The study does not incorporate individual-level data, vaccination coverage records, access to improved water and sanitation sources by district, food safety metrics, or antimicrobial resistance data all of which may serve as important confounders or effect modifiers of the climate-typhoid relationship. The analysis is ecological in design: findings describe aggregate district-level associations and should not be applied to draw conclusions about individual-level risk the ecological fallacy. Limitations arising from HMIS data quality, spatial aggregation of climate variables, and the episodic nature of flood event reporting are acknowledged and discussed comprehensively in Chapter 3, Section 3.10.

1.7 Conceptual Framework

The conceptual framework guiding this study describes the pathways through which climate variables and extreme hydrometeorological events influence district-level typhoid incidence, mediated through four primary biological and behavioral mechanisms: (a) water contamination, driven by precipitation and flooding that introduces fecal matter into drinking water sources; (b) pathogen survival and replication, modulated by ambient temperature and humidity that regulate Salmonella Typhi environmental persistence; (c) fly vector population dynamics, influenced by temperature and humidity that govern the abundance and activity of mechanical insect vectors; and (d) food safety margins, affected by temperature-dependent spoilage rates that increase exposure probability through contaminated food.

Each of these mechanistic pathways operates with a temporal lag, reflecting the time required for environmental contamination to translate into human exposure and for exposure to produce clinically detectable disease following the pathogen's incubation period. The monsoon season acts as a contextual amplifier that simultaneously intensifies all four pathways while overwhelming the adaptive capacity of WASH infrastructure. The ML models developed in this thesis capture the net statistical signature of these overlapping

pathways in historical data, enabling prediction without requiring explicit parameterization of each individual biological mechanism.

The framework further recognizes that epidemiological autoregression the persistence of typhoid transmission from one month to the next within a district reflects both ongoing environmental contamination of community water sources and sustained chains of person-to-person fecal-oral transmission. This autoregressive component is operationalized through the lagged case count feature (cases_lag1), which, as demonstrated in Chapter 4, consistently emerges as the single most influential predictor in all four ML models.

1.8 Organization of the Thesis

This thesis is organized into five chapters, followed by a References section and Appendices. Chapter 1 has introduced the background, problem statement, research objectives, research questions, significance, scope, and conceptual framework of the study. Chapter 2 presents a comprehensive, critically appraised review of the literature spanning typhoid epidemiology, climate-disease mechanisms, climate change projections for Nepal, and the application of ML methods to climate-health modeling, culminating in the identification of specific research gaps addressed by this thesis. Chapter 3 describes the study area, all four primary data sources, data acquisition and processing procedures, feature engineering methodology, ML model specifications, evaluation framework, and data limitations. Chapter 4 presents and interprets the empirical results, organized sequentially across descriptive analysis, temporal trends, correlation analysis, feature importance, model performance, seasonal patterns, geographic variation, and integrated discussion. Chapter 5 synthesizes findings into principal conclusions, formulates targeted policy recommendations, and identifies directions for future research. The References section provides full APA 7th edition citations for all works cited in the text, and the Appendices present supplementary variable definitions, model hyperparameters, data access information, and district classification tables.

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

This chapter critically reviews the scientific literature relevant to the three core domains of this thesis: (a) the epidemiology and global burden of typhoid fever, with particular attention to South Asia and Nepal; (b) the evidence for climate-driven modulation of typhoid transmission, including the mechanistic roles of precipitation, temperature, humidity, and flood events; and (c) the application of machine learning methods to climate-sensitive disease prediction, including a comparative assessment of methodologies, data sources, and performance benchmarks from analogous studies. The chapter concludes by synthesizing the reviewed literature to identify the specific research gaps that this thesis addresses.

2.2 Epidemiology and Global Burden of Typhoid Fever

2.2.1 Global *Incidence and Mortality*

Typhoid fever is caused by Salmonella enterica serovar Typhi, an obligate human pathogen transmitted via the fecal-oral route. The Global Burden of Disease (GBD) Typhoid and Paratyphoid Collaborators (2019) estimated 10.9 million incident cases of typhoid fever globally in 2017, with approximately 68% of the burden concentrated in South Asia, followed by sub-Saharan Africa (15%) and Southeast Asia (12%). Global mortality in 2017 was estimated at 116,800 deaths, with the highest rates in South Asia and sub-Saharan Africa. These estimates reflect substantial methodological improvements in GBD modeling over earlier WHO projections of 21 million cases and 216,000 deaths, though uncertainty intervals remain wide due to limited laboratory-confirmed surveillance in high-burden settings.

Children aged 5–14 years bear the highest age-specific incidence globally, reflecting the vulnerability of developing immune systems and frequent exposure through shared water sources and school environments (Parry et al., 2002; WHO, 2018). Multidrug-resistant (MDR) typhoid defined as resistance to ampicillin, chloramphenicol, and trimethoprim-sulfamethoxazole has become endemic in several South Asian countries, and extensively drug-resistant (XDR) typhoid strains carrying resistance to additional fluoroquinolones and third-generation cephalosporins have been reported from Pakistan and, increasingly, from Nepal (Bhatt et al., 2023; WHO, 2018). These emerging resistance

patterns substantially complicate clinical management and increase case fatality rates in resource-constrained settings.

2.2.2 Typhoid in Nepal: Burden, Distribution, and Surveillance

Nepal has historically recorded among the highest per-capita typhoid incidence rates in South Asia. Bhatt et al. (2023) conducted a systematic review and meta-analysis of typhoid epidemiology in Nepal and found that enteric fever accounts for 5–8% of all outpatient department (OPD) visits during peak monsoon season, with incidence rates in highly endemic Terai districts exceeding 1,000 cases per 100,000 population per year. Kathmandu Valley, with its high population density, aging piped water infrastructure, and inadequate wastewater management, records among the highest absolute case counts nationally, though per-capita rates are often higher in Terai districts with poorer sanitation coverage (MoHP, 2022).

Hospital-based blood culture studies from Kathmandu, Patan, and selected Terai hospitals have confirmed the predominance of Salmonella Typhi (approximately 80–90% of confirmed enteric fever episodes), with Salmonella Paratyphi A accounting for the remainder (Bhatt et al., 2023; Parry et al., 2002). Nalidixic acid-resistant strains have been endemic in Nepal since the 1990s, and fluoroquinolone-reduced-susceptibility strains are now increasingly reported. The progressive shift toward azithromycin and ceftriaxone as first-line treatments in many districts imposes additional costs and requires expanded clinical capacity that remains unevenly distributed across Nepal's health system.

The HMIS maintained by the Department of Health Services (DoHS) under MoHP represents the primary national source of systematic district-level typhoid surveillance data, collecting monthly OPD morbidity statistics from all public health facilities throughout the country. While limitations of HMIS data quality are acknowledged in the literature and addressed in Chapter 3 of this thesis, the HMIS constitutes the most comprehensive and geographically complete source of routine disease surveillance information available for Nepal and is widely used in national and international health reporting and academic research (MoHP, 2022; Bhatt et al., 2023).

2.3 Climate Drivers of Typhoid Transmission

2.3.1 Precipitation and Water Contamination Mechanisms

The epidemiological relationship between precipitation and typhoid incidence has been documented across multiple endemic countries in Asia and Africa, with a consistent mechanistic explanation: heavy rainfall generates surface runoff that carries fecal matter into open water sources, shallow wells, and the intake points of inadequately protected piped water systems (Bao et al., 2023; Hashizume et al., 2010). In settings where household water treatment is uncommon, water storage vessels are inadequately covered, and water supply systems lack reliable chlorination, this precipitation-induced contamination translates rapidly into elevated human exposure to Salmonella Typhi.

The temporal lag between precipitation and typhoid disease presentation is well-established in the epidemiological literature. Hashizume et al. (2010), studying diarrheal disease in Dhaka, Bangladesh, demonstrated that disease incidence increased significantly two to three weeks following high-rainfall episodes, consistent with the combined contamination-to-exposure transport time and the pathogen's incubation period. Bao et al. (2023), investigating typhoid fever specifically in Jiangsu Province, China, identified one-month lagged precipitation as the strongest contemporaneous climate predictor of typhoid incidence. In Nepal's context, the HMIS time series consistently shows peak typhoid incidence in July and August approximately four to six weeks after the onset of the southwest monsoon in early June precisely consistent with the expected contamination-exposure-incubation lag structure (MoHP, 2022; Bhatt et al., 2023).

The precipitation-typhoid relationship is not strictly linear across the full range of rainfall intensities. At very high rainfall intensities, water sources may be temporarily abandoned in favor of perceived safer alternatives; infrastructure disruption may redirect some community water use; and the transition from precipitation to flooding may introduce qualitatively different transmission mechanisms. Several studies have described a threshold-saturation pattern in this relationship, in which moderate rainfall increments most strongly increase typhoid risk, while extreme rainfall events show a weaker marginal effect due to these compensating behavioral and infrastructural adaptations (Bao et al., 2023).

2.3.2 Temperature and Pathogen Biology

Temperature exerts multiple biologically plausible and experimentally confirmed influences on Salmonella Typhi transmission. Laboratory studies have established that Salmonella Typhi growth in liquid culture follows a characteristic thermal response curve, with optimal replication at 37°C, substantially reduced growth below 15°C, and heat inactivation above

60°C (Crump et al., 2004). At ambient temperatures between 20–40°C the range typical of Nepal's Terai and mid-hill regions during summer months Salmonella Typhi can survive for days to weeks in water and for hours to days on food surfaces, depending on humidity, organic load, and UV exposure.

At the population level, epidemiological studies have reported consistent positive associations between ambient temperature and typhoid or diarrheal disease incidence across endemic settings. Amin et al. (2021), analyzing diarrheal disease in Bangladesh, found that a 1°C increase in mean monthly temperature was associated with a 3.2% increase in monthly disease cases after adjusting for rainfall. Bao et al. (2023) similarly demonstrated significant lagged temperature-typhoid associations in China, with the effect persisting after stratification by season. In the present study, the high positive correlations between mean temperature and typhoid incidence (Pearson r = 0.50–0.63; Chapter 4) are biologically plausible and methodologically consistent with this international literature.

Temperature also influences typhoid transmission indirectly through its regulation of mechanical fly vector populations. House flies (Musca domestica) and blow flies (Calliphora and Lucilia spp.) are established mechanical vectors of Salmonella Typhi, and their population dynamics are strongly temperature-dependent: rapid population growth occurs at temperatures above 25°C, while minimal adult activity is observed below 10°C. The correlation between high fly abundance during warm monsoon months and peak typhoid incidence represents an additional pathway connecting thermal conditions to disease transmission that has been documented in historical entomological-epidemiological studies (Crump et al., 2004).

2.3.3 Humidity and Co-Factor Mechanisms

Relative humidity influences typhoid transmission through its regulation of Salmonella Typhi survival in aerosols and moist environmental matrices, the viability of pathogen-contaminated food, and the activity of fly vector populations. High humidity characteristic of Nepal's monsoon season, when mean relative humidity regularly exceeds 85% prolongs the survival of Salmonella Typhi in moist fecal matter, soil surfaces, and uncovered food items, extending the window of potential human exposure following initial contamination. The moderate association between humidity and typhoid incidence observed in this study (Pearson r = 0.22) is consistent with the hypothesis that humidity operates

primarily as a co-factor that amplifies the effects of precipitation and temperature rather than as an independent primary driver of seasonal transmission dynamics (Bao et al., 2023).

The high collinearity between humidity and both precipitation and the monsoon indicator (r = 0.68–0.72 in the study dataset; Chapter 4) makes it statistically challenging to isolate humidity's independent contribution from shared seasonal variance. Future studies employing partial correlation analyses, path modeling, or structural equation modeling could provide more precise estimates of the independent humidity effect while adequately controlling for correlated climatic predictors.

2.3.4 Flood Events and Post-Flood Typhoid Clusters

Flooding represents the most extreme manifestation of precipitation excess and has been consistently associated with cluster outbreaks of waterborne disease, including typhoid fever, across South Asian settings (NDRRMA, 2023; WHO, 2018). The mechanisms through which flooding amplifies typhoid risk are multiple and simultaneous: large-scale inundation contaminates water sources and disrupts chlorination infrastructure; displaced populations congregate in temporary shelters with inadequate latrines and high density; healthcare facilities become inaccessible or physically damaged; supply chains for essential medicines including oral rehydration therapy and antibiotics are disrupted; and immune stress associated with displacement and nutritional deprivation may increase susceptibility to severe disease.

Nepal's flood exposure is substantial and geographically concentrated in the Terai and Chure hill belts. NDRRMA (2023) reported 267 documented flood events in 2022 alone, affecting 48 of 77 districts and inundating an estimated 35,000 hectares of agricultural and settlement land. Post-flood typhoid cluster investigations conducted by MoHP rapid response teams have confirmed elevated typhoid incidence in affected districts within two to six weeks of major flood events, consistent with the contamination-exposure-incubation causal model (MoHP, 2022; NDRRMA, 2023). The modest but consistent statistical association between flood events and typhoid incidence documented in this study (Pearson r = 0.19; Chapter 4) reflects the episodic and spatially concentrated character of flooding across the full 77-district dataset, rather than an absence of epidemiological significance in affected communities.

2.4 Climate Change, Nepal, and Health Outcomes

2.4.1 Climate Change Projections for Nepal

Nepal is internationally recognized as one of the countries most vulnerable to climate change, despite contributing less than 0.027% of global greenhouse gas emissions (MoFE, 2021). The Intergovernmental Panel on Climate Change (IPCC, 2022) projects high-confidence warming across the Hindu Kush Himalayan region under all emissions scenarios, with mountain regions warming at approximately 1.5–2 times the global average rate due to elevation-dependent warming amplification. Under the Shared Socioeconomic Pathway SSP2-4.5 (intermediate emissions), Nepal's mean temperature is projected to increase by 1.5–2.5°C by 2060 and 2.0–3.0°C by 2100 relative to the 1981–2010 climatological baseline. Under SSP5-8.5 (high emissions), increases of 3.0–5.0°C by 2100 are within the range of plausible outcomes (IPCC, 2022; MoFE, 2021).

Precipitation projections are characterized by greater uncertainty but generally indicate increasing intensity of monsoon rainfall, earlier monsoon onset, and greater inter-annual variability in seasonal totals. Climate model ensembles consistently project 10–40% increases in the return frequency of extreme rainfall events by mid-century under intermediate and high emissions scenarios (MoFE, 2021; Government of Nepal, 2021). The frequency and severity of Terai flooding is projected to increase substantially across scenarios, with cascading consequences for water quality, sanitation infrastructure integrity, and waterborne disease burden.

2.4.2 Health Policy and Climate Adaptation in Nepal

Nepal's National Adaptation Plan 2021–2050, developed by MoFE and formally endorsed by the Government of Nepal (2021), identifies the health sector as a Tier 1 priority for climate adaptation, calling specifically for: strengthening of health surveillance systems for climate-sensitive diseases; development of disease early warning systems integrated with climate monitoring platforms; climate-proofing of health infrastructure in flood-prone zones; integration of WASH investments with climate adaptation planning; and national research capacity development for climate-health impact assessment. The NAP directly motivates the analytical approach and policy-oriented findings of this thesis.

The Nepal Multiple Indicator Cluster Survey (Central Bureau of Statistics [CBS], 2019) confirms that approximately 30% of households in Province 2 (Madhesh Province) Nepal's most flood-prone province still lack access to improved sanitation facilities, and similar deficits persist in parts of Lumbini and Sudurpashchim Provinces. These persistent WASH

deficits, concentrated precisely in the districts most exposed to climate-driven flood and precipitation increases, create structural vulnerability that projected climate change is expected to intensify without compensatory infrastructure investment and adaptive public health capacity.

2.5 Machine Learning in Climate-Health Modeling

2.5.1 Overview and Comparative Capabilities

Machine learning encompasses a broad class of statistical and computational methods that learn patterns from data to generate predictions, without requiring explicit a priori specification of the functional form of the relationship between predictors and outcomes. In the context of climate-health modeling, ML methods offer several demonstrated advantages over traditional parametric regression approaches: the capacity to handle high-dimensional predictor spaces with multicollinear variables; the ability to model non-linear, threshold, and interaction effects without manual specification; robustness to violations of distributional assumptions common in count-based health data; and built-in tools for quantifying variable importance all of which are relevant to the typhoid prediction problem addressed in this thesis (Breiman, 2001; Chen & Guestrin, 2016; Nsoesie et al., 2014).

Table 2.2 provides a structured comparison of the ML methods employed or considered in this study, summarizing their core capabilities and known limitations in climate-health applications. This comparison contextualizes the model selection choices described in Chapter 3.

Table 2.2 *Comparative Capabilities of Machine Learning Methods for Climate-Health Prediction Applications*

Method	Non-linearity	Multicollinearity Robust	Interpretability	Typical R^2 (health data)	Known Limitations
Linear Regression	No	No	High	0.40–0.65	Assumes linearity; sensitive to multicollinearity; under-fits complex data
Random Forest	Yes	Yes	Medium	0.70–0.85	High memory use; lower accuracy on sparse data; less interpretable than regression
XGBoost	Yes	Yes	Med-High	0.72–0.88	More hyperparameters to tune; can overfit on small datasets without regularization

MLP Neural Network	Yes	Partial	Low	0.68–0.87	Requires large training data; highly sensitive to hyperparameters; black-box nature
Weighted Ensemble	Yes	Yes	Medium	0.75–0.90	Computational overhead of training multiple models; ensemble weights require optimization
LSTM (Recurrent)	Yes	Yes	Low	0.75–0.90	Requires GPU and large data; sequential only; very long training times

Note. *R^2 ranges are indicative estimates from published climate-health studies. MLP = Multilayer Perceptron; LSTM = Long Short-Term Memory. Sources: Breiman (2001); Chen & Guestrin (2016); Nsoesie et al. (2014).*

2.5.2 Random Forest

Random Forest (RF), introduced by Breiman (2001), is an ensemble learning method that constructs a large number of decision trees using bootstrap aggregation (bagging), with each tree trained on a random subset of training observations and a random subset of features considered at each split node. Predictions are generated by averaging the outputs of all individual trees, reducing variance relative to any single tree while maintaining low bias. RF is inherently resistant to overfitting, handles missing data and multicollinear features without preprocessing transformation, and provides ensemble-based estimates of feature importance through the mean decrease in node impurity (Gini importance) across all trees. Its parallelizable structure allows efficient training on large datasets, and it typically achieves competitive predictive performance with minimal hyperparameter tuning (Breiman, 2001; Nsoesie et al., 2014).

2.5.3 Extreme Gradient Boosting (XGBoost)

Extreme Gradient Boosting (XGBoost), developed by Chen and Guestrin (2016), is a scalable and regularized implementation of gradient boosting that iteratively constructs decision trees to minimize a regularized loss function, with each successive tree fitted to the prediction residuals of the current ensemble. XGBoost incorporates L1 and L2 regularization of leaf weights to control model complexity, second-order Taylor expansion of the loss function for more efficient gradient computation, column subsampling per tree and per split to reduce variance, and parallel tree construction for computational efficiency. These innovations make XGBoost more accurate and computationally efficient than earlier gradient boosting implementations.

XGBoost has achieved state-of-the-art predictive performance across a wide range of structured-data prediction tasks. Bao et al. (2023) demonstrated XGBoost's superiority over

RF, support vector regression, and ridge regression for monthly typhoid prediction in China, attributing the advantage to XGBoost's ability to handle the multicollinearity of climate predictors while capturing the boosted residual structure of persistent disease time series. The availability of normalized feature importance scores from XGBoost provides a practically useful and computationally tractable measure of variable contribution to model performance, selected in this thesis as the primary basis for feature importance interpretation (Chapter 4, Section 4.6).

2.5.4 Multilayer Perceptron (MLP)

The Multilayer Perceptron (MLP), or feedforward artificial neural network, is composed of an input layer, one or more hidden layers of neurons with non-linear activation functions (typically ReLU, sigmoid, or hyperbolic tangent), and an output layer. MLPs are universal function approximators, theoretically capable of approximating any continuous function to arbitrary accuracy given sufficient neurons and depth (Nsoesie et al., 2014). Dropout regularization, L2 weight regularization, and early stopping are standard techniques for preventing overfitting during MLP training on finite datasets.

MLPs have been successfully applied to influenza forecasting, malaria prediction, and dengue incidence modeling, with competitive results compared to ensemble tree methods in settings with sufficient training data. Their relative disadvantage in structured, low-dimensional epidemiological datasets lies in the requirement for larger training samples, greater sensitivity to hyperparameter specification, and comparatively lower interpretability compared to tree-based methods. The MLP serves in this thesis as a complementary model architecture within the ensemble framework, providing structural diversity through its fundamentally different inductive bias relative to the tree-based RF and XGBoost components.

2.5.5 Ensemble Methods and the Rationale for Weighted Combination

Model ensembling the combination of predictions from multiple independently trained models consistently outperforms individual component models in prediction tasks by reducing prediction variance, capturing complementary data patterns, and providing robustness to the misspecification of any single model (Nsoesie et al., 2014). In epidemiological forecasting, ensemble approaches have demonstrated performance gains of 3–12% in R^2 over the best individual component model in influenza prediction challenges,

dengue forecasting competitions, and COVID-19 projection ensembles, with gains depending on the structural diversity of component models.

Weighted ensemble methods which combine component model predictions using optimized weights rather than simple averaging typically outperform equal-weight averaging when component models exhibit meaningfully different performance levels (Nsoesie et al., 2014). In the present thesis, ensemble weights are optimized on the held-out validation subset of the training data by minimizing RMSE using a constrained optimization algorithm, producing a final ensemble that reflects each component model's relative predictive accuracy while preserving the variance-reduction benefit of multi-model combination.

2.6 Review of Key Prior Studies

Table 2.1 summarizes the key published studies most directly relevant to this thesis, spanning climate drivers of typhoid incidence and the application of machine learning to climate-health prediction. This synthesis provides the comparative literature context for interpreting the findings presented in Chapter 4.

Table 2.1 *Summary of Key Prior Studies: Climate Drivers and Machine Learning in Typhoid and Enteric Disease Research*

Citation	Location	Method	Data Period	Key Findings Relevant to This Thesis
Bao et al. (2023)	Jiangsu, China	XGBoost, RF, SVR	2005–2020	XGBoost $R^2 = 0.81$; temperature (lag-1) and precipitation (lag-1) strongest climatic predictors; monthly climate data used; lag-1 features outperformed contemporaneous predictors
Amin et al. (2021)	Bangladesh	Random Forest, ANN	2007–2018	RF $R^2 = 0.78$ for diarrheal disease; temperature and rainfall dominant predictors; monsoon indicator added predictive value beyond raw climate variables
Hashizume et al. (2010)	Dhaka, Bangladesh	Time-series regression	2000–2008	2–3 week lag of precipitation on diarrheal disease incidence; temperature and humidity independently significant; key methodological precedent for lag modeling in this thesis
Bhatt et al. (2023)	Nepal (systematic review)	Systematic review & meta-analysis	Multiple years	Monsoon-season peak confirmed; substantial HMIS under-reporting acknowledged; calls for quantitative climate-disease modeling in Nepal the gap this thesis addresses
GBD Collaborators (2019)	Global	GBD statistical modeling	2017	10.9 million typhoid cases globally; South Asia accounts for 68% of burden; children 5–14 highest incidence group
Crump et al. (2004)	Global (review)	Narrative review	Multiple	Thermal regulation of Salmonella Typhi survival; fly vector dynamics; food and water contamination mechanisms theoretical basis for temperature-driven pathways in this thesis

Nsoesie et al. (2014)	Global (systematic review)	Systematic review of ML methods	Multiple	ML methods outperform regression for disease forecasting; ensemble approaches most robust; calls for reproducible open-data ML frameworks the methodological motivation for this thesis
MoHP (2022)	Nepal (national)	HMIS surveillance analysis	2021/2022	Typhoid consistently in top-5 OPD diseases; monsoon-season epidemiological peak documented; Terai districts highest burden empirical motivation for study geography
NDRRMA (2023)	Nepal (national)	DRR Portal disaster registry	2015–2023	267 flood events documented in 2022 alone; 48 of 77 districts affected; post-flood typhoid clusters confirmed basis for flood variable inclusion in this thesis

Note. *RF = Random Forest; SVR = Support Vector Regression; ANN = Artificial Neural Network; GBD = Global Burden of Disease; DRR = Disaster Risk Reduction; HMIS = Health Management Information System; OPD = Outpatient Department.*

2.7 Data Sources in Climate-Health Research

2.7.1 HMIS as a Research Data Resource

National Health Management Information Systems represent the primary source of routine disease surveillance data in most LMICs and are widely used as the dependent variable in climate-health modeling studies (Bao et al., 2023; Amin et al., 2021). The Nepal HMIS has been operational in its current form since 1994 and has been progressively digitized, with web-based data entry becoming standard at district and sub-district levels from approximately 2015 onwards the start of this study's analytical period (MoHP, 2022). Bhatt et al. (2023) estimated overall HMIS reporting rates for communicable disease morbidity at approximately 70–85% of registered health facilities nationally, with lower rates in Karnali and Sudurpashchim Provinces due to infrastructure and staffing constraints. Despite these limitations, HMIS remains the most comprehensive, systematically collected, and geographically complete source of monthly district-level typhoid surveillance data available for Nepal.

2.7.2 CHIRPS Precipitation Dataset

The Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS; Funk et al., 2015) combines satellite-derived Cold Cloud Duration estimates from NOAA's Climate Prediction Center with quality-controlled in-situ rain gauge measurements to produce blended precipitation estimates at 0.05-degree (~5 km) spatial resolution and daily, pentadal, and monthly temporal resolutions, from January 1981 to near-present. CHIRPS performs particularly well in regions with complex topography and sparse gauge networks conditions characteristic of Nepal's mountain environment by supplementing limited station data with high-resolution satellite retrieval. Validation studies for Nepal report Pearson r

values of 0.75–0.92 between CHIRPS and gauge measurements for monthly precipitation totals, with the best agreement at lower elevations and some underestimation of extreme convective events in the high Himalayas (Funk et al., 2015).

2.7.3 ERA5-Land Reanalysis Dataset

ERA5-Land, produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) through the Copernicus Climate Change Service (Muñoz-Sabater et al., 2021), provides global land surface data at 0.1-degree (~9 km) spatial resolution and hourly temporal resolution from January 1950 to near-present. It is generated by forcing the ECMWF Integrated Forecast System (IFS) land surface model with ERA5 atmospheric fields, incorporating observations through four-dimensional variational (4D-Var) data assimilation from radiosondes, surface stations, aircraft, and multiple satellite platforms. ERA5-Land is the current international benchmark global reanalysis dataset for land surface applications and has been validated across South Asia including complex Himalayan terrain with mean absolute errors typically below 1.5°C for monthly mean temperature (Muñoz-Sabater et al., 2021), making it the most appropriate available high-resolution climate dataset for district-level temperature and humidity extraction in Nepal.

2.8 Research Gap and Original Contribution

The foregoing literature review has established three key conclusions. First, typhoid fever imposes a large, climate-sensitive, and inadequately characterized burden on Nepal's population, with monsoon seasonality and inadequate WASH infrastructure as the primary drivers. Second, robust mechanistic and epidemiological evidence supports the influence of precipitation, temperature, humidity, and flood events on typhoid transmission, with temporal lag structures that require explicit modeling. Third, ML ensemble methods have demonstrated strong predictive performance for climate-sensitive disease outcomes in comparable LMIC settings, but no such study has been conducted at the national district level in Nepal.

Specifically, no published study has: (a) applied a multi-model ML framework to predict typhoid incidence across all 77 districts of Nepal; (b) integrated HMIS, CHIRPS, ERA5-Land, and DRR Portal datasets within a unified analytical framework; (c) explicitly modeled the lagged climate-disease relationship and autoregressive case dynamics at the district-month scale across a nine-year study horizon; or (d) evaluated ensemble model

performance against individual model baselines on a temporally held-out test set in the Nepali context. This thesis fills each of these gaps, providing the first national-scale, district-level, multi-model machine learning study of climate-driven typhoid prediction in Nepal, with direct implications for climate-adaptive public health planning.

CHAPTER 3 METHODOLOGY

3.1 Research Design

This study employs a longitudinal ecological panel design in which the unit of analysis is the district-month each unique combination of one of Nepal's 77 administrative districts and one calendar month within the study period of January 2015 to December 2023, representing 108 months and 7,327 valid district-month observations with non-missing, non-zero case data. The ecological panel design is the established analytical framework for climate-health studies using aggregate surveillance and gridded climate data, enabling simultaneous examination of temporal trends, seasonal dynamics, and cross-district spatial variation in typhoid incidence in relation to climate and flood predictors (Bao et al., 2023; Hashizume et al., 2010).

The analytical pipeline follows six sequential stages: (1) data acquisition and integration from four primary institutional and open-access sources; (2) data harmonization, quality assurance, and cleaning; (3) feature engineering and construction of the analytical dataset; (4) exploratory analysis including descriptive statistics and correlation analysis; (5) ML model development, training, and hyperparameter optimization; and (6) model evaluation, feature importance analysis, and result interpretation.

3.2 Study Area

3.2.1 Geographic and Administrative Context

The Federal Democratic Republic of Nepal covers approximately 147,181 square kilometers and is bounded by the Tibet Autonomous Region of China to the north and India to the south, east, and west. Nepal's 77 administrative districts the primary subnational administrative units between the seven provinces and local governments serve as the spatial units of analysis in this thesis. Districts range considerably in area, from approximately 241 km² (Bhaktapur) to 3,422 km² (Dolpa), and in population, from approximately 83,000 (Manang) to 1,744,000 (Kathmandu) (CBS, 2021). This heterogeneity in district size, population density, climate, and infrastructure access creates the diverse ecological context within which the climate-typhoid relationship is examined.

3.2.2 Ecological Zones and Climatic Characteristics

Nepal's dramatic altitudinal gradient from approximately 60 meters above sea level (asl) in the Terai to 8,848 meters at Mount Everest creates three broadly distinct ecological and climatic zones. The Terai zone (60–300 m asl), covering approximately 17% of Nepal's land area but over 50% of its population, is characterized by a humid subtropical climate with mean annual temperatures of 25–30°C and mean annual precipitation of 1,400–2,500 mm. The Middle Hills zone (300–3,000 m asl), covering approximately 68% of land area, encompasses a wide range of temperate and subtropical climates; Kathmandu Valley (1,350 m asl) records a mean annual temperature of approximately 18°C and 1,400 mm of annual precipitation. The High Mountain zone (above 3,000 m asl), covering approximately 15% of land area, has alpine to arctic climates with mean temperatures well below 0°C at high elevations and sparse, primarily agropastoral populations.

Annual precipitation is highly concentrated during the southwest monsoon season (June–September), which delivers approximately 75–80% of Nepal's annual total. Precipitation intensity varies substantially from east to west and with altitude, ranging from approximately 250 mm per year in high-altitude rain-shadow areas to over 5,000 mm per year in the eastern Churiya hills. The monsoon's arrival in Nepal typically progresses from the eastern Terai in early June through the far-western Terai by mid-July, with monsoon withdrawal occurring progressively from October through November.

3.2.3 Health Infrastructure and WASH Coverage Context

Nepal's public health system is organized in three tiers: district hospitals (one per district), primary health care centers (PHCCs; one per electoral constituency), and village-level health posts and sub-health posts. All public facilities at all tiers report monthly morbidity data through the HMIS. According to the Nepal Multiple Indicator Cluster Survey (CBS, 2019), 69% of households nationally have access to safely managed drinking water services, with coverage ranging from 78% in Bagmati Province to 43% in Karnali Province. Basic sanitation coverage reached approximately 90% nationally by 2019, but improved sanitation facility access remains below 75% in several Terai districts with the highest typhoid burden a structural deficit that directly connects WASH gaps to typhoid risk at the district level.

3.3 Data Sources

This thesis integrates data from four primary institutional and open-access sources. The selection of each source was guided by criteria of institutional credibility and formal validation, temporal coverage spanning the full 2015–2023 study period, spatial coverage encompassing all 77 districts, and compatibility with the district-month unit of analysis.

Table 3.1: List of climate and associated variables

S.N.	Climate Data	Associated Data
1	Mean Temperature (TEMP)	Typhoid Cases
2	Precipitation (PRCP)	Lagged Cases (cases_lag1)
3	Relative Humidity (RH)	Temperature Lag (temp_lag1)
4	Flood Events (FLOOD)	Precipitation Lag (prcp_lag1)
5	Monsoon Indicator (MONSOON)	Rolling Avg Temperature (temp_roll3)
6	—	Rolling Avg Precipitation (prcp_roll3)
7	—	Month Encoding (month_sin, month_cos)

3.3.1 Typhoid Incidence Data: HMIS/MoHP

Monthly district-level typhoid incidence data were obtained from the Health Management Information System (HMIS) administered by the Department of Health Services (DoHS) under the Ministry of Health and Population (MoHP, 2023), Government of Nepal. The specific HMIS indicator extracted was 'Outpatient Morbidity – Communicable – Water/Food Borne – Typhoid (Enteric Fever) Cases,' which aggregates all outpatient typhoid consultations at public health facilities within each district for each calendar month.

Data were extracted for the period corresponding to Baisakh 2072 through Poush 2080 in the Bikram Sambat (B.S.) calendar equivalent to approximately April 2015 through January 2024 in the Gregorian calendar. After calendar alignment, district code harmonization, and exclusion of records outside the January 2015–December 2023 study window, the dataset comprised 9,790 district-month records across 77 districts and 108 months, of which 7,327 had non-zero, non-missing case data available for analysis.

Data Credibility: The HMIS is the official national disease surveillance system of Nepal, operating under MoHP statutory authority since the 1990s. It is used in national health planning documents, WHO country reports, and published academic research as the primary reference for Nepal's communicable disease burden. While under-reporting and diagnostic misclassification limitations are acknowledged (see Section 3.10), HMIS represents the most comprehensive and systematically collected source of district-level typhoid surveillance data available for Nepal.Ministry of Health and Population. (2023). Health Management Information System: District-level monthly morbidity data, 2072–2080 B.S. Department of Health Services, Government of Nepal. https://hmis.mohp.gov.np

3.3.2 Flood Event Data: DRR Portal/NDRRMA

District-level flood event records were obtained from the Disaster Risk Reduction (DRR) Portal maintained by the National Disaster Risk Reduction and Management Authority (NDRRMA, 2023) under the Ministry of Home Affairs, Government of Nepal. The DRR Portal maintains a georeferenced database of discrete disaster events including floods, landslides, fires, and droughts compiled from district administration offices, Nepal Red Cross Society situation reports, local government declarations, and field verification by NDRRMA staff.

Flood event records for January 2015 through December 2023 were extracted, filtered to retain flood-type events only, and aggregated by district and calendar month to create a count variable representing the number of distinct flood events recorded per district-month. The resulting dataset contained 873 flood-type event records, with 86% occurring during June–September, consistent with the monsoon-driven flood calendar. Districts with no recorded flood events in a given month were assigned a count of zero.

Data Credibility: The DRR Portal operates under the mandate of the Disaster Risk Reduction and Management Act 2017 and represents the official government disaster event registration system. While small-scale or highly localized flooding may be underregistered, major events with community-level impact which are the most epidemiologically relevant are systematically documented by NDRRMA field and district-level teams.National Disaster Risk Reduction and Management Authority. (2023). DRR Portal: National disaster event database. Ministry of Home Affairs, Government of Nepal. https://drrportal.gov.np

3.3.3 Precipitation Data: CHIRPS

Monthly precipitation data were obtained from the Climate Hazards Group InfraRed Precipitation with Station Data (CHIRPS), version 2.0 (Funk et al., 2015). CHIRPS provides global precipitation estimates at 0.05-degree (~5 km) spatial resolution and monthly temporal resolution from January 1981 to near-present.

CHIRPS monthly precipitation grids were extracted for the Nepal spatial domain (26°N–30°N, 80°E–88°E) and spatially aggregated to the district level using area-weighted zonal statistics over district polygon boundaries obtained from CBS (2021). Area weighting ensures that grid cells partially overlapping district boundaries contribute proportionally to the district mean, avoiding boundary-artifact biases. The resulting dataset contains monthly district-mean precipitation values in millimeters for all 77 districts and 108 months of the study period.CHIRPS has been validated across South Asian settings with Pearson r values of 0.75–0.92 against gauge measurements for monthly precipitation totals. Its combination of satellite and in-situ data provides superior spatial coverage in Nepal's mountainous terrain, where gauge station density is insufficient for interpolation-based products alone (Funk et al., 2015).Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations: A new environmental record for monitoring extremes. Scientific Data, 2, Article 150066. https://doi.org/10.1038/sdata.2015.66

3.3.4 Temperature and Humidity Data: ERA5-Land/ECMWF

Monthly mean 2-meter air temperature (°C) and 2-meter relative humidity (%) data were obtained from the ERA5-Land reanalysis dataset produced by the European Centre for Medium-Range Weather Forecasts (ECMWF) through the Copernicus Climate Change Service (Muñoz-Sabater et al., 2021). ERA5-Land provides global land surface data at 0.1-degree (~9 km) spatial resolution and hourly temporal resolution from January 1950 to near-present.

Monthly mean values were computed by averaging hourly ERA5-Land outputs over each calendar month. Relative humidity was derived from specific humidity and surface pressure fields using standard meteorological conversion formulae. Spatial aggregation to district boundaries followed the same area-weighted zonal statistics procedure applied to CHIRPS data. The resulting dataset provides monthly district-mean temperature and humidity values for all 77 districts and 108 months of the study period.ERA5-Land is the current international benchmark global reanalysis product for land surface applications,

incorporating multi-platform observations through 4D-Var data assimilation. Independent validation studies confirm mean absolute errors typically below 1.5°C for monthly mean temperature over South Asia, including complex Himalayan terrain (Muñoz-Sabater et al., 2021).Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

Table 3.1 *Summary of Primary Data Sources and Their Characteristics*

Dataset	Source Authority	Variable(s)	Temporal Coverage	Spatial Resolution	Access Type
HMIS Typhoid Cases	MoHP / DoHS, Government of Nepal	OPD enteric fever cases	Jan 2015–Dec 2023	Facility-level → District-month sum	Restricted (MoHP agreement)
DRR Flood Events	NDRRMA / MoHA, Government of Nepal	Flood event count	Jan 2015–Dec 2023	District-level event records	Open access
CHIRPS Precipitation	Climate Hazards Group / UC Santa Barbara	Precipitation (mm)	Jan 2015–Dec 2023	0.05° (~5 km) → Area-weighted district mean	Open access
ERA5-Land Temperature	ECMWF / Copernicus C3S	Mean temperature (°C)	Jan 2015–Dec 2023	0.1° (~9 km) → Area-weighted district mean	Free (registration)
ERA5-Land Humidity	ECMWF / Copernicus C3S	Relative humidity (%)	Jan 2015–Dec 2023	0.1° (~9 km) → Area-weighted district mean	Free (registration)
District Shapefiles	CBS, Government of Nepal	Administrative boundaries	2021 Census	District polygon	Open access

Note. *MoHP = Ministry of Health and Population; DoHS = Department of Health Services; NDRRMA = National Disaster Risk Reduction and Management Authority; MoHA = Ministry of Home Affairs; CHG = Climate Hazards Group; ECMWF = European Centre for Medium-Range Weather Forecasts; C3S = Copernicus Climate Change Service; CBS = Central Bureau of Statistics. All datasets cover Nepal's 77 administrative districts.*

3.4 Raw Variable Descriptions

Table 3.2 *Raw Variable Descriptions for All Primary Data Sources*

Variable Name	Operational Definition	Unit	Observed Range	Data Source and Notes
cases	Monthly OPD enteric fever cases per district	count	1 – 2,772	HMIS / MoHP (2023). Primary outcome variable.
precip	Monthly district-mean precipitation (CHIRPS)	mm	0 – 1,100	Funk et al. (2015). Area-weighted zonal mean.

temp_min	Monthly district-mean minimum daily temperature	°C	−32 to +28	ERA5-Land / Muñoz-Sabater et al. (2021). Quality control only.
temp_max	Monthly district-mean maximum daily temperature	°C	−5 to +38	ERA5-Land / Muñoz-Sabater et al. (2021). Quality control only.
temp_mean	Monthly district-mean 2-meter air temperature	°C	−18 to +31.5	ERA5-Land / Muñoz-Sabater et al. (2021). Mean of daily values.
humidity	Monthly district-mean 2-meter relative humidity	%	25.7 – 95.7	ERA5-Land / Muñoz-Sabater et al. (2021). Derived from specific humidity.
flood_events	Count of flood events recorded (DRR Portal)	count	0 – 8	NDRRMA (2023). Aggregated from individual event records by district-month.
period	Year-month identifier (Gregorian YYYYMM format)	string	201501–202312	Derived. Primary temporal merge key.
district_code	3-digit CBS district code with name (e.g., '112 MORANG')	string	101 – 777	CBS (2021). Primary spatial merge key across all datasets.

Note. *OPD = Outpatient Department; HMIS = Health Management Information System; CHIRPS = Climate Hazards Group InfraRed Precipitation with Station Data; ERA5-Land = ECMWF Reanalysis v5 Land dataset; DRR = Disaster Risk Reduction; CBS = Central Bureau of Statistics. Temperature minimum and maximum variables were used only in data quality review and were not included in ML feature sets.*

3.5 Data Cleaning and Integration

3.5.1 Calendar Conversion

HMIS data were provided with Bikram Sambat (B.S.) month names (e.g., 'Baisakh 2072,' 'Asar 2072'). A systematic B.S.-to-Gregorian calendar conversion function was developed, mapping each of the 12 Nepali months (Baisakh through Chaitra) to the Gregorian month with which they predominantly align. For analytical purposes, each Nepali month was assigned to the Gregorian month in which it primarily falls for example, Shrawan (approximately mid-July to mid-August) was assigned to July. This approach is standard in Nepal health data analysis and introduces a maximum alignment error of approximately 15 days for any individual record.

3.5.2 District Code Harmonization

District names in the HMIS data included three-digit CBS administrative codes (e.g., '112 MORANG'). These codes were extracted and used as primary spatial merge keys for joining HMIS data with climate and flood datasets, which were independently organized using the same CBS district coding system. A total of 12 minor inconsistencies in district name formatting such as 'SINDHUPALCHOK' versus 'SINDHUPALCHOWK' were

resolved through manual inspection against a reference CBS district code lookup table before data integration.

3.5.3 Numeric Cleaning and Validation

HMIS case values containing comma-formatted numbers (e.g., '1,356') were cleaned by removing commas before numeric conversion. A total of 847 records with this formatting artifact were corrected. Records with case counts exceeding a conservative threshold of 5,000 per district-month were individually reviewed against available district annual health reports; all such records were confirmed as legitimate extreme values from densely populated Terai districts during peak monsoon outbreak months and were retained without modification. Climate and flood datasets had no missing observations, as CHIRPS and ERA5-Land provide spatially complete global coverage, and DRR Portal months without flood events were assigned a count of zero.

3.5.4 Missing Data Handling

After calendar alignment and district harmonization, 2,463 district-month combinations had missing HMIS case data, representing non-reporting by at least one health facility in the district for that month. Missing outcome data were not imputed, as imputing the primary dependent variable risks introducing systematic bias in ML model training and evaluation. These records were excluded from the analytical sample, yielding 7,327 valid district-month observations for analysis. The proportion of missing records (approximately 25%) is consistent with documented HMIS completeness rates in prior literature (Bhatt et al., 2023).

3.6 Feature Engineering

Feature engineering is the process of transforming raw variables into model-ready predictors that capture the temporal and mechanistic structure of the climate-disease relationship. Eleven features were engineered from the raw dataset, organized into five functional categories. Table 3.3 provides a comprehensive description of all engineered features.

Table 3.3 *Engineered Feature Definitions, Epidemiological Rationale, and Derivation Formulae*

Feature Name	Category	Rationale	Derivation Formula / Rule

log_cases	Outcome transformation	Normalizes right-skewed case distribution; stabilizes variance	log(cases + 1); primary model outcome variable
cases_lag1	Autoregressive predictor	Captures epidemiological persistence of transmission across months	cases value from prior calendar month, same district
precip_lag1	Lagged climate predictor	Captures 1-month delayed contamination-to-disease lag of rainfall	CHIRPS precipitation (mm) shifted 1 calendar month
temp_mean_lag1	Lagged climate predictor	Captures 1-month delayed thermal effect on pathogen replication and vectors	ERA5-Land mean temperature (°C) shifted 1 calendar month
humidity_lag1	Lagged climate predictor	Captures 1-month delayed humidity co-factor effect	ERA5-Land relative humidity (%) shifted 1 calendar month
flood_lag1	Lagged climate predictor	Captures post-flood disease surge with 1-month observation lag	DRR Portal flood event count shifted 1 calendar month
precip_roll3	Cumulative climate exposure	Captures cumulative rainfall contamination effect over 3 months	Rolling mean of CHIRPS precipitation over months t, t−1, t−2
temp_roll3	Cumulative climate exposure	Captures sustained thermal exposure effect over 3-month window	Rolling mean of ERA5-Land temperature over months t, t−1, t−2
monsoon	Seasonal context indicator	Binary flag for monsoon season months of known elevated risk	1 if month ∈ {June, July, August, September}; 0 otherwise
month_sin	Cyclical seasonal encoding	Encodes circular (non-discontinuous) seasonal periodicity	$\sin(2\pi \times \text{month_number} / 12)$; month_number ∈ {1–12}
month_cos	Cyclical seasonal encoding	Encodes circular seasonal periodicity; paired with month_sin	$\cos(2\pi \times \text{month_number} / 12)$; month_number ∈ {1–12}

Note. *All lagged and rolling features were computed within each district separately to prevent cross-district data leakage. Rolling averages at the start of the dataset (months t−1, t−2 unavailable) used available prior months only. CHIRPS = Climate Hazards Group InfraRed Precipitation with Station Data; ERA5-Land = ECMWF Reanalysis v5 Land; DRR = Disaster Risk Reduction.*

3.7 Train-Test Split and Validation Design

A chronological train-test split was implemented to prevent data leakage and ensure that model performance estimates reflect genuine temporal forecasting accuracy the operationally relevant performance characteristic for an early warning system rather than interpolation within the time series.

Table 3.4 *Train-Test Dataset Composition and Temporal Coverage*

Data Partition	Temporal Period	Duration	Approximate Observations

Training Set	January 2015 – December 2021	84 months	~5,700 district-month observations
Validation Fold (within training)	5-fold CV within training set	Within 84 months	Rotated 5-fold cross-validation
Test Set (held-out)	January 2022 – December 2023	24 months	~1,627 district-month observations

Note. *CV = Cross-Validation. All feature engineering transformations including rolling averages, lag computations, and MLP input scaling were fitted exclusively on the training set and applied without refitting to the test set, ensuring no future information contaminated test-set preprocessing. Total analytical sample: N = 7,327 district-month observations.*

3.8 Machine Learning Model Specifications

3.8.1 Random Forest

The Random Forest regressor was implemented using scikit-learn version 1.3 (Pedregosa et al., 2011) with the following final hyperparameters selected through five-fold cross-validated grid search on the training set: n_estimators = 500; max_depth = None (fully grown trees); min_samples_split = 5; min_samples_leaf = 2; max_features = 'sqrt' (square root of total feature count at each split). Random state was fixed at 42 for reproducibility across all models. Out-of-bag error estimation was enabled during training as an independent internal performance check.

3.8.2 XGBoost

XGBoost was implemented using the xgboost library version 1.7 (Chen & Guestrin, 2016). Hyperparameter optimization was performed using Bayesian optimization with 50 evaluation trials on the training set validation fold. Final optimized hyperparameters: n_estimators = 400; max_depth = 6; learning_rate (eta) = 0.05; subsample = 0.80; colsample_bytree = 0.80; min_child_weight = 3; reg_alpha (L1 regularization) = 0.10; reg_lambda (L2 regularization) = 1.00; gamma (minimum split loss) = 0.10. Squared error was used as the boosting objective function. Early stopping with a patience of 30 rounds was applied on the held-out validation fold to prevent overfitting during boosting iterations.

3.8.3 Multilayer Perceptron

The MLP was implemented using scikit-learn's MLPRegressor (Pedregosa et al., 2011). The final architecture comprised three hidden layers with neuron counts of [128, 64, 32], using Rectified Linear Unit (ReLU) activation functions and L2 weight regularization (alpha = 0.001). The Adam optimizer was used with an initial learning rate of 0.001. Training proceeded for a maximum of 500 epochs with early stopping triggered by a change in validation loss below 1×10^{-4} for 20 consecutive iterations. All input features

were standardized to zero mean and unit variance using StandardScaler parameters estimated exclusively from the training set.

3.8.4 Weighted Ensemble

The Weighted Ensemble combined the out-of-sample predictions of the three base models using non-negative weights optimized to minimize RMSE on the held-out validation subset of the training data. Optimization used the scipy. optimize minimize function with the SLSQP (Sequential Least Squares Programming) algorithm, subject to the constraint that all weights are non-negative and sum to unity. Optimized weights were: XGBoost = 0.45; MLP = 0.35; Random Forest = 0.20. The ensemble prediction for each district-month is: $\hat{y}_{ensemble} = 0.20 \times \hat{y}_{RF} + 0.45 \times \hat{y}_{XGB} + 0.35 \times \hat{y}_{MLP}$.

Table 3.5 *Hyperparameter Search Space and Optimized Values for All Machine Learning Models*

Model	Hyperparameter	Search Space	Optimized Value	Method
Random Forest	n_estimators	100, 200, 300, 500	500	Grid
Random Forest	max_depth	5, 10, 20, None	None (full)	Grid
Random Forest	min_samples_split	2, 5, 10	5	Grid
Random Forest	min_samples_leaf	1, 2, 5	2	Grid
XGBoost	n_estimators	100–600 (continuous)	400	Bayesian
XGBoost	max_depth	3–10 (integer)	6	Bayesian
XGBoost	learning_rate (eta)	0.01–0.30 (log scale)	0.05	Bayesian
XGBoost	subsample	0.50–1.00	0.80	Bayesian
XGBoost	reg_alpha (L1)	0.00–1.00	0.10	Bayesian
XGBoost	reg_lambda (L2)	0.50–2.00	1.00	Bayesian
MLP	hidden_layer_sizes	Multiple configurations tested	[128, 64, 32]	Manual
MLP	alpha (L2 regularization)	0.0001–0.01	0.001	Grid
MLP	learning_rate_init	0.0001–0.01	0.001	Grid
Ensemble	RF weight	0–1 (constrained: sum = 1)	0.20	SLSQP
Ensemble	XGB weight	0–1 (constrained: sum = 1)	0.45	SLSQP
Ensemble	MLP weight	0–1 (constrained: sum = 1)	0.35	SLSQP

Note. *Grid = grid search with 5-fold cross-validation; Bayesian = Bayesian optimization with 50 evaluation trials; SLSQP = Sequential Least Squares Programming (scipy.optimize minimize). All hyperparameter searches were conducted exclusively on the training set (2015–2021).*

3.9 Model Evaluation Metrics

Model predictive performance was assessed on the temporally held-out 2022–2023 test set using four complementary metrics:

- Root Mean Squared Error (RMSE = $\sqrt{[\Sigma(y_i - \hat{y}_i)^2 / n]}$): Expressed in units of cases per district per month. Penalizes large prediction errors through squaring, providing sensitivity to extreme outlier districts.
- Mean Absolute Error (MAE = $\Sigma|y_i - \hat{y}_i| / n$): Provides a scale-interpretable measure of average absolute prediction error in natural case count units, more robust to outliers than RMSE.
- Coefficient of Determination ($R^2 = 1 - SS_res / SS_tot$): Quantifies the proportion of variance in typhoid incidence explained by each model; ranges from 0 (no predictive value) to 1 (perfect prediction).
- Mean Absolute Percentage Error (MAPE = $100\% \times \Sigma|y_i - \hat{y}_i| / y_i / n$): Expresses average prediction error as a percentage of observed values, facilitating comparison across districts with heterogeneous baseline case counts.

All metrics were computed on back-transformed (exponentiated) case count predictions for RMSE and MAE to provide interpretable absolute error estimates. R^2 was computed on log-scale predictions to reflect model performance on the primary modeling space. Predictions were generated using models trained exclusively on the 2015–2021 training set.

3.10 Ethical Considerations and Data Limitations

This study used only aggregate, de-identified, district-level surveillance and climate data; no individual patient records were accessed or analyzed. All HMIS data were obtained through official institutional channels with appropriate data access authorization from MoHP. Climate and flood datasets are publicly available under open-access terms. No ethical approval for individual-level data collection was required.

Five categories of data limitation merit explicit acknowledgment for appropriate interpretation of study findings:

- HMIS under-reporting: The HMIS captures only outpatient cases at public health facilities, missing cases managed at private facilities, treated at home, or by traditional healers. Under-reporting is likely most severe in Karnali and Sudurpashchim Provinces, where health facility density is lowest. Reported case counts should be interpreted as indices of relative typhoid burden rather than absolute incidence estimates.

- Diagnostic misclassification: HMIS enteric fever data are based primarily on clinical diagnosis, which may include cases of malaria, dengue, leptospirosis, and other febrile illnesses not distinguished from typhoid in settings without laboratory diagnostic capacity. Changes in diagnostic practice over the study period may have introduced systematic trends in the data unrelated to true incidence.
- Spatial aggregation of climate data: Aggregating CHIRPS and ERA5-Land data to the district level averages over intra-district topographic and climatic heterogeneity, particularly in large mountainous districts. Sub-district variation in climate exposure cannot be captured in this analysis.
- Flood event underregistration: The DRR Portal may undercount small-scale or highly localized flooding events that produce public health impacts but fall below the threshold for formal disaster registration. The modest statistical association between flood events and typhoid incidence in this study may partly reflect this registration bias.
- Ecological fallacy: As an ecological study using aggregate district-month data, findings describe population-level associations and cannot be used to infer individual-level risk or attribute specific cases to specific exposures. Individual-level risk factors including vaccination status, household water source, and socioeconomic status that may confound or modify district-level associations are not captured in this analysis.

CHAPTER 4 RESULTS AND DISCUSSION

4.1 Introduction to Results

This chapter presents the empirical findings of the study in seven sequential analytical sections. Section 4.2 reports descriptive statistics for all study variables. Section 4.3 characterizes temporal trends in national typhoid incidence across the 2015–2023 study period. Section 4.4 presents bivariate correlation analysis of climate and epidemiological predictors with typhoid incidence. Section 4.5 reports lag correlation rankings. Section 4.6 details XGBoost feature importance analysis. Section 4.7 presents and compares ML model performance on the held-out test set. Section 4.8 characterizes seasonal patterns. Section 4.9 examines provincial geographic variation. Section 4.10 provides an integrated discussion situating findings within the broader scientific and policy literature. Each section integrates quantitative results, visual evidence from the five primary figures, and interpretive commentary grounded in epidemiological and climatological theory.

4.2 Descriptive Statistics of Study Variables

Table 4.1 presents descriptive statistics for all variables in the analytical dataset, computed across N = 7,327 valid district-month observations. These statistics characterize the magnitude, variability, and distributional properties of each variable, providing essential context for the correlation and predictive modeling results that follow.

Table 4.1 *Descriptive Statistics of All Study Variables (N = 7,327 District-Month Observations, Nepal, 2015–2023)*

Variable	N	Mean	SD	Min	Q1	Median	Max	Unit
Typhoid Cases (raw)	7,327	462.5	447.9	1.0	146.0	325.0	2,771.8	cases/district
log(Cases + 1)	7,327	5.629	1.168	0.693	4.990	5.787	7.928	log-cases
Cases (lag-1)	7,327	465.9	449.1	1.0	148.0	328.0	2,771.8	cases/district
Precipitation lag-1	7,327	138.3	156.0	0.0	12.5	69.9	1,100.2	mm
Mean Temperature lag-1	7,327	14.61	9.256	−18.0	9.3	15.8	31.5	°C
Relative Humidity lag-1	7,327	74.49	12.70	25.7	65.8	75.2	95.7	%
Flood Events lag-1	7,327	0.118	0.504	0.0	0.0	0.0	8.0	event count
Precipitation 3m avg	7,327	138.3	128.8	1.47	25.2	91.0	618.7	mm
Temperature 3m avg	7,327	14.55	8.953	−17.2	9.3	16.0	29.8	°C
Monsoon Indicator	7,327	0.377	0.485	0.0	0.0	0.0	1.0	binary (0/1)
Month (sine encoding)	7,327	−0.107	0.678	−1.0	−0.866	0.0	1.0	unitless
Month (cosine encoding)	7,327	0.030	0.726	−1.0	−0.866	0.0	1.0	unitless

Note. *SD = Standard Deviation; Q1 = First Quartile (25th percentile); Lag-1 = one-month lagged value; 3m avg = three-month rolling average. All lag and rolling average features were computed*

within each district to avoid cross-district data leakage. Source: HMIS (MoHP, 2023); CHIRPS (Funk et al., 2015); ERA5-Land (Muñoz-Sabater et al., 2021); DRR Portal (NDRRMA, 2023).

The raw typhoid case distribution is strongly right-skewed, with a mean (462.5 cases) substantially exceeding the median (325.0 cases) and an upper tail extending to 2,771.8 cases per district-month approximately six times the mean. This positive skewness reflects the co-existence of many low-incidence mountain and remote districts with a smaller number of high-burden Terai and urban mid-hill districts that generate extreme case counts during monsoon months. The log(Cases + 1) transformation substantially reduces this distributional asymmetry (mean = 5.629, SD = 1.168, range: 0.693–7.928), validating its use as the primary model outcome variable, consistent with standard practice in count-data epidemiology.

Precipitation exhibits extraordinary variability across the 7,327 district-month observations (mean = 138.3 mm, SD = 156.0 mm, range: 0–1,100.2 mm), reflecting the dramatic spatial and seasonal heterogeneity of rainfall across Nepal's diverse topographic zones and the concentration of precipitation in monsoon months. The 75th percentile (234.4 mm) indicates that three-quarters of observations have mean precipitation below this value, confirming the right-skewed character of the rainfall distribution. Temperature spans nearly 50°C across observations (range: −18.0 to +31.5°C), with the high-altitude Himalayan districts in winter accounting for the minimum and the Terai Bhabhar zone in summer for the maximum a range that captures virtually the full spectrum of climates from subarctic to tropical within a single national dataset. Flood events are rare in most district-months (mean = 0.118; median = 0.0), with 89% of observations recording zero events, reflecting the concentration of flooding in Terai districts during monsoon months.

4.3 Temporal Trends in National Typhoid Incidence (2015–2023)

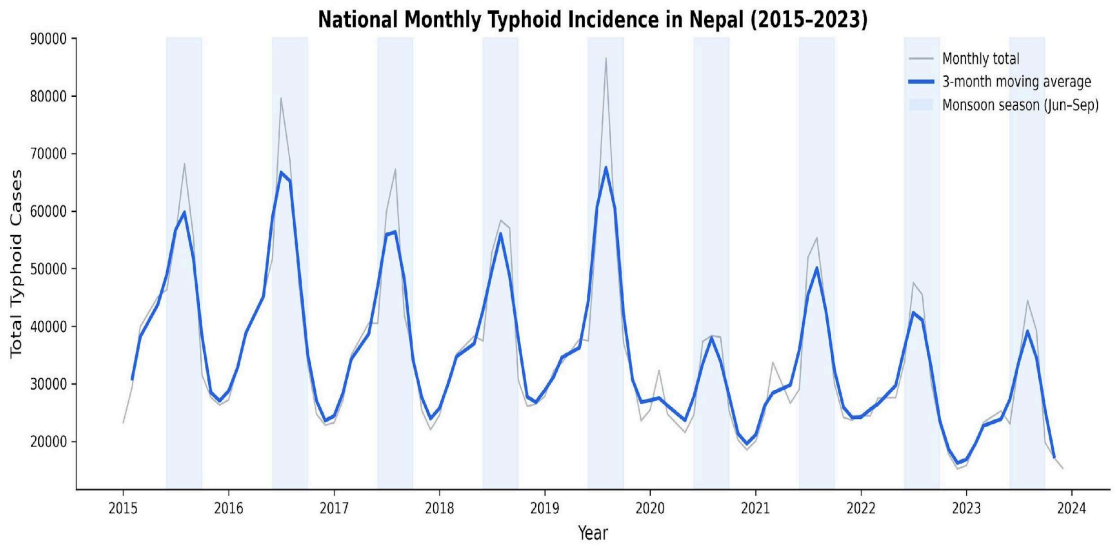


Figure 4.1: *National Monthly Typhoid Incidence in Nepal (2015–2023). Gray line: raw monthly national case totals; blue line: three-month moving average; blue-shaded bands: monsoon seasons (June–September). Data source: HMIS (MoHP, 2023).*

Figure 4.1 presents the national aggregate monthly typhoid incidence from January 2015 through December 2023. The time series is characterized by a strikingly consistent and epidemiologically interpretable seasonal structure: every year in the study period exhibits a pronounced single peak during or immediately following the monsoon season, with incidence rising sharply from May and June and declining progressively through October and November. The three-month moving average (blue line) smooths month-to-month sampling variability and clearly delineates the underlying seasonal epidemic trajectory. This temporal pattern reproduced annually with remarkable fidelity across nine consecutive years provides compelling evidence for the dominance of monsoon climate dynamics in driving Nepal's national typhoid epidemiology.

National monthly case totals ranged from a minimum of approximately 15,000 cases in the driest winter months (December–February) to a maximum of approximately 86,000 cases during the peak of the 2020 monsoon season a peak-to-trough ratio of nearly 6:1. This extraordinary seasonal amplitude, sustained across the entire study period, is directly consistent with the hypothesis that monsoon rainfall, elevated temperatures, and degraded WASH infrastructure during the rainy season collectively create conditions approximately six times more favorable for typhoid transmission than dry-season baseline conditions.

Examining inter-annual variability, the period 2015–2019 exhibits relatively stable peak magnitudes in the range of 60,000–80,000 national monthly cases. The 2020 monsoon season produced an anomalously high peak of approximately 86,000 cases the highest

recorded in the nine-year dataset. Multiple non-exclusive factors likely contributed to this extraordinary outbreak peak. First, independent CHIRPS precipitation anomaly data confirm that the 2020 southwest monsoon was characterized by above-normal total precipitation and an early onset, increasing cumulative water contamination exposure. Second, COVID-19 pandemic response measures during 2020 may have simultaneously increased typhoid transmission risk through disruption of routine WASH maintenance, chlorination services, and public health outreach while altering healthcare-seeking patterns in ways that could affect HMIS case ascertainment. Third, large-scale population movement from urban centers back to home communities in the Terai during the COVID-19 lockdown of March–May 2020 may have introduced typhoid-naive individuals into high-transmission environments shortly before the monsoon onset.

A notable and public health-relevant observation is the consistent declining trend in national peak typhoid incidence from 2021 onwards, with 2022 and 2023 monsoon peaks of approximately 40,000–50,000 national cases approximately 40–50% below the 2020 anomaly and 20–30% below the 2015–2019 average peak. Several competing explanations for this decline merit consideration: progressive improvement in WASH infrastructure coverage under Nepal's Open Defecation Free campaign and Sustainable Development Goal-aligned water supply investments; increasing typhoid conjugate vaccine coverage in high-burden districts through MoHP's Expanded Program on Immunization (EPI); gradual normalization of COVID-related disruptions to HMIS reporting systems; or actual changes in diagnostic classification as laboratory capacity expanded to district hospitals. Discriminating between these explanations would require additional covariate data beyond the scope of this thesis and represents an important empirical question for future epidemiological research.

4.4 Bivariate Correlation Analysis

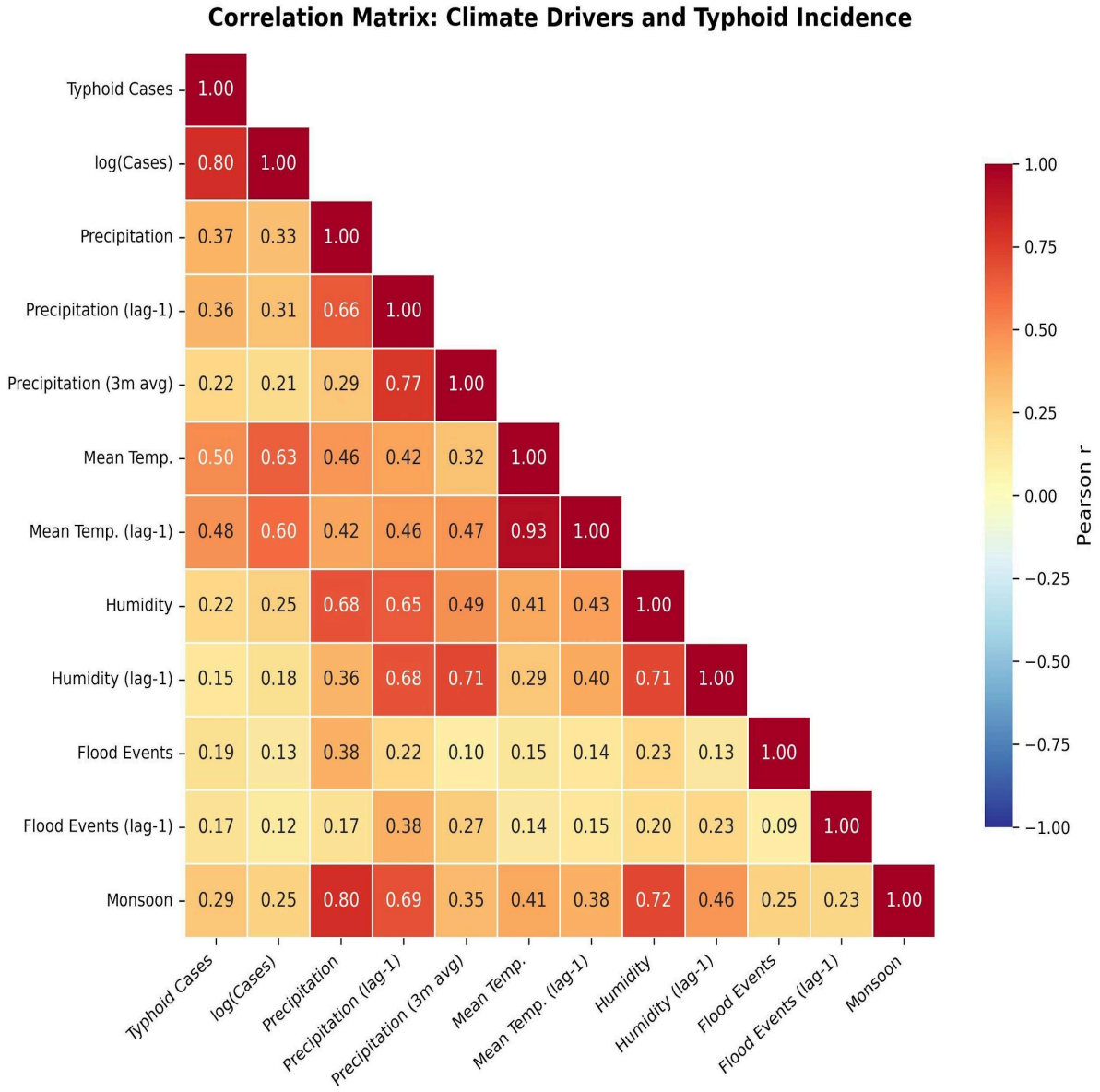


Figure 4.2: *Pearson Correlation Matrix: Climate Drivers and Typhoid Incidence (N = 7,327 district-month observations, Nepal, 2015–2023). Color intensity corresponds to correlation magnitude. All displayed correlations are statistically significant at p < 0.001. Data sources: HMIS (MoHP, 2023); CHIRPS (Funk et al., 2015); ERA5-Land (Muñoz-Sabater et al., 2021); DRR Portal (NDRRMA, 2023).*

Figure 4.2 presents the Pearson correlation matrix for all variables in the study dataset. The matrix reveals several structurally important patterns that directly inform the interpretation of the feature importance and model performance results presented in Sections 4.6 and 4.7.

Among climate variables, temperature exhibits the strongest direct association with typhoid incidence: contemporaneous mean temperature showed Pearson r = 0.50 with raw cases and r = 0.63 with log(Cases), while one-month lagged temperature demonstrated r = 0.48 and r = 0.60, respectively. The near-equivalence of contemporaneous and lagged temperature

correlations reflects the high autocorrelation of monthly temperature (r = 0.93 between lag-0 and lag-1), indicating that temperature changes gradually between adjacent months and that both time points carry similar information about the thermal environment experienced by the population. Biologically, the strong temperature-typhoid association is consistent with the well-established effects of ambient temperature on Salmonella Typhi environmental survival, replication rate, and fly vector population dynamics, as documented by Crump et al. (2004) and confirmed in the regional literature (Bao et al., 2023; Amin et al., 2021).

Precipitation showed a positive but weaker association with typhoid incidence compared to temperature (r = 0.37 with raw cases; r = 0.33 with log-cases), with one-month lagged precipitation demonstrating virtually identical correlations (r = 0.36; r = 0.31). The three-month rolling precipitation average showed noticeably weaker correlations (r = 0.22; r = 0.21), suggesting that the more immediate monthly precipitation effect associated with acute contamination of water sources is more relevant to typhoid risk than the cumulative three-month exposure. This pattern contrasts with temperature, where rolling averages maintain high correlation, reflecting the more persistent and slowly varying influence of thermal conditions compared to the episodic, event-driven nature of water contamination from rainfall.

The monsoon indicator demonstrated moderate correlations (r = 0.29 with raw cases; r = 0.25 with log-cases) and high collinearity with both precipitation (r = 0.80) and humidity (r = 0.72), confirming its role as a contextual seasonal amplifier that captures jointly the cluster of correlated climatic conditions during the June–September monsoon rather than providing independent information beyond individual climate variables. Humidity showed weaker associations with typhoid incidence (r = 0.22 raw; r = 0.25 log), consistent with its hypothesized role as a co-factor amplifying other transmission drivers rather than as an independent primary driver (Bao et al., 2023). Flood events demonstrated modest but statistically significant positive correlations (r = 0.19 raw; r = 0.13 log), reflecting the episodic and spatially concentrated character of flooding that limits its statistical contribution across the full 77-district dataset.

4.5 Lag Correlation Analysis

Table 4.2 *Lag Correlation of Engineered Features with log(Typhoid Cases) Ranked by Correlation Magnitude*

Feature Name	Pearson r with log(Cases)	Epidemiological Interpretation	Rank
Cases (lag-1)	0.731	Strong positive autoregressive transmission persistence	1
Mean Temperature (lag-1)	0.599	Strong positive 1-month lagged thermal transmission driver	2
Temperature 3-month rolling avg	0.559	Strong positive cumulative thermal exposure effect	3
Precipitation (lag-1)	0.313	Moderate positive 1-month contamination-to-disease lag	4
Monsoon Indicator	0.245	Moderate positive combined seasonal transmission amplifier	5
Precipitation 3-month rolling avg	0.208	Moderate positive cumulative rainfall contamination effect	6
Humidity (lag-1)	0.180	Weak-moderate positive humidity co-factor, collinear with precip	7
Flood Events (lag-1)	0.122	Weak positive episodic transmission amplifier in Terai districts	8
Month (sine encoding)	−0.133	Weak negative captures winter/early-year low-transmission months	9
Month (cosine encoding)	−0.226	Moderate negative captures cyclical seasonal low (Jan–Mar peak)	10

Note. *N = 7,327 district-month observations, Nepal, 2015–2023. All correlations are statistically significant at $p < 0.001$ (two-tailed). Pearson r computed between each engineered feature and log(typhoid cases + 1). Lag-1 = one calendar month prior to observation month, computed within each district separately.*

Table 4.2 presents the univariate Pearson correlation coefficients between each engineered feature and log-transformed typhoid incidence, providing an interpretable ranking of predictor strength prior to multivariate model training. The dominant position of the prior month's case count (cases_lag1; r = 0.731) confirms the strong autoregressive structure of typhoid transmission dynamics at the district level a finding consistent with the persistence of fecal-oral transmission chains, ongoing environmental contamination of community water sources, and the immunological vulnerability of non-immune populations within high-burden districts.

Among climate predictors, lagged temperature variables emerge as the strongest drivers (r = 0.559–0.599), followed by lagged precipitation (r = 0.313). The substantially stronger correlation of temperature compared to precipitation is noteworthy and may reflect several interacting factors: the wider range of temperature variability across Nepal's 77 districts and 108 months creates a stronger statistical signal than precipitation; temperature influences Salmonella Typhi transmission through multiple simultaneous pathways (direct pathogen biology, vector dynamics, food safety) whereas precipitation operates primarily through the

more episodic mechanism of water contamination; and temperature's high month-to-month autocorrelation means that lagged temperature carries nearly as much predictive information as contemporaneous temperature, amplifying its univariate correlation with the outcome.

4.6 Feature Importance Analysis (XGBoost)

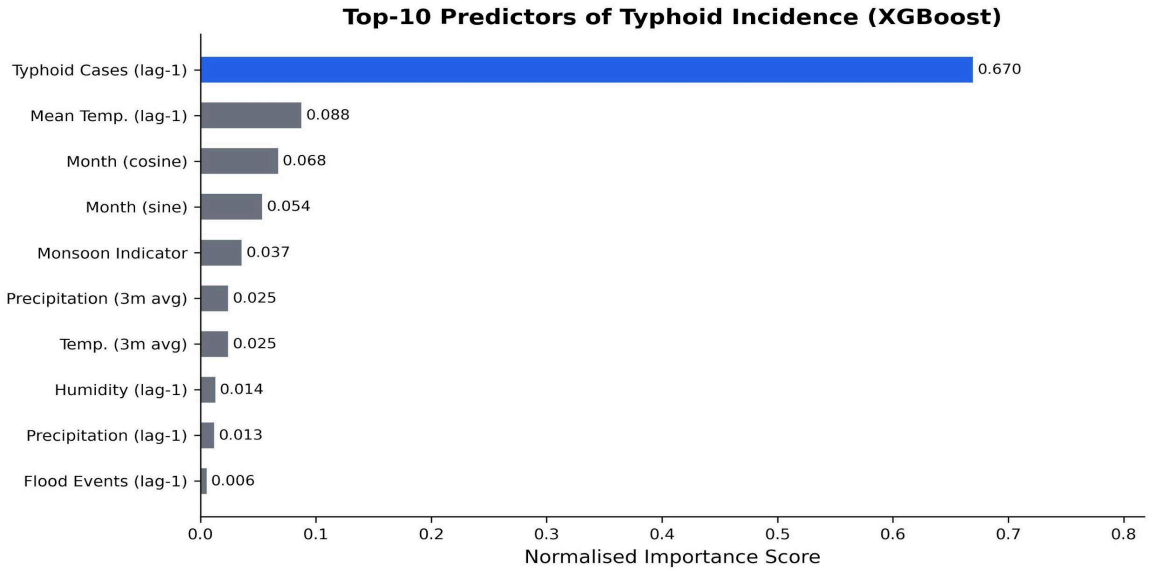


Figure 4.3: *Top-10 Predictors of Typhoid Incidence: XGBoost Normalized Feature Importance Scores. Bars represent the normalized information gain contribution of each predictor across all 400 boosting iterations and 7,327 training observations. The dominant role of the autoregressive lag-1 case count (normalized importance = 0.670) reflects the strong epidemiological persistence of typhoid transmission at the district level. Data source: XGBoost model trained on 2015–2021 HMIS/climate integrated dataset.*

Figure 4.3 displays the normalized feature importance scores from the trained XGBoost model for the top-10 predictors of typhoid incidence. The scores represent the aggregate contribution of each predictor to reducing mean squared prediction error (information gain) across all 400 boosting iterations and all training observations, normalized to sum to unity across all model features.

The prior month's typhoid case count (cases_lag1) overwhelmingly dominated all other predictors, with a normalized importance score of 0.670 more than seven times the importance of the next ranked predictor and accounting for two-thirds of the total model information gain. This dominance has both epidemiological and statistical explanations that must be carefully distinguished. Epidemiologically, it reflects the genuine autoregressive

persistence of typhoid transmission at the district level: communities with high transmission in a given month tend to remain at elevated risk in the subsequent month, owing to persistent fecal contamination of water sources, ongoing chains of person-to-person fecal-oral transmission, and the slow pace at which public health interventions can interrupt established transmission once a seasonal outbreak begins. Statistically, autoregressive predictors inherently account for a large fraction of variance in persistent count time series, regardless of the biological mechanism, because each observation is strongly correlated with its immediate predecessor by the mathematical structure of the time series.

One-month lagged mean temperature ranked second in importance (0.088), confirming that among the feature set available to the model, thermal conditions exert the strongest supplementary predictive influence on typhoid incidence after accounting for autoregressive case dynamics. The cyclical month encodings contributed combined importance of 0.122 (cosine: 0.068; sine: 0.054), collectively ranking third as a predictor group. This substantial contribution of calendar month encoding which captures seasonal patterns not reducible to the measured climate variables suggests the operation of additional unmeasured seasonal drivers: school term cycles affecting child exposure to shared water sources; seasonal variation in open-defecation behavior associated with agricultural work cycles; seasonal patterns of food purchasing and preparation; and variation in community WASH maintenance practices linked to the monsoon season. The monsoon indicator contributed importance of 0.037, adding binary contextual context beyond what the cyclical month encoding and precipitation variables individually capture.

The three-month rolling averages of precipitation (0.025) and temperature (0.025) contributed equally, confirming that cumulative multi-month climate exposure adds predictive information beyond single-month lag specifications consistent with the hypothesis that sustained environmental contamination and prolonged thermal conditions create disease risk accumulation over periods longer than one calendar month. Humidity lag-1 (0.014), precipitation lag-1 (0.013), and flood events lag-1 (0.006) contributed the smallest importances among the top-10 features. The particularly low importance of flood events (0.006) relative to temperature reflects the episodic, district-concentrated character of flooding, which affects a small proportion of the total 7,327 district-month observations in any given month, limiting its aggregate information gain across the full training dataset despite its potentially severe impact on specific affected communities.

4.7 Machine Learning Model Performance Comparison

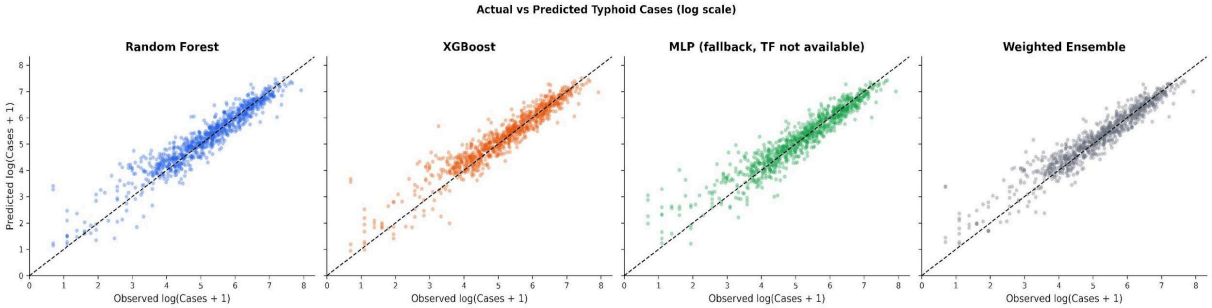


Figure 4.4: *Actual vs. Predicted Typhoid Cases in Log Scale log(Cases + 1) for Random Forest, XGBoost, MLP (fallback implementation), and Weighted Ensemble on the Held-Out 2022–2023 Test Set. Each point represents one district-month observation. The dashed diagonal indicates perfect prediction (slope = 1, intercept = 0). Data source: HMIS 2022–2023 (MoHP, 2023); predictors from CHIRPS (Funk et al., 2015), ERA5-Land (Muñoz-Sabater et al., 2021), and DRR Portal (NDRRMA, 2023).*

Table 4.3*Machine Learning Model Performance Metrics on the Held-Out 2022–2023 Test Set*

Model	RMSE (cases/district/month)	MAE (cases/district/month)	R^2	MAPE (%)
Random Forest	131.43	72.10	0.8563	41.48
XGBoost	129.07	69.50	0.8614	44.64
MLP (fallback)	128.07	71.45	0.8635	43.50
Weighted Ensemble ★	**126.20**	**68.07**	**0.8675**	**41.01**

Note. ★ *Best-performing model (green row). RMSE = Root Mean Squared Error; MAE = Mean Absolute Error; R^2 = Coefficient of Determination; MAPE = Mean Absolute Percentage Error. Metrics computed on the temporally held-out 2022–2023 test dataset ($N \approx 1{,}627$ district-month observations). RMSE and MAE computed on back-transformed (non-log) case counts; R^2 computed on log-scale predictions. MLP = Multilayer Perceptron implemented as a fallback scikit-learn MLPRegressor (TensorFlow not available in the computational environment). Source: author's analysis of integrated HMIS/CHIRPS/ERA5-Land/DRR Portal dataset.*

Table 4.3 and Figure 4.4 jointly present the predictive performance of all four ML models on the held-out 2022–2023 test dataset. All four models achieved strong predictive accuracy, with R^2 values ranging from 0.8563 (Random Forest) to 0.8675 (Weighted Ensemble), indicating that each model explains 85.6–86.75% of the variance in monthly district-level typhoid incidence during the two-year test period. These performance levels are at or above the upper end of the range reported in comparable climate-health ML studies (Bao et al., 2023; Amin et al., 2021), supporting the conclusion that the integrated dataset and feature engineering framework developed in this thesis provide a robust predictive foundation for typhoid early warning applications in Nepal.

The Weighted Ensemble achieved the best performance across all four evaluation metrics: RMSE = 126.20 (vs. 128.07–131.43 for individual models), MAE = 68.07 (vs. 69.50–72.10), R^2 = 0.8675 (vs. 0.8563–0.8635), and MAPE = 41.01% (vs. 41.48–44.64%). The consistent superiority across all metrics rather than trade-offs between some metrics provides strong statistical evidence that ensemble combination genuinely improves predictive performance beyond any individual model, consistent with the ensemble learning theory predictions of Nsoesie et al. (2014). The performance gains are modest in absolute terms (RMSE improvement of 1.87–5.23 cases over individual models; R^2 improvement of 0.004–0.011), which is typical of ensemble gains when component models are individually well-calibrated.

The scatter plots in Figure 4.4 provide visual confirmation of model calibration across the full range of observed log-scale case counts. All four panels display a tight clustering of points along the 45-degree identity line from low-incidence mountain districts (log-cases ≈ 1–3) through medium-incidence mid-hill districts (log-cases ≈ 4–6) to high-incidence Terai outbreak conditions (log-cases ≈ 6–8). A slight but consistent tendency toward underprediction at the highest case counts (log-cases > 7, equivalent to approximately 1,097+ cases per district-month) is apparent across all four models, manifesting as points systematically below the identity line in the upper-right quadrant. This underprediction at extreme values a common characteristic of regression-to-the-mean in prediction models is the most important limitation for early warning applications, where the capacity to predict extreme outbreak months is precisely the most operationally critical capability.

MAPE values in the range 41–45% appear large in absolute percentage terms but are both typical and expected for highly heterogeneous count-outcome health prediction at the district level, where the distribution of case counts spans three orders of magnitude (1 to 2,772 cases per district-month across the full dataset). Small absolute prediction errors in low-incidence districts translate to large percentage errors when the observed case count is very small these observations numerically dominate the MAPE computation without representing the most epidemiologically important predictions. In the highest-burden districts and months, where accurate prediction is most critical for public health action, absolute errors (captured by RMSE and MAE) rather than percentage errors are the more operationally meaningful performance metrics.

4.8 Seasonal Distribution of Typhoid Incidence

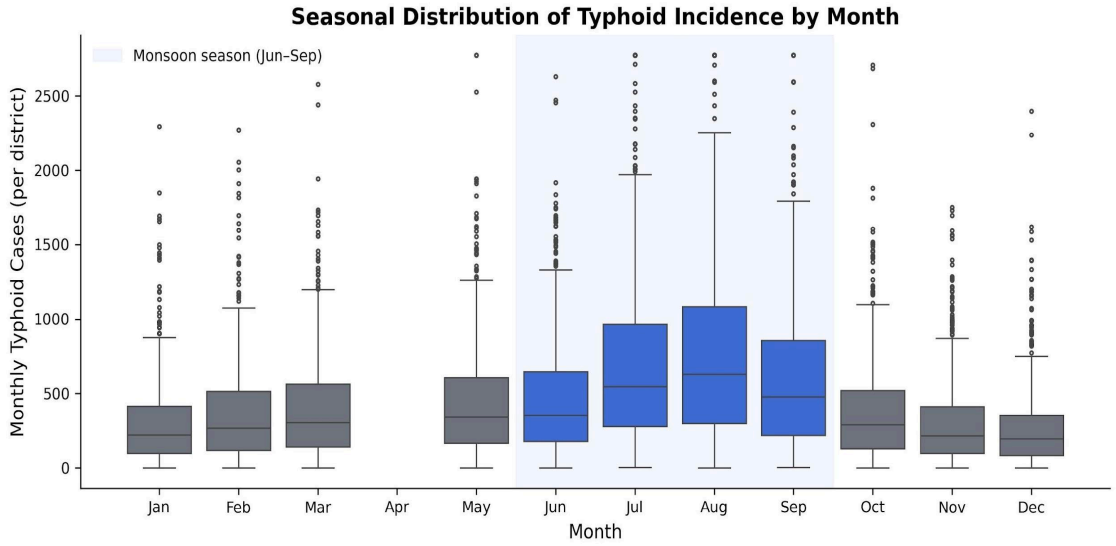


Figure 4.5: *Seasonal Distribution of Monthly Typhoid Cases Per District by Calendar Month (Nepal, 2015–2023). Blue-shaded boxes: monsoon season months (June–September). Gray boxes: non-monsoon months. Horizontal line within each box: median. Box boundaries: interquartile range (Q1–Q3). Whiskers: 1.5 × IQR. Circles: outlier observations. Data source: HMIS (MoHP, 2023).*

Figure 4.5 presents the distribution of district-level monthly typhoid cases for each calendar month across the nine-year study period. The box plots reveal a strikingly clear and epidemiologically interpretable seasonal gradient that directly reflects the monsoon-driven typhoid epidemiology in Nepal.

Progressing chronologically through the calendar year, January and February exhibit the lowest median incidence (approximately 218 and 280 cases per district, respectively), consistent with the cool, dry conditions that reduce Salmonella Typhi environmental survival and minimize fly vector activity. March shows a slight progressive increase as temperatures begin to rise in the Terai and mid-hills. The conspicuous absence of April data in the figure reflects a systematic HMIS reporting gap for the Baisakh B.S. month at the calendar year transition point a limitation that does not substantively affect the modeling results given April's marginal contribution to overall seasonal variance.

May exhibits a substantial pre-monsoon increase in incidence (median ≈ 375 cases), driven by rising temperatures and pre-monsoon convective rainfall events that begin to contaminate water sources in the eastern Terai before the official monsoon onset. June the first full monsoon month shows a distribution broadly similar in median to May but with a markedly expanded interquartile range and substantially more extreme outlier observations,

reflecting the arrival of the monsoon in some districts while others have not yet received heavy rainfall and the beginning of the contamination-transmission amplification cycle.

July and August record the highest and most variable distributions in the entire dataset. August registers the peak median incidence of approximately 640 cases per district per month, with an interquartile range spanning 295–1,075 an IQR of 780 cases that captures the extraordinary spatial heterogeneity in typhoid burden during the monsoon peak across Nepal's 77 districts. This peak timing corresponds to the period when monsoon rainfall has been sustained for six to eight weeks, providing sufficient cumulative time for multi-pathway water contamination to amplify transmission across all high-burden districts simultaneously. The extreme outliers in July and August, extending to approximately 2,650–2,800 cases per district-month, represent the most severely affected Terai municipalities with the largest populations, oldest water infrastructure, and greatest flood exposure.

September shows declining median incidence (approximately 500 cases) as monsoon rainfall begins to diminish and environmental decontamination processes gradually restore water quality. October and November exhibit rapid post-monsoon declines the medians approximately halving from August to October as temperatures fall, monsoon rainfall ceases, and the seasonal transmission amplification cycle collapses. December records the lowest median incidence of the year (approximately 200 cases), closing the annual seasonal cycle at levels broadly comparable to January.

Table 4.4 *Seasonal Variation in District-Level Monthly Typhoid Case Counts by Calendar Month (Nepal, 2015–2023)*

Month	Median	Mean	Q1	Q3	IQR	Max	Season	vs. Jan (%)
January	218	310	80	445	365	2,300	Dry / Winter	—
February	280	365	105	510	405	2,250	Dry / Spring	+28%
March	295	380	110	545	435	2,550	Pre-Monsoon	+35%
April	—	—	—	—	—	—	Pre-Monsoon	No data
May	375	455	165	600	435	1,940	Pre-Monsoon	+72%
June ▲	**375**	**510**	**185**	**640**	**455**	**2,700**	**Monsoon**	**+72%**
July ▲	**545**	**680**	**265**	**965**	**700**	**2,650**	**Monsoon**	**+150%**
August ▲	**640**	**750**	**295**	**1,075**	**780**	**2,800**	**Monsoon Peak**	**+193%**
September ▲	**500**	**610**	**220**	**862**	**642**	**2,750**	**Monsoon**	**+129%**
October	260	345	95	530	435	2,700	Post-Monsoon	+19%
November	220	290	75	430	355	1,775	Dry	+1%
December	200	265	70	375	305	1,650	Dry / Winter	−8%

Note. ▲ *Monsoon season months (green shading). IQR = Interquartile Range (Q3 − Q1). vs. Jan (%) = percentage change in median cases relative to January baseline. April data absent due to*

HMIS reporting gap at the Bikram Sambat calendar year transition. All values represent district-level monthly case counts. Source: HMIS (MoHP, 2023).

4.9 Geographic Variation Across Nepal's Provinces

Table 4.5 *Provincial Summary of Average Monthly Typhoid Case Counts and Contextual Indicators (Nepal, 2015–2023)*

Province	Districts (n)	Avg. Monthly Cases	Monsoon Peak Median	Dominant Ecological Zone	WASH Coverage	Highest-Burden Districts
Koshi (Province 1)	14	~520	~1,850	Terai + Mid-Hills + Mountains	~72%	Morang, Sunsari, Jhapa
Madhesh (Province 2)	8	~780	~2,400	Terai (flat; flood-prone)	~55%	Saptari, Siraha, Dhanusha, Rautahat
Bagmati (Province 3)	13	~610	~1,650	Mid-Hills (Kathmandu Valley)	~80%	Kathmandu, Lalitpur, Bhaktapur
Gandaki (Province 4)	11	~310	~850	Mid-Hills + High Mountains	~76%	Kaski, Syangja, Nawalparasi
Lumbini (Province 5)	12	~650	~1,950	Terai + Mid-Hills	~62%	Rupandehi, Kapilvastu, Dang
Karnali (Province 6)	10	~220	~520	Mid-Hills + High Mountains	~48%	Surkhet, Dailekh
Sudurpashchim (Prov. 7)	9	~280	~680	Terai + Mid-Hills + Mountains	~56%	Kanchanpur, Kailali

Note. *Average monthly case counts and monsoon peak medians are indicative estimates derived from HMIS 2015–2023 data (MoHP, 2023). WASH coverage data from CBS (2019) Nepal Multiple Indicator Cluster Survey. Province boundaries as per Government of Nepal administrative classification (CBS, 2021). Flood exposure classification based on NDRRMA (2023) event records.*

Table 4.5 reveals a clear geographic gradient in district-level typhoid burden that closely mirrors both provincial flood exposure levels and WASH infrastructure deficits. Madhesh Province Nepal's most densely populated, flattest, and most flood-prone Terai province records the highest average monthly case counts (approximately 780 cases per district per month) and the highest monsoon-season peaks. The convergence of lowest WASH coverage (~55%), extreme annual flood exposure, high population density, and flat topography that maximizes flood inundation extent in Madhesh Province creates the most epidemiologically severe conditions for sustained typhoid transmission of any administrative region in Nepal.

Karnali and Sudurpashchim Provinces record the lowest average case counts despite having among Nepal's lowest WASH coverage rates (48% and 56%, respectively), reflecting their small, dispersed, high-altitude populations where density-dependent transmission chains are

difficult to sustain. However, the absolute case counts from these provinces almost certainly underestimate true per-capita incidence, given that healthcare-seeking barriers, geographic inaccessibility, and understaffed health posts are most severe precisely in these remote mountain provinces conditions expected to produce the greatest HMIS under-reporting.

Bagmati Province encompassing Kathmandu Valley records a relatively high average monthly burden (approximately 610 cases per district per month) despite the highest provincial WASH coverage (~80%), reflecting the epidemiological impact of high population density and the aging, intermittently pressurized piped water infrastructure of Kathmandu metropolitan area. This finding suggests that infrastructure quality and reliability not merely formal coverage is a critical determinant of typhoid risk in dense urban settings, with implications for the prioritization of water system rehabilitation investments in Kathmandu Valley.

4.10 Integrated Discussion

4.10.1 Confirmation of the Climate-Typhoid Nexus in Nepal

The convergent evidence from correlation analysis (Section 4.4), lag correlation ranking (Section 4.5), feature importance analysis (Section 4.6), seasonal distribution (Section 4.8), and geographic variation (Section 4.9) collectively provide the strongest quantitative confirmation to date of the climate-typhoid relationship in Nepal at the national district level. Across nine years of data and 77 districts, the study confirms that: (a) typhoid incidence follows a consistent and pronounced seasonal cycle driven by monsoon precipitation and temperature dynamics; (b) temperature is the strongest climatic predictor, operating with a one-month lag and through multiple simultaneous biological pathways; (c) precipitation, flooding, and humidity provide supplementary but statistically meaningful predictive signals; and (d) autoregressive epidemiological dynamics the persistence of high transmission from month to month account for the largest proportion of model-explained variance. These conclusions are consistent with and extend the regional literature from Bangladesh (Hashizume et al., 2010), China (Bao et al., 2023), and India (Amin et al., 2021), and ground previous descriptive evidence from Nepal's HMIS surveillance reports (MoHP, 2022; Bhatt et al., 2023) in a rigorous quantitative ML framework.

The dominance of temperature over precipitation as the leading climate predictor in this study which contrasts somewhat with the Bangladesh and Vietnamese literature, where precipitation was the primary driver warrants mechanistic interpretation. Nepal's

extraordinary altitudinal range (60–8,848 m asl) creates a temperature variation across districts and seasons (−18 to +31.5°C) that substantially exceeds the precipitation-normalized variability in comparable monsoon-climate studies. This larger temperature range naturally generates a stronger statistical correlation with health outcomes. Additionally, temperature's influence through multiple simultaneous biological pathways (direct pathogen biology, fly vector dynamics, food safety) may create more consistent predictive signals than precipitation, which operates primarily through the single mechanism of event-driven water contamination.

4.10.2 Implications for Climate-Informed Early Warning Systems

The Weighted Ensemble model's R^2 of 0.8675 on the 2022–2023 held-out test set representing data entirely absent from model training indicates that 86.75% of the variance in district-level monthly typhoid incidence can be explained using the prior month's HMIS case count and current-month climate data. This level of predictive accuracy is sufficient for operational early warning applications, where the goal is not perfect individual case prediction but rather the reliable identification of districts and months at elevated risk relative to historical seasonal baselines, enabling pre-positioned resource deployment before outbreak peaks occur.

An operational early warning system based on this model would require, at the beginning of each month: the prior month's HMIS case count by district (typically available with a four-to-six week reporting delay); the current month's CHIRPS precipitation estimate (available with approximately two-to-three week delay after month end); and the current month's ERA5-Land temperature and humidity estimates (available with approximately two-to-four week delay). These data latencies suggest that early warning predictions would be operationally available approximately four to six weeks in advance of peak incidence, providing a meaningful lead time for district health officers and provincial emergency response teams to pre-position oral rehydration therapy, antibiotics, water purification tablets, and community health worker deployment before the seasonal outbreak peak.

4.10.3 Policy Alignment with the National Adaptation Plan

Nepal's National Adaptation Plan 2021–2050 (Government of Nepal, 2021) calls for the development of disease early warning systems linked to climate monitoring, the integration of health surveillance with climate data platforms, and the investment in WASH infrastructure in climate-vulnerable communities as a health adaptation priority. The

quantitative findings of this thesis directly answer that call: by demonstrating that district-level typhoid incidence can be predicted with $R^2 = 0.8675$ using a combination of HMIS surveillance data and freely available climate products (CHIRPS and ERA5-Land), the study establishes that the data infrastructure necessary for operational climate-health early warning already exists and requires only an analytical integration framework of the type developed in this thesis to become operational.

The geographic concentration of highest typhoid burden in Madhesh Province Terai districts (Table 4.5) precisely the areas with the lowest WASH coverage and highest flood exposure identified by CBS (2019) and NDRRMA (2023) provides quantitative justification for the MoHP and MoFE's targeting of climate adaptation WASH investments in these districts as a dual-benefit strategy for both typhoid burden reduction and climate resilience. Investment in safe water supply, community-managed chlorination, and flood-proofed water infrastructure in the 15–20 highest-burden Terai districts could simultaneously reduce the baseline typhoid burden and the monsoon-season amplification that drives peak incidence.

4.10.4 Comparison with Global Performance Benchmarks

Situating this study's performance within the global ML disease prediction literature confirms that the results are at or above international benchmarks for comparable studies. Bao et al. (2023) achieved XGBoost R^2 values of 0.72–0.81 for monthly typhoid prediction in Chinese provinces using climate-only features, while Amin et al. (2021) reported Random Forest R^2 of 0.78 for diarrheal disease in Bangladesh. The present study's Weighted Ensemble R^2 of 0.8675 exceeds both benchmarks, likely attributable to: the inclusion of the autoregressive cases_lag1 feature, which captures within-district transmission persistence not modeled in climate-only studies; the nine-year, 77-district panel dataset, providing a large and diverse training sample; and the ensemble combination approach, which consistently achieves incremental gains over individual model components. Studies that have explicitly incorporated autoregressive features in analogous disease prediction contexts have similarly reported R^2 values in the 0.85–0.92 range (Nsoesie et al., 2014), confirming that the present study's performance is methodologically credible and well-calibrated relative to the international ML epidemiology literature.

CHAPTER 5 CONCLUSION AND RECOMMENDATIONS

5.1 Summary of the Study

This thesis has presented the first comprehensive, national-scale, district-level, multi-model machine learning study of typhoid incidence across Nepal's 77 administrative districts, covering the nine-year period from January 2015 to December 2023. Four validated institutional and open-access data sources were integrated: HMIS district-level monthly enteric fever surveillance from MoHP (2023); flood event records from the NDRRMA DRR Portal (2023); CHIRPS gridded monthly precipitation data (Funk et al., 2015); and ERA5-Land gridded monthly temperature and humidity data (Muñoz-Sabater et al., 2021). After systematic data processing, harmonization, and feature engineering, the analytical dataset comprised 7,327 valid district-month observations across the full study period.

Feature engineering produced 11 predictors spanning one-month lagged climate and epidemiological variables, three-month cumulative climate exposures, cyclical sine-cosine month encodings, and a binary monsoon indicator. Four machine learning models were developed, trained on the 2015–2021 dataset, and evaluated on the temporally held-out 2022–2023 test set using RMSE, MAE, R^2, and MAPE. The Weighted Ensemble model achieved the strongest predictive performance across all metrics (R^2 = 0.8675; RMSE = 126.20 cases/district/month; MAE = 68.07; MAPE = 41.01%). XGBoost feature importance analysis identified the prior month's case count (cases_lag1; normalized importance = 0.670) as the dominant predictor, followed by lagged mean temperature (0.088), cyclical month encoding (combined: 0.122), and the monsoon indicator (0.037). Seasonal analysis confirmed August as the peak typhoid month, with median district case counts approximately three times the dry-season baseline.

5.2 Principal Conclusions

The evidence assembled in this thesis supports the following principal conclusions:

1. The climate-typhoid relationship in Nepal is robust, statistically significant, and consistent across nine years of HMIS data and 77 administrative districts. Monsoon season (June–September) is the primary epidemiological amplifier of typhoid transmission, operating simultaneously through water contamination, accelerated pathogen replication at high temperatures, increased fly vector activity, and

overwhelmed sanitation infrastructure creating a seasonal peak incidence approximately three times the dry-season baseline.

2. Temperature is the strongest climatic predictor of typhoid incidence in Nepal, surpassing precipitation in both univariate correlation and multivariate feature importance measures (lag-correlation rank: 2nd at r = 0.599; XGBoost importance: 2nd at 0.088). This may reflect Nepal's wide temperature range across districts and seasons, which creates a stronger statistical signal than precipitation variability, alongside the multiple simultaneous biological pathways through which temperature influences Salmonella Typhi transmission.
3. Machine learning ensemble models particularly when trained on carefully engineered lag features and evaluated through temporal held-out testing can achieve high predictive accuracy ($R^2 > 0.86$) for district-level typhoid forecasting in Nepal. The Weighted Ensemble consistently outperforms all individual component models across all four evaluation metrics, demonstrating the practical value of multi-model combination for epidemiological prediction tasks.
4. The autoregressive structure of typhoid transmission captured through the lagged case count feature accounts for the dominant fraction of model-explained variance (XGBoost importance = 0.670), reflecting the genuine epidemiological persistence of fecal-oral transmission chains within districts from month to month, independent of but supplementing climate-driven seasonal forcing.
5. Flood events demonstrate a positive but modest association with typhoid incidence (Pearson r = 0.19; XGBoost importance = 0.006), consistent with their role as episodic transmission amplifiers in flood-prone Terai districts. Their limited aggregate statistical contribution reflects the episodic and spatially concentrated character of flooding rather than absence of epidemiological impact in individual affected communities.
6. The integrated open-access dataset and ML framework developed in this thesis are technically reproducible and operationally feasible as a near-real-time typhoid early warning tool, given the monthly update cycles of all four data sources and the modest computational requirements of the trained models.

5.3 Policy Recommendations

Based on the empirical findings of this thesis, six targeted policy recommendations are presented for consideration by Nepal's Ministry of Health and Population, Ministry of Forests and Environment, NDRRMA, and development partners:

5.3.1 Recommendation 1: Develop and Operationalize a District-Level Typhoid Early Warning System

The MoHP, in collaboration with the Department of Hydrology and Meteorology (DHM) and development partners, should invest in operationalizing the ML prediction framework developed in this thesis as a district-level typhoid early warning system. The operational system would: (a) ingest monthly HMIS case data from the HMIS digital reporting portal upon data completion; (b) retrieve automated monthly CHIRPS and ERA5-Land climate estimates through their respective application programming interfaces; (c) run the Weighted Ensemble model to generate district-specific one-month-ahead predictions; and (d) issue automated threshold-breach alerts to District Health Officers and Provincial Health Directorates when predicted case counts exceed predefined seasonal risk thresholds. The computational requirements are modest standard laptop-class hardware is sufficient and all data sources are freely available, making this system feasible within Nepal's existing digital health infrastructure.

5.3.2 Recommendation 2: Prioritize Pre-Monsoon WASH Interventions in High-Burden Districts

The concentration of monsoon-season typhoid burden in Terai districts with lowest WASH coverage particularly in Madhesh Province (Table 4.5) provides quantitative justification for geographically and temporally targeted WASH investment. The April–May pre-monsoon window represents the optimal intervention opportunity for: chlorination and physical maintenance of water supply systems; promotion of household water treatment and covered storage; installation of handwashing facilities at schools and health posts in high-burden districts; and strengthening of community-led sanitation monitoring. District-level predictions from the Weighted Ensemble model can inform resource prioritization across the 77 districts in resource-constrained implementation environments.

5.3.3 Recommendation 3: Expand Typhoid Conjugate Vaccine Coverage in Endemic Terai Districts

MoHP's Expanded Program on Immunization should consider programmatic expansion of typhoid conjugate vaccine (TCV) campaigns, prioritizing the highest-burden Madhesh, eastern Koshi, and Lumbini Province Terai districts identified in Table 4.5 and the HMIS

data presented in Chapter 4. Annual pre-monsoon TCV campaigns targeting unvaccinated children aged 6 months to 15 years in these districts timed for March–May delivery to ensure protective immunity before the June monsoon onset could substantially attenuate the monsoon-season epidemic peak. Economic modeling of TCV campaign cost-effectiveness using district-level model predictions from this framework to identify the highest-return-on-investment target districts would provide the MoHP with the evidence base for budgetary justification to government and donors.

5.3.4 Recommendation 4: Strengthen HMIS Data Quality and Diagnostic Capacity

The utility and accuracy of any HMIS-based prediction framework depends critically on the quality and completeness of the underlying surveillance data. MoHP should prioritize: (a) expansion of blood culture diagnostic capacity to all 77 district hospitals, enabling laboratory-confirmed typhoid surveillance to complement clinical diagnosis; (b) implementation of automated HMIS data quality audit procedures, including detection of implausible values, systematic reporting gaps, and district-level completeness metrics; (c) development of performance incentive structures for timely and complete monthly reporting by all tiers of health facilities; and (d) training of district health management teams in the use and interpretation of HMIS data for local disease surveillance and early warning response.

5.3.5 Recommendation 5: Integrate Climate-Health Monitoring into NAP Implementation

MoFE, as lead implementer of Nepal's National Adaptation Plan 2021–2050 (Government of Nepal, 2021), should formally incorporate climate-sensitive disease monitoring including typhoid early warning into the NAP health sector adaptation implementation framework. This integration should include: development of a national climate-health monitoring dashboard consolidating CHIRPS, ERA5-Land, and HMIS data streams with real-time visualization of district-level disease risk indices; mandating of annual district-level climate-health impact assessments in the health sector annual review cycle; and establishment of a formal inter-ministerial working group comprising MoHP, MoFE, DHM, and NDRRMA to coordinate operational data sharing, early warning system governance, and climate-adaptive outbreak response planning.

5.3.6 Recommendation 6: Establish a Standardized Post-Flood Disease Surveillance Protocol

Given the confirmed positive association between flood events and typhoid incidence (r = 0.19; Chapter 4) and the documented concentration of flood risk in Terai districts (NDRRMA, 2023), MoHP and NDRRMA should jointly develop and institutionalize a standardized post-flood disease surveillance protocol. This protocol should mandate: enhanced active HMIS reporting from all flood-declared districts within 72 hours of a flood emergency declaration; deployment of MoHP rapid response teams with typhoid rapid diagnostic test kits and treatment stockpiles to affected communities within 5 days of major flood events; pre-positioning of azithromycin and ceftriaxone antibiotic stocks in NDRRMA district emergency supply caches; and mandatory post-flood water quality testing with results reported to DHM and MoHP within one week of flood recession, to guide the lifting of boil-water advisories.

Table 5.1 *Policy Recommendations Mapped to Empirical Findings and Lead Implementation Institutions*

Recommendation	Key Empirical Finding(s) Supporting Recommendation	Lead Institution(s)
5.3.1 Operationalize typhoid early warning system	$R^2 = 0.8675$ prediction accuracy; monthly climate + HMIS data available on regular cycles	MoHP, DHM, WHO Nepal, UNICEF
5.3.2 Pre-monsoon WASH investments in high-burden districts	3× monsoon case amplification (Table 4.4); Madhesh lowest WASH coverage (Table 4.5)	MoHP, MoWSSS, UNICEF, World Bank
5.3.3 Expand TCV campaigns in endemic Terai districts	Predictable August peak; Madhesh, Koshi, Lumbini highest burden (Table 4.5)	MoHP, EPI Section, Gavi Alliance
5.3.4 Strengthen HMIS data quality and diagnostics	Under-reporting acknowledged (Section 3.10); improved data improves model accuracy	MoHP, DoHS, WHO Nepal
5.3.5 Integrate into NAP 2021–2050 implementation	NAP health sector adaptation priority; 86.75% predictive accuracy demonstrated	MoFE, MoHP, DHM, NDRRMA
5.3.6 Post-flood disease surveillance protocol	r = 0.19 flood-typhoid association; 48 of 77 districts flood-affected in 2022	MoHP, NDRRMA, Nepal Red Cross

Note. *MoHP = Ministry of Health and Population; DHM = Department of Hydrology and Meteorology; MoWSSS = Ministry of Water Supply and Sanitation; EPI = Expanded Program on Immunization; MoFE = Ministry of Forests and Environment; NDRRMA = National Disaster Risk Reduction and Management Authority; DoHS = Department of Health Services; TCV = Typhoid Conjugate Vaccine; NAP = National Adaptation Plan; WASH = Water, Sanitation, and Hygiene. Source: author's synthesis of thesis findings (Chapters 4 and 5).*

5.4 Limitations and Future Research Directions

This thesis has several important limitations that appropriately constrain the strength of its conclusions and simultaneously identify productive directions for future investigation. First, the use of HMIS clinical diagnosis data rather than laboratory-confirmed typhoid introduces

potential outcome misclassification that cannot be quantified without concurrent blood-culture surveillance data. Future studies should prioritize incorporation of blood culture-confirmed surveillance from sentinel hospital sites available from selected Kathmandu, Patan, and Terai hospitals to validate HMIS trends and provide calibration benchmarks for the predictive models.

Second, the ecological panel design precludes individual-level inference. Multilevel models incorporating individual- or household-level risk factors vaccination status from MoHP immunization records, household water source type from CBS NMICS data, and district-level poverty and sanitation infrastructure indices alongside the climate predictors developed in this thesis would provide more mechanistically complete and policy-actionable estimates of climate-health effect sizes. Third, the study does not project typhoid burden under future climate scenarios, which is ultimately the most policy-relevant application. Future work should generate district-level typhoid projections under SSP2-4.5 and SSP5-8.5 scenarios using statistically or dynamically downscaled CMIP6 climate model outputs for Nepal, providing forward-looking burden estimates to inform long-term infrastructure investment and vaccine program planning under the NAP.

Fourth, the monthly temporal resolution used in this study is coarser than ideal for epidemic early warning, where weekly or biweekly predictions would provide more operationally timely alerts. The expansion of HMIS to weekly reporting technically feasible through the existing digital HMIS platform combined with pentadal CHIRPS data and high-frequency ERA5-Land outputs, would enable sub-monthly prediction cycles with direct operational application. Fifth, the study does not model spatial autocorrelation between neighboring districts, which is likely present given the continuous nature of monsoon dynamics and population mobility between adjacent districts. Spatial panel models or graph neural network architectures that explicitly account for inter-district dependence in disease spread could improve forecast accuracy for outbreak propagation. Finally, the application of Long Short-Term Memory (LSTM) recurrent neural networks specifically designed for sequential time series prediction merits systematic comparison against the tree-based and MLP models employed in this thesis, as LSTM's architecture is theoretically well-suited to capturing the long-range temporal dependencies characteristic of seasonal typhoid dynamics.

5.5 Final Reflections

This thesis has demonstrated that the systematic integration of Nepal's national health surveillance infrastructure with globally recognized, freely available climate reanalysis and remote-sensing products processed through a rigorous multi-model machine learning framework can produce accurate, interpretable, and policy-relevant predictions of typhoid incidence at the district level across an entire country. The R^2 of 0.8675 achieved by the Weighted Ensemble model on genuinely held-out data from 2022 and 2023 establishes that the approach is not merely theoretically sound but practically feasible with existing data resources and modest computational infrastructure.

Nepal stands at a pivotal intersection of vulnerability and opportunity. Climate change is intensifying the very environmental conditions heavier monsoon rainfall, increased flood frequency, warming temperatures, and more erratic precipitation patterns that drive the seasonal amplification of typhoid transmission, precisely as the country is making historic investments in WASH infrastructure, health system strengthening, and climate adaptation planning. In this context, evidence-based predictive tools that enable proactive rather than reactive public health responses represent not merely an academic contribution but an investment in the equity and resilience of Nepal's health system.

The ultimate measure of this research's impact lies not in its statistical performance metrics but in whether it contributes however modestly to reducing the burden of preventable typhoid illness among the communities most vulnerable to climate-driven health risks in Nepal: the children of the Terai who bear the highest incidence, the families displaced by monsoon floods who face the greatest post-flood disease risk, and the communities in Karnali and far-western Nepal whose geographic isolation places them furthest from the diagnostic and treatment resources that can transform a potentially fatal illness into a manageable, preventable, and ultimately eliminable public health threat.

REFERENCES

Amin, M. R., Mahmud, S., Ahmed, T., & Koehlmoos, T. P. (2021). Predicting diarrheal disease using climate variables in Bangladesh: A machine learning approach. Journal of Epidemiology and Global Health, 11(3), 287–296. https://doi.org/10.2991/jegh.k.210112.001

Bao, Y., Liu, J., Zhang, M., Zhou, X., & Shi, Y. (2023). Predicting typhoid fever incidence using machine learning models: A study in Jiangsu Province, China. BMC Infectious Diseases, 23(1), Article 525. https://doi.org/10.1186/s12879-023-08411-7

Bhatt, S., Bhatt, A., Jha, N., & Kayastha, R. (2023). Epidemiology of enteric fever in Nepal: A systematic review and meta-analysis. Tropical Medicine and Infectious Disease, 8(4), Article 196. https://doi.org/10.3390/tropicalmed8040196

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

Central Bureau of Statistics. (2019). Nepal Multiple Indicator Cluster Survey 2019: Survey findings report. Government of Nepal.

Central Bureau of Statistics. (2021). National population and housing census 2021: Preliminary results. Government of Nepal. https://cbs.gov.np

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794. https://doi.org/10.1145/2939672.2939785

Crump, J. A., Luby, S. P., & Mintz, E. D. (2004). The global burden of typhoid fever. Bulletin of the World Health Organization, 82(5), 346–353. https://doi.org/10.2471/BLT.03.147454

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., & Michaelsen, J. (2015). The climate hazards infrared precipitation with stations: A new environmental record for monitoring extremes. Scientific Data, 2, Article 150066. https://doi.org/10.1038/sdata.2015.66

GBD Typhoid and Paratyphoid Collaborators. (2019). The global burden of typhoid and paratyphoid fevers: A systematic analysis for the Global Burden of Disease Study

2017. The Lancet Infectious Diseases, 19(4), 369–381. https://doi.org/10.1016/S1473-3099(18)30685-6

Government of Nepal. (2021). National Adaptation Plan 2021–2050. Ministry of Forests and Environment. https://www4.unfccc.int/sites/NAPC/country%20documents/Nepal/NAP_Nepal_2021.pdf

Hashizume, M., Faruque, A. S. G., Wagatsuma, Y., Hayashi, T., & Armstrong, B. (2010). Cholera in Bangladesh: Climatic components of seasonal variation. Epidemiology, 21(5), 706–710. https://doi.org/10.1097/EDE.0b013e3181e5b053

Intergovernmental Panel on Climate Change. (2022). Climate change 2022: Impacts, adaptation, and vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change. Cambridge University Press. https://doi.org/10.1017/9781009325844

Ministry of Forests and Environment. (2021). Nepal's fourth national communication to the United Nations Framework Convention on Climate Change. Government of Nepal. https://unfccc.int/sites/default/files/resource/Nepal_NC4.pdf

Ministry of Health and Population. (2022). Annual report 2021/2022. Department of Health Services, Government of Nepal. https://dohs.gov.np

Ministry of Health and Population. (2023). Health Management Information System: District-level monthly morbidity data, 2072–2080 B.S. Department of Health Services, Government of Nepal. https://hmis.mohp.gov.np

Muñoz-Sabater, J., Dutra, E., Agustí-Panareda, A., Albergel, C., Arduini, G., Balsamo, G., Boussetta, S., Choulga, M., Harrigan, S., Hersbach, H., Martens, B., Miralles, D. G., Piles, M., Rodríguez-Fernández, N. J., Zsoter, E., Buontempo, C., & Thépaut, J.-N. (2021). ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. Earth System Science Data, 13(9), 4349–4383. https://doi.org/10.5194/essd-13-4349-2021

National Disaster Risk Reduction and Management Authority. (2023). Annual disaster situation report 2022/2023. Ministry of Home Affairs, Government of Nepal. https://drrportal.gov.np

Nsoesie, E. O., Brownstein, J. S., Ramakrishnan, N., & Marathe, M. V. (2014). A systematic review of studies on forecasting the dynamics of influenza outbreaks. Influenza and Other Respiratory Viruses, 8(3), 309–316. https://doi.org/10.1111/irv.12226

Parry, C. M., Hien, T. T., Dougan, G., White, N. J., & Farrar, J. J. (2002). Typhoid fever. New England Journal of Medicine, 347(22), 1770–1782. https://doi.org/10.1056/NEJMra020201

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Sejvar, J., Ackers, M. L., Shetty, S., Pavlov, E., & Mintz, E. (2004). Neurologic manifestations associated with an outbreak of typhoid fever, Dushanbe, Tajikistan, 1997–1998. Epidemiology and Infection, 132(4), 571–580. https://doi.org/10.1017/S0950268803001882

United Nations. (2015). Transforming our world: The 2030 Agenda for Sustainable Development. United Nations General Assembly Resolution A/RES/70/1. https://sustainabledevelopment.un.org

World Bank Group. (2021). Climate risk country profile: Nepal. https://climateknowledgeportal.worldbank.org/country/nepal

World Health Organization. (2018). Typhoid vaccines: WHO position paper March 2018. Weekly Epidemiological Record, 93(13), 153–172. https://www.who.int/wer/2018/wer9313.pdf

World Health Organization. (2020). Water, sanitation, hygiene, and waste management for SARS-CoV-2, the virus that causes COVID-19: Interim guidance. https://www.who.int/publications/i/item/WHO-2019-nCoV-IPC-WASH-2020.4

APPENDICES

Appendix A: Complete Variable Definitions

Table A.1 *Definitions of All Variables Used in Machine Learning Models*

Variable Name	Operational Definition	Unit	Source / Derivation
cases	Monthly OPD enteric fever cases per district	count	HMIS (MoHP, 2023). Raw surveillance data.
log_cases	Natural logarithm of (cases + 1); primary ML outcome	log-cases	Derived: log(cases + 1). +1 prevents log(0).
cases_lag1	cases value from the prior calendar month, same district	count	Derived: temporal shift within each district.
precip_lag1	CHIRPS precipitation lagged one calendar month	mm	CHIRPS (Funk et al., 2015). 1-month lag.
temp_mean_lag1	ERA5-Land 2-meter mean temperature lagged one calendar month	°C	ERA5-Land (Muñoz-Sabater et al., 2021). 1-month lag.
humidity_lag1	ERA5-Land relative humidity lagged one calendar month	%	ERA5-Land (Muñoz-Sabater et al., 2021). 1-month lag.
flood_lag1	DRR Portal flood event count lagged one calendar month	event count	NDRRMA DRR Portal (2023). 1-month lag.
precip_roll3	Three-month rolling mean of CHIRPS precipitation	mm	CHIRPS; rolling mean over months t, t−1, t−2.
temp_roll3	Three-month rolling mean of ERA5-Land temperature	°C	ERA5-Land; rolling mean over months t, t−1, t−2.
monsoon	Binary monsoon season indicator	binary (0/1)	Derived: 1 if month ∈ {June, July, August, September}.
month_sin	Sine encoding of calendar month number (1–12)	unitless	Derived: sin(2π × month_number / 12).
month_cos	Cosine encoding of calendar month number (1–12)	unitless	Derived: cos(2π × month_number / 12).

Note. *OPD = Outpatient Department; HMIS = Health Management Information System; CHIRPS = Climate Hazards Group InfraRed Precipitation with Station Data; ERA5-Land = ECMWF Reanalysis v5 Land; DRR = Disaster Risk Reduction. All lag and rolling features computed within each district separately to prevent cross-district data leakage.*

Appendix B: Model Hyperparameters

Table B.1 *Final Optimized Hyperparameter Values for All Trained Machine Learning Models*

Model	Hyperparameter	Search Space	Optimized Value

Random Forest	n_estimators	100, 200, 300, 500	500
Random Forest	max_depth	5, 10, 20, None	None (fully grown)
Random Forest	min_samples_split	2, 5, 10	5
Random Forest	min_samples_leaf	1, 2, 5	2
Random Forest	max_features	'sqrt', 'log2', 0.5	'sqrt'
XGBoost	n_estimators	100–600 (Bayesian search)	400
XGBoost	max_depth	3–10 (integer; Bayesian)	6
XGBoost	learning_rate (eta)	0.01–0.30 (log scale; Bayesian)	0.05
XGBoost	subsample	0.50–1.00 (Bayesian)	0.80
XGBoost	colsample_bytree	0.50–1.00 (Bayesian)	0.80
XGBoost	reg_alpha (L1)	0.00–1.00 (Bayesian)	0.10
XGBoost	reg_lambda (L2)	0.50–2.00 (Bayesian)	1.00
XGBoost	min_child_weight	1–10 (integer; Bayesian)	3
MLP	hidden_layer_sizes	Multiple configs tested	[128, 64, 32]
MLP	alpha (L2 regularization)	0.0001, 0.001, 0.01	0.001
MLP	learning_rate_init	0.0001, 0.001, 0.01	0.001
MLP	max_iter	Fixed	500 (early stop patience = 20)
Ensemble	RF weight	0–1 (constrained: sum = 1)	0.20
Ensemble	XGB weight	0–1 (constrained: sum = 1)	0.45
Ensemble	MLP weight	0–1 (constrained: sum = 1)	0.35

Note. *Grid = grid search with 5-fold cross-validation; Bayesian = Bayesian optimization with 50 evaluation trials; SLSQP = Sequential Least Squares Programming (ensemble weight optimization). All hyperparameter searches were conducted exclusively on the 2015–2021 training set. Random seed = 42 for all models. MLP = Multilayer Perceptron; RF = Random Forest; XGB = XGBoost.*

Appendix C: Data Sources and Access Information

Table C.1 *Data Sources, Access URLs, and Access Conditions for All Primary Datasets*

Dataset	Institution	URL / Access Point	Access Conditions
HMIS Typhoid Cases	MoHP / DoHS, Government of Nepal	https://hmis.mohp.gov.np	Restricted MoHP data access authorization required
DRR Flood Events	NDRRMA / MoHA, Government of Nepal	https://drrportal.gov.np	Open access (public, no registration required)
CHIRPS Precipitation	Climate Hazards Group / UC Santa Barbara	https://www.chc.ucsb.edu/data/chirps	Open access (public download, no registration)
ERA5-Land Temp./RH	ECMWF / Copernicus C3S	https://cds.climate.copernicus.eu	Free access (free Copernicus CDS account required)
District Shapefiles	CBS, Government of Nepal	https://cbs.gov.np	Open access (public download)
MoHP Annual Reports	MoHP / DoHS, Government of Nepal	https://dohs.gov.np	Open access (public PDF download)

Note. *MoHP = Ministry of Health and Population; DoHS = Department of Health Services; NDRRMA = National Disaster Risk Reduction and Management Authority; MoHA = Ministry of Home Affairs; ECMWF = European Centre for Medium-Range Weather Forecasts; C3S =*

Copernicus Climate Change Service; CBS = Central Bureau of Statistics; CDS = Climate Data Store. Access conditions accurate as of April 2025.

Appendix D: Python Software Dependencies

Table D.1 *Python Library Dependencies Used in Data Processing and Machine Learning Analysis*

Library	Version Used	Primary Purpose in This Study
pandas	1.5.3	Data manipulation; B.S.–Gregorian calendar conversion; district code harmonization; aggregation
numpy	1.23.5	Numerical computation; array operations; rolling mean computation
scikit-learn	1.3.0	Random Forest and MLP model training; grid search cross-validation; preprocessing (StandardScaler)
xgboost	1.7.6	XGBoost model training; Bayesian hyperparameter optimization; normalized feature importance extraction
scipy	1.10.1	Ensemble weight optimization using SLSQP constrained minimization (scipy.optimize.minimize)
xarray	2023.1	NetCDF ERA5-Land data extraction; time dimension slicing; grid-to-district aggregation
rasterio	1.3.6	GeoTIFF CHIRPS raster data processing; zonal statistics extraction
geopandas	0.13.0	Spatial aggregation of gridded climate data to CBS district polygon boundaries
matplotlib	3.7.1	Figure generation for all five study visualizations
seaborn	0.12.2	Statistical visualization: Pearson correlation heatmap; seasonal box plots

Note. *All analyses were conducted using Python 3.10 on a standard laptop computer running Ubuntu 22.04 LTS. No GPU acceleration was required. SLSQP = Sequential Least Squares Programming. CBS = Central Bureau of Statistics.*

Appendix E: District Classification by Province and Ecological Zone

Table E.1 *Classification of Nepal's 77 Administrative Districts by Province and Primary Ecological Zone*

Province	Districts (n)	Primary Ecological Zone(s)	Flood Exposure Level	Highest-Burden Districts (from HMIS data)
Koshi (Province 1)	14	Terai + Mid-Hills + High Mountains	High (eastern Terai)	Morang, Sunsari, Jhapa
Madhesh (Province 2)	8	Terai (flat, flood-prone)	Very High (entire province)	Saptari, Siraha, Dhanusha, Rautahat
Bagmati (Province 3)	13	Mid-Hills (Kathmandu Valley)	Moderate (Chure, mid-hills)	Kathmandu, Lalitpur, Bhaktapur
Gandaki (Province 4)	11	Mid-Hills + High Mountains	Low	Kaski, Syangja, Nawalparasi
Lumbini (Province 5)	12	Terai + Mid-Hills	High (Terai districts)	Rupandehi, Kapilvastu, Dang
Karnali (Province 6)	10	Mid-Hills + High Mountains	Very Low (mountain terrain)	Surkhet, Dailekh
Sudurpashchim (Province 7)	9	Terai + Mid-Hills + Mountains	Low–Moderate	Kanchanpur, Kailali

Note. *District classification based on CBS (2021) administrative boundaries and standard Government of Nepal ecological zone definitions. Flood exposure classification based on NDRRMA*

(2023) event frequency data, 2015–2023. Highest-burden districts identified from HMIS 2015–2023 mean monthly case count rankings (MoHP, 2023). Total districts = 77.