DISSERTATION

# HYBRID INTELLIGENCE IN ENVIRONMENTAL SYSTEMS: BRIDGING PHYSICS-BASED MODELING AND DEEP LEARNING FOR INTEGRATED WATER RESOURCE MANAGEMENT AND URBAN PLANNING

Submitted by

Mohamed Fawzy Mahmoud

Department of Civil & Environmental Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2026

Doctoral Committee:

Advisor: Mazdak Arabi

Ryan Bailey

Ryan Morrison

Christopher Goemans

Kendall DeJonge

# ABSTRACT


## HYBRID INTELLIGENCE IN ENVIRONMENTAL SYSTEMS: BRIDGING PHYSICS-BASED MODELING AND DEEP LEARNING FOR INTEGRATED WATER RESOURCE MANAGEMENT AND URBAN PLANNING

Nutrient pollution represents a critical water quality challenge requiring accurate prediction of nutrient transport through complex river networks under alternative management scenarios. Traditional process-based models provide mechanistic rigor but face computational limitations constraining comprehensive scenario analysis, while pure machine learning approaches offer efficiency but frequently violate fundamental physical principles and have limited interpretability. This dissertation develops a hybrid intelligence framework that strategically integrates multiple artificial intelligence techniques with physics-based principles to achieve both computational tractability and scientific validity in watershed nutrient modeling.

This dissertation contributes a hybrid intelligence framework that strategically integrates multiple AI techniques with physics-based principles to enable comprehensive watershed nutrient assessment under alternative future scenarios. The framework comprises four methodological contributions: (1) an ensemble machine learning approach using Bayesian Model Averaging that provides uncertainty-quantified predictions from natural landscapes; (2) a ConvLSTM-based deep learning framework integrating scenario planning with policy constraints for flexible long-term land use prediction; (3) a Knowledge-Guided Machine Learning architecture embedding

agricultural physics to ensure mass balance consistency; and (4) a calibrated SWIFT annual routing system enabling rapid source attribution analysis.

The framework comprises four interconnected components. Ensemble machine learning combining Random Forest, Gradient Boosting, and Extreme Gradient Boosting through Bayesian Model Averaging predicts nutrient loads from natural landscapes (forests and rangelands), achieving cross-validation $R^2$ values of 0.70-0.83 while reducing prediction uncertainty by 46% compared to individual algorithms. A Convolutional Long Short-Term Memory neural network trained on 39 years of annual land cover data predicts urban spatial expansion through 2050 under five Colorado Water Plan scenarios, revealing that 20-year forecasts (F1 = 0.87) substantially outperform 1-year predictions (F1 = 0.27) when dense temporal sequences enable detection of persistent growth patterns. A Knowledge-Guided Machine Learning framework incorporating physics-informed architecture and scenario-aware loss functions predicts agricultural nutrient loads across management scenarios, reducing mass balance violations by 95% while maintaining Nash-Sutcliffe Efficiency exceeding 0.85. The SWIFT routing model integrates source-specific predictions with wastewater facility data through streamlined mass balance algorithms, achieving computational speedups of approximately 1,900× relative to traditional SWAT+ applications while maintaining calibration performance (NSE > 0.90, $R^2$ > 0.91) across 200 monitoring stations.

Beyond computational gains, these findings advance scientific understanding: ensemble methods reveal that prediction uncertainty derives primarily from model structural differences rather than parameter uncertainty; the 'Temporal Depth Paradox', where 20-year forecasts outperform 1-year predictions, demonstrates that urban growth follows detectable multi-decadal patterns obscured by short-term stochasticity; and physics-informed constraints serve as beneficial regularization for nutrients, especially in scarce data situations.

Application across Colorado watersheds demonstrates that hybrid frameworks enable comprehensive scenario-based assessment in hours rather than weeks, transforming operational feasibility of regional water quality management. Source-specific delivery ratio quantification reveals systematic differences among nutrient sources (wastewater: 0.80; agriculture: 0.74), informing cost-effective management targeting. Results establish design principles for hybrid intelligence systems that preserve physical consistency while enabling practical application at scales relevant for regulatory decision support and strategic planning under uncertainty.

# ACKNOWLEDGMENTS

The journey toward a doctoral degree is neither solitary nor straightforward. Over these five years, I have been sustained by a remarkable constellation of mentors, colleagues, friends, and family whose contributions extend far beyond what these words can adequately capture. This work stands as a testament not only to my efforts but to the generosity, wisdom, and patience of those who believed in me even when the path forward seemed uncertain.

First and foremost, I extend my deepest gratitude to my advisor, Dr. Mazdak Arabi, whose guidance has been the compass by which I navigated this doctoral journey. For five years, his door remained open—not just for questions, but for ideas, opportunities, and encouragement. Dr. Arabi's ability to see potential in nascent concepts and his unwavering support through both triumphs and setbacks have shaped not only this dissertation but the researcher I have become.

I am profoundly grateful to Dr. Ryan Bailey, whose technical expertise proved invaluable at every turn. His patient instruction in the intricacies of SWAT and SWAT+ codes, and his tireless assistance in debugging SWIFT when we encountered seemingly insurmountable obstacles, were instrumental to the completion of this work. I am particularly moved to recall that Dr. Bailey taught my very first class upon arriving in the United States five years ago—a fitting full circle as I now complete this degree under his continued mentorship.

My sincere thanks extend to my committee members -Dr. Ryan Morrison, Dr. Christopher Goemans, and Dr. Kendall DeJonge- whose thoughtful guidance and constructive feedback strengthened this research immeasurably. Their diverse perspectives challenged me to think more broadly and rigorously about the implications of this work.

To my colleagues, who transformed what could have been an isolating experience into a collaborative adventure: Dr. Ahmed Gharib has been far more than a colleague; he has been a true friend who guided me through both academic challenges and the complexities of life in a new country from my very first days in the United States. Dr. Mahshid Mohammadzadeh walked beside me through every phase of this journey—we discussed, debated, and discovered together, and our shared work enriched us both. Dr. Donya Dezfooli offered wisdom at critical junctures that helped me navigate difficult decisions. Though they have all graduated and moved forward in their careers, their influence remains with me as I follow the path they helped illuminate.

I am grateful, too, for friends beyond my department whose presence reminded me that life extends beyond research. Their support provided the energy and perspective necessary to persist through difficult seasons.

My family has been my foundation. To my wife, Doaa Mahmoud: your understanding, patience, and unwavering belief in me have been the bedrock upon which this achievement rests. To my two children: you gave me reason to continue when surrender seemed easier. Though our time together meant time away from research, those hours spent with you recharged my spirit in ways nothing else could. You are my greatest accomplishment, and this degree pales in comparison to the privilege of being your father.

Finally, I dedicate this work to the memory of my younger brother, Ahmed, whose unexpected passing during my doctoral studies marked the darkest moment of my life. Miles separated us when I received that devastating news, and nothing -not this degree, not any achievement- will ever be worth more than one more conversation, one more moment with him. This dissertation stands as a tribute to his memory, ensuring that his presence remains woven into everything I accomplish. Ahmed, you are loved, you are missed, and you will never be forgotten.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

# CHAPTER 1 INTRODUCTION

## 1.1 BACKGROUND AND MOTIVATION

Nutrient pollution represents one of the most pervasive and challenging water quality issues confronting watersheds worldwide. Excess nitrogen and phosphorus loading to surface waters drives eutrophication, harmful algal blooms, hypoxic zones, and degradation of aquatic ecosystems (Vadas et al., 2007; Ouyang et al., 2007), with economic damages in the United States alone estimated at hundreds of millions of dollars annually (Colorado River Basin Salinity Control Forum, 2023; U.S. Department of the Interior, 2011). The complexity of nutrient dynamics in large river basins, involving multiple sources, diverse transport pathways, biogeochemical transformations, and interactions across spatial and temporal scales, necessitates sophisticated modeling frameworks capable of supporting evidence-based management decisions.

Effective nutrient management requires accurate prediction of nutrient loads from agricultural lands, urban areas, wastewater treatment facilities, and natural landscapes, along with quantification of how these loads are transported, transformed, and delivered through river networks to downstream receiving waters. Traditional approaches to watershed nutrient modeling have relied primarily on process-based models such as the Soil and Water Assessment Tool (SWAT), which simulate the physical, chemical, and biological processes governing nutrient cycling and transport (Gassman et al., 2007; White et al., 2010). While these models provide mechanistic understanding and can represent complex watershed processes, they face fundamental limitations. Calibration demands extensive data and computational resources, with high-dimensional parameter spaces creating equifinality problems where multiple parameter sets yield similar predictive performance but different process representations (Arabi et al., 2007; Jha, 2011).

Most critically, the computational burden of process-based models severely constrains their application for comprehensive uncertainty analysis, extensive scenario evaluation, and systematic sensitivity assessment—activities essential for robust water quality planning but often requiring thousands of model simulations (Kumar & Merwade, 2009; Cibin & Chaubey, 2015; Jeon et al., 2014).

Recent advances in machine learning offer alternative approaches that can identify complex patterns in environmental data with computational efficiency orders of magnitude greater than traditional simulation models. Techniques including Random Forests, Gradient Boosting, and deep neural networks have demonstrated strong predictive performance for water quality applications (Solomatine et al., 2007; Nourani & Kalantari, 2010; Zhang et al., 2009; Kratzert et al., 2019), with the capacity to capture nonlinear relationships between watershed characteristics and nutrient exports. However, purely data-driven machine learning approaches face their own critical limitations. These models often operate as "black boxes" with limited interpretability (Gill et al., 2007), making it difficult for stakeholders to understand the basis for predictions or trust model outputs for regulatory applications. More fundamentally, unconstrained machine learning models frequently violate basic physical principles such as mass conservation, potentially predicting nutrient exports that exceed possible inputs or generating predictions that are inconsistent with established hydrological and biogeochemical understanding (Solomatine & Shrestha, 2009). Such violations undermine scientific credibility and limit the utility of these models for informing management decisions that require physically defensible projections.

This tension between the computational efficiency of machine learning and the physical consistency of process-based models represents a central challenge in contemporary environmental modeling. The emergence of hybrid approaches that integrate domain knowledge

with data-driven learning offers a promising path forward (Bulygina et al., 2012; Faybishenko, 2010; Velázquez et al., 2011), enabling development of modeling frameworks that achieve both computational tractability and scientific rigor.

## 1.2 THE COLORADO CONTEXT AND WATER PLAN SCENARIOS

Colorado's watersheds provide an ideal setting for developing and testing integrated nutrient modeling frameworks. The state's river basins span dramatic environmental gradients from high-elevation montane headwaters exceeding 4,000 meters to semi-arid plains below 1,000 meters, with mean annual precipitation ranging from over 1,000 mm in mountain regions to less than 400 mm in eastern plains. Land use varies from extensive forests and rangelands in mountainous areas to intensive irrigated agriculture in river valleys and rapidly expanding urban development along the Front Range corridor, where population has grown substantially over recent decades.

These diverse conditions create complex spatial patterns of nutrient sources and transport processes. High-elevation watersheds with minimal human development contribute relatively low nutrient loads dominated by natural landscape processes, while downstream reaches receive substantial inputs from municipal wastewater treatment facilities, agricultural fertilizer application and livestock operations, and urban stormwater runoff. The transition from snowmelt-dominated hydrologic regimes in mountains to heavily managed irrigation systems in valleys introduces additional complexity through altered flow patterns, return flows, and seasonal variations in nutrient mobilization and delivery.

The Colorado Water Plan, developed through extensive stakeholder engagement, establishes a framework for water resource management under uncertainty by defining five alternative future scenarios spanning different assumptions about demographic, economic, and policy trajectories

(Kim et al., 2022; Gandini et al., 2021). These scenarios—Business as Usual, Weak Economy, Cooperative Growth, Adaptive Innovation, and Hot Growth—represent plausible future conditions characterized by varying population growth rates (ranging from 34% to 57% increases by 2050), urban development densities, agricultural intensities, and regulatory frameworks. This scenario-based structure acknowledges fundamental uncertainties in long-term projections and enables exploration of how different future pathways might influence water quality outcomes (Chu et al., 2020; Cong et al., 2023).

Application of consistent scenarios across all components of this research transforms the framework from producing singular predictions to creating an analytical system capable of exploring multiple plausible futures. This approach directly addresses a limitation of traditional watershed modeling studies that typically generate point predictions rather than characterizing the range of possible outcomes under alternative assumptions. By maintaining scenario consistency from urban growth prediction through agricultural management to downstream water quality impacts, the framework enables systematic evaluation of questions such as: How do nutrient loads differ if development follows compact transit-oriented patterns versus dispersed growth? What are the relative water quality benefits of agricultural best management practices compared to wastewater treatment upgrades? How do delivery ratios from different sources vary under alternative hydrologic regimes?

## 1.3 RESEARCH OBJECTIVES AND INTEGRATED FRAMEWORK

This dissertation develops a comprehensive, multi-scale framework for predicting watershed nutrient dynamics under alternative future scenarios by strategically integrating multiple machine learning approaches with physically based routing models. The overarching contribution is

demonstrating that the perceived trade-off between computational efficiency and scientific validity in watershed modeling is not fundamental but rather reflects limitations in how modeling approaches have traditionally been deployed. Strategic integration of machine learning with physical principles achieves both objectives simultaneously by matching modeling approaches to the characteristics of specific prediction challenges.

The framework consists of four interconnected components addressing distinct aspects of watershed nutrient prediction. First, an ensemble machine learning methodology predicts nutrient loads from natural landscapes (forests and rangelands) by combining multiple algorithms through Bayesian Model Averaging (Hoeting et al., 1999; Raftery et al., 1997) to achieve predictions that are both accurate and robust across diverse environmental conditions. This approach leverages twenty years of water quality data from 150 watersheds to learn complex relationships between climatic, physiographic, and soil characteristics and nutrient export rates, while systematic uncertainty quantification and spatial cross-validation ensure reliability when applied to ungauged locations (Wang et al., 2023).

Second, a deep learning framework based on Convolutional Long Short-Term Memory neural networks predicts spatially explicit urban land cover change through 2050 under the five Colorado Water Plan scenarios (Kim et al., 2022; Peng et al., 2024). By processing a uniquely dense 39-year annual time series of land cover data, this component captures long-term spatiotemporal patterns in urban growth that traditional cellular automata models struggle to represent (Li & and Yeh, 2002; Mustafa et al., 2017; Tang & Bennett, 2010). Integration of conservation constraints and transit-oriented development policies ensures predictions respect ecological priorities and infrastructure planning objectives (Allam et al., 2022), while scenario-specific parameterization

enables quantification of how alternative development trajectories influence urban spatial form and associated stormwater nutrient loading.

Third, a Knowledge-Guided Machine Learning framework addresses agricultural nutrient loading by integrating domain knowledge directly into neural network architectures and training procedures (Bulygina et al., 2012; Ghosh & Ghosh, 2010). This physics-informed approach embeds representations of hydrological processes—runoff generation, erosion dynamics, and nutrient cycling—within the network structure, while scenario-aware loss functions enforce mass balance constraints and process consistency. A novel Three-Phase Progressive Training strategy enables stable convergence by gradually introducing physical constraints during optimization (Auxepaules et al., 2008), allowing the model to learn fundamental statistical patterns before imposing stringent scientific requirements. Application to approximately 21,000 irrigated fields across Colorado under factorial combinations of tillage and irrigation practices demonstrates that this approach achieves strong predictive accuracy while dramatically reducing violations of physical principles compared to unconstrained machine learning.

Fourth, an integrated watershed routing framework combines the source-specific load predictions with wastewater treatment facility data through the SWIFT modeling system (Arnold et al., 2018) to simulate constituent transport through Colorado's river networks. This hybrid framework accepts flow outputs from calibrated SWAT+ models and applies streamlined mass balance algorithms to route nutrients through channel networks, achieving computational speedups of approximately 1,900× relative to full process-based simulations while maintaining high predictive accuracy. Systematic calibration using Sobol global sensitivity analysis (Saltelli et al., 2008; Sobol, 2001) and validation against 200 monitoring stations establishes model reliability, while load substitution experiments quantify source-specific delivery ratios that reveal how efficiently

nutrients from different sources are transported to downstream locations (Wollheim et al., 2008; Jones et al., 2015).

The integration of these four components creates a complete prediction system: ensemble machine learning generates forest and rangeland nutrient loads at the subwatershed scale; ConvLSTM predictions of future urban extent are translated to stormwater nutrient contributions; Knowledge-Guided Machine Learning provides agricultural loads under alternative management scenarios; and SWIFT routes these source-specific loads through calibrated river networks to predict in-stream concentrations and downstream deliveries. Consistent application of Colorado Water Plan scenarios across all components enables direct comparison of how alternative futures influence nutrient dynamics from source generation through transport to downstream water quality outcomes.

## 1.4 METHODOLOGICAL INTEGRATION AND DATA FLOW

The utility of this research seeks to extend beyond the individual modeling innovations presented in each chapter, relying instead on their strategic integration into a cohesive system. As illustrated in Figure 1- 1, the data flow and dependencies between chapters are designed to ensure methodological consistency across varying domains. To facilitate this integration, spatial and temporal scales are harmonized throughout the dissertation. Spatially, all predictions maintain a resolution of HUC12 watersheds (Chapter 2) or finer (Chapters 3 and 4), a granularity that enables the aggregation of diverse inputs into the common spatial units required for the routing mechanisms detailed in Chapter 5. Temporally, historical model training is standardized to the period spanning 2000 through 2023 across all chapters, while future projections are uniformly

targeted toward 2050. This temporal alignment ensures that model outputs correspond directly with the planning horizons and scenario frameworks established by the Colorado Water Plan.

Within this framework, the five Colorado Water Plan scenarios provide a consistent set of assumptions regarding population growth, urban density, agricultural intensity, and regulatory environments for Chapters 3, 4, and 5. While the ensemble machine learning approach utilized in Chapter 2 remains scenario-agnostic -focusing on predicting contributions from current natural landscapes- it serves the critical function of quantifying uncertainty bounds that inform the sensitivity analyses in subsequent chapters.

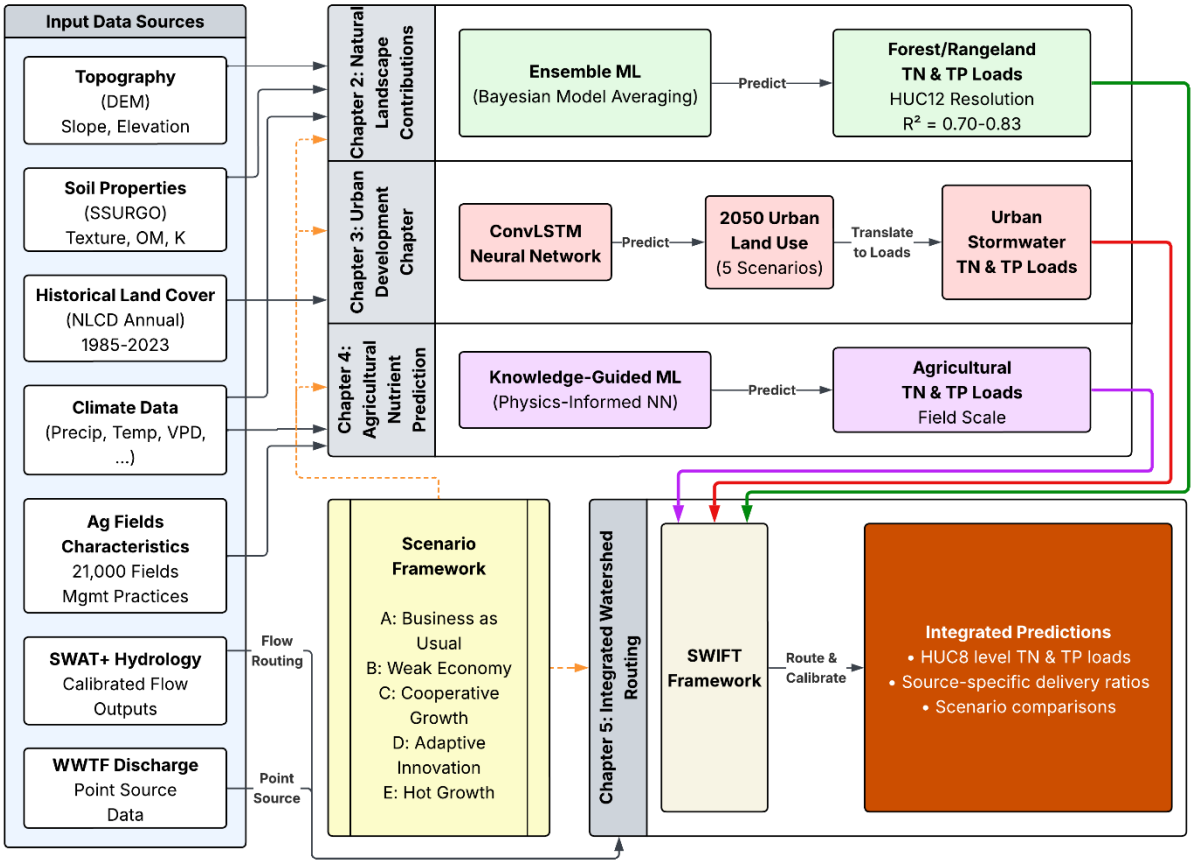


Figure 1- 1 Schematic representation of the integrated modeling framework.

To ensure the validity of these integrated projections, the framework incorporates physical constraints appropriate to each specific domain. This includes the application of mass balance principles within the Knowledge-Guided Machine Learning (KGML) model (Chapter 4) and the SWIFT channel routing system (Chapter 5), as well as the strict preservation of known conservation areas within the ConvLSTM architecture (Chapter 3). These constraints are intended to ensure that statistical predictions respect fundamental physical principles. Furthermore, the selection of specific model architecture reflects the computational requirements necessary for such integration. The efficiency of the chosen model ranging from the Ensemble ML and KGML approaches that predict loads for approximately 21,000 fields in seconds, to the SWIFT model's rapid routing capabilities, is a deliberate design choice. This computational efficiency facilitates the comprehensive 5-scenario by 4-source analyses that might otherwise prove computationally prohibitive under traditional, more process-intensive modeling paradigms.

## 1.5 RESEARCH CONTRIBUTIONS

This dissertation advances the field of watershed nutrient modeling through several fundamental contributions spanning methodological innovation, operational feasibility, and decision-support capabilities.

Chapter 2 contributes an ensemble machine learning framework using Bayesian Model Averaging that provides robust, uncertainty-quantified nutrient load predictions from natural landscapes, enabling reliable baseline estimates for ungauged watersheds. Chapter 3 contributes a deep learning framework integrating ConvLSTM neural networks with scenario-based planning and policy constraint embedding, enabling flexible generation of spatially-explicit land use maps under alternative futures through 2050. Chapter 4 contributes a Knowledge-Guided Machine Learning

framework that integrates agricultural domain knowledge into neural network architecture and training, enabling physically-consistent nutrient predictions across management scenarios while reducing data requirements. Chapter 5 contributes an integrated annual routing framework combining SWIFT with source-specific load predictions, enabling comprehensive scenario analysis and source attribution at regional scales.

Methodological advances include demonstration that hybrid frameworks strategically combining multiple artificial intelligence and machine learning approaches with physically based models can simultaneously achieve computational efficiency and scientific rigor. The ensemble machine learning methodology establishes that Bayesian Model Averaging provides superior stability and uncertainty quantification compared to individual algorithms while maintaining minimal computational overhead (Hoeting et al., 1999; Duan et al., 2007). The ConvLSTM framework reveals a counterintuitive "Temporal Depth Paradox" where long-term predictions substantially outperform short-term forecasts when models are trained on dense multi-decadal time series (Al-Raeei, 2025; Peng et al., 2024), challenging conventional assumptions about forecast uncertainty established in earlier cellular automata studies (Gui et al., 2025; Yaagoubi et al., 2024). The Knowledge-Guided Machine Learning approach demonstrates that progressive integration of physics constraints prevents the training instabilities that plague multi-objective optimization (Tosserams et al., 2006; Rocha & Fernandes, 2011), enabling neural networks to learn solutions satisfying both statistical accuracy and conservation laws. The SWIFT application shows that dramatic computational speedups need not compromise predictive performance when hydrologic simulation is strategically decoupled from constituent routing (Arnold et al., 2018).

Scientific insights emerge from applying these methods to Colorado watersheds. Ensemble model weights vary systematically across landscape types, with tree-based methods receiving higher

weights in threshold-dominated headwaters while boosting methods dominate in agricultural lowlands, suggesting different algorithmic structures capture different process characteristics (Chapter 2). The Temporal Depth Paradox indicates that urban growth follows persistent multi-decadal trajectories temporarily obscured by short-term stochasticity, with the ConvLSTM architecture effectively filtering noise to identify underlying growth signals (Chapter 3). Differential constraint responses between nitrogen and phosphorus -where phosphorus predictions actually improve under mass balance guidance (NSE: 0.845 to 0.850)- reveal that sediment-associated transport processes are more prone to overfitting than dissolved transport pathways (Chapter 4). Furthermore, physics-informed models demonstrate superior performance in data-scarce regimes, achieving reliable predictions with only 1% of training data while pure ML maintains violation rates above 2.7% even with full datasets (Chapter 4). Source-specific delivery ratios quantify transport pathway effects hypothesized but rarely measured at watershed scales, revealing systematic differences reflecting landscape retention opportunities (Chapter 5).

Operational feasibility is established through demonstration that comprehensive scenario-based watershed assessment is computationally tractable at regional scales. The framework transforms computational requirements from days or weeks to minutes for state-wide nutrient predictions, enabling activities previously impractical with traditional approaches (Huang et al., 2024; Neumann et al., 2021). Systematic sensitivity analysis using thousands of model evaluations becomes feasible for identifying critical parameters and quantifying uncertainty propagation. Extensive "what-if" scenario evaluation supports exploration of alternative management strategies without prohibitive computational burdens. This efficiency advance transforms watershed modeling from a research exercise requiring specialized expertise and substantial computing resources to an operational decision-support tool accessible for routine planning applications.

Physical consistency and stakeholder trust are maintained through integration of domain knowledge at multiple levels. Conservation constraints embedded in urban growth predictions ensure development respects protected areas and ecological priorities (Stralberg et al., 2020; Langpap et al., 2008). Physics-informed neural network architectures explicitly represent agricultural processes including runoff generation, erosion dynamics, and nutrient cycling, with mass balance loss functions preventing predictions that violate fundamental principles. Rigorous calibration of the routing framework against extensive monitoring data validates that predictions match observed nutrient dynamics across diverse watershed conditions (Moriasi et al., 2007). This demonstrated adherence to scientific principles builds stakeholder confidence necessary for operational deployment in regulatory contexts, addressing a critical barrier to machine learning adoption in environmental management.

Uncertainty characterization moves beyond point predictions to provide decision-makers with multiple plausible futures under alternative assumptions. Ensemble machine learning quantifies prediction uncertainty through distributions across multiple model structures (Raftery et al., 2005; Yang et al., 2021). Scenario-based prediction explicitly acknowledges irreducible uncertainties in long-term projections of population growth, land use change, agricultural practices, and climate conditions. Sensitivity analysis identifies which parameters and assumptions most strongly influence predictions (Saltelli et al., 2002), guiding priorities for data collection and model refinement. This comprehensive uncertainty assessment enables robust decision-making that accounts for limitations in predictive capability while still providing actionable insights for water quality planning.

Source-specific delivery quantification enables targeted management strategies by revealing how efficiently nutrients from different sources reach downstream monitoring locations. Load

substitution experiments isolate contributions from wastewater treatment facilities, agricultural lands, urban stormwater, and natural landscapes, quantifying distinct delivery ratios reflecting different transport pathways and retention opportunities (Wollheim et al., 2008; Jones et al., 2015; Natho et al., 2020). Systematic differences—with wastewater showing delivery ratios near 0.80 compared to agricultural ratios near 0.74—inform cost-effectiveness analysis by accounting for source-specific transport efficiency (Maringanti et al., 2009). Spatial variation in delivery ratios across subwatersheds identifies locations where management actions would produce greatest downstream benefits, supporting optimization of nutrient reduction investments.

Transferability and generalization are addressed through framework design that separates methodology from geographic application. While developed and tested for Colorado watersheds, the modular structure enables adaptation to other regions by retraining models with local data while maintaining the overall analytical approach. Systematic evaluation procedures including spatial cross-validation assess how well models predict in ungauged locations (Wang et al., 2023; Brenning, 2023), quantifying the extent to which learned relationships transfer across watershed conditions. Documentation of training strategies, hyperparameter optimization, and calibration protocols provides guidance for applying similar frameworks in different geographic contexts.

## 1.6 DISSERTATION ORGANIZATION

The remainder of this dissertation is organized as follows. Chapter 2 develops the ensemble machine learning methodology for predicting salinity and nutrient loads from natural landscapes, establishing foundational approaches for uncertainty quantification and spatial prediction that inform subsequent chapters. Chapter 3 presents the ConvLSTM framework for urban land cover prediction under alternative scenarios, demonstrating how deep learning can capture long-term

spatiotemporal patterns while respecting conservation and infrastructure planning constraints. Chapter 4 introduces the Knowledge-Guided Machine Learning approach for agricultural nutrient prediction, showing how domain knowledge integration enables physically consistent predictions across management scenarios. Chapter 5 describes the SWIFT routing framework and its calibration across Colorado watersheds, demonstrating computational efficiency gains and quantifying source-specific delivery ratios. Chapter 6 synthesizes findings across all components, discusses broader implications for water quality management and machine learning applications in environmental science, identifies limitations, and proposes directions for future research. Collectively, these chapters establish a comprehensive framework that bridges the gap between computational efficiency and scientific validity in watershed nutrient modeling, providing tools that support evidence-based decision-making under uncertainty for water quality management.

## 1.7 REFERENCES


Abbaspour, K., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., ... & Srinivasan, R. (2007). Modelling hydrology and water quality in the pre-alpine/alpine thur watershed using swat. Journal of Hydrology, 333(2-4), 413-430. https://doi.org/10.1016/j.jhydrol.2006.09.014

Allam, Z., Bibri, S. E., Jones, D. S., Chabaud, D., & Moreno, C. (2022). Unpacking the '15-Minute City' via 6G, IoT, and Digital Twins: Towards a New Narrative for Increasing Urban Efficiency, Resilience, and Sustainability. Sensors, 22(4), Article 4. https://doi.org/10.3390/s22041369

Al-Raeei, M. (2025). The smart future for sustainable development: Artificial intelligence solutions for sustainable urbanization. Sustainable Development, 33(1), 508-517. https://doi.org/10.1002/sd.3131

Arabi, M., Govindaraju, R., Engel, B., & Hantush, M. (2007). Multiobjective sensitivity analysis of sediment and nitrogen processes with a watershed model. Water Resources Research, 43(6). https://doi.org/10.1029/2006wr005463

Arnold, J., Bieger, K., White, M., Srinivasan, R., Dunbar, J., & Allen, P. (2018). Use of decision tables to simulate management in swat+. Water, 10(6), 713. https://doi.org/10.3390/w10060713

Auxepaules, L., Py, D., & Lemeunier, T. (2008). A diagnosis method that matches class diagrams in a learning environment for object-oriented modeling. 2008 Eighth IEEE International Conference on Advanced Learning Technologies, 26-30. https://doi.org/10.1109/icalt.2008.187

Brenning, A. (2023). Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. International Journal of Geographical Information Science, 37(3), 584-606. https://doi.org/10.1080/13658816.2022.2131789

Bulygina, N., Ballard, C., McIntyre, N., O'Donnell, G., & Wheater, H. (2012). Integrating different types of information into hydrological model parameter estimation: application to ungauged catchments and land use scenario analysis. Water Resources Research, 48(6). https://doi.org/10.1029/2011wr011207

Chu, X., Shi, Z., Yang, L., & Guo, S. (2020). Evolutionary Game Analysis on Improving Collaboration in Sustainable Urban Regeneration: A Multiple-Stakeholder Perspective. Journal of Urban Planning and Development, 146(4), 04020046. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000630

Cibin, R. and Chaubey, I. (2015). A computationally efficient approach for watershed scale spatial optimization. Environmental Modelling & Software, 66, 1-11. https://doi.org/10.1016/j.envsoft.2014.12.014

Colorado River Basin Salinity Control Forum. (2023). Water Quality Standards for Salinity—Colorado River System. https://barnettwater.sharepoint.com/:b:/g/ERm4dI86t1JEndfeaozZT4cBmfb8JoqmHh48Y4Vm3JA03w?e=vuThnW

Cong, C., Pan, H., Page, J., Barthel, S., & Kalantari, Z. (2023). Modeling place-based nature-based solutions to promote urban carbon neutrality. Ambio, 52(8), 1297-1313. https://doi.org/10.1007/s13280-023-01872-x

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources, 30(5), 1371-1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Faybishenko, B. (2010). Fuzzy-probabilistic calculations of water-balance uncertainty. Stochastic Environmental Research and Risk Assessment, 24(6), 939-952. https://doi.org/10.1007/s00477-010-0379-y

Gandini, A., Quesada, L., Prieto, I., & Garmendia, L. (2021). Climate change risk assessment: A holistic multi-stakeholder methodology for the sustainable development of cities. Sustainable Cities and Society, 65, 102641. https://doi.org/10.1016/j.scs.2020.102641

Gassman, P., Reyes, M., Green, C., & Arnold, J. (2007). The soil and water assessment tool: historical development, applications, and future research directions. Transactions of the Asabe, 50(4), 1211-1250. https://doi.org/10.13031/2013.23637

Ghosh, R. and Ghosh, S. (2010). DNS of compressible turbulent flows. In Turbulence and Interactions (pp. 537-547). Springer. https://doi.org/10.1007/978-90-481-3652-0_80

Gill, M., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. Water Resources Research, 43(7). https://doi.org/10.1029/2006wr005298

Gui, B., Bhardwaj, A., & Sam, L. (2025). Cellular automata models for simulation and prediction of urban land use change: Development and prospects. Artificial Intelligence in Geosciences, 6(2), 100142. https://doi.org/10.1016/j.aiig.2025.100142

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14(4), 382-401. https://www.jstor.org/stable/2676803

Huang, J., Arhonditsis, G., Zhang, Y., Zhang, S., Ji, Y., Paerl, H., ... & Gao, J. (2024). The critical role of hydrological distance in shaping nutrient dynamics along the watershed-lake continuum. Earth's Future, 12(10). https://doi.org/10.1029/2024ef004860

Jeon, J., Park, C., & Engel, B. (2014). Comparison of performance between genetic algorithm and sce-ua for calibration of scs-cn surface runoff simulation. Water, 6(11), 3433-3456. https://doi.org/10.3390/w6113433

Jha, M. (2011). Evaluating hydrologic response of an agricultural watershed for watershed analysis. Water, 3(2), 604-617. https://doi.org/10.3390/w3020604

Jones, C., Scott, D., Guth, C., Hester, E., & Hession, W. (2015). Seasonal variation in floodplain biogeochemical processing in a restored headwater stream. Environmental Science & Technology, 49(22), 13190-13198. https://doi.org/10.1021/acs.est.5b02426

Kim, J. M., Park, J. S., Lee, C. Y., & Lee, S. G. (2022). PREDICTING OF URBAN EXPANSION USING CONVOLUTIONAL LSTM NETWORK MODEL: THE CASE OF SEOUL METROPOLITAN AREA, KOREA. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-4-W3-2022, 113-118. https://doi.org/10.5194/isprs-annals-X-4-W3-2022-113-2022

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research, 55(12), 11344-11354. https://doi.org/10.1029/2019WR026065

Kumar, S. and Merwade, V. (2009). Impact of watershed subdivision and soil data resolution on swat model calibration and parameter uncertainty. JAWRA Journal of the American Water Resources Association, 45(5), 1179-1196. https://doi.org/10.1111/j.1752-1688.2009.00353.x

Langpap, C., Hascic, I., & Wu, J. (2008). Protecting Watershed Ecosystems through Targeted Local Land Use Policies. American Journal of Agricultural Economics, 90(3), 684-700. https://doi.org/10.1111/j.1467-8276.2008.01145.x

Li, X., & and Yeh, A. G.-O. (2002). Neural-network-based cellular automata for simulating multiple land use changes using GIS. International Journal of Geographical Information Science, 16(4), 323-343. https://doi.org/10.1080/13658810210137004

Maringanti, C., Chaubey, I., & Popp, J. (2009). Development of a multiobjective optimization tool for the selection and placement of best management practices for nonpoint source pollution control. Water Resources Research, 45(6). https://doi.org/10.1029/2008wr007094

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.

Mustafa, A., Cools, M., Saadi, I., & Teller, J. (2017). Coupling agent-based, cellular automata and logistic regression into a hybrid urban expansion model (HUEM). Land Use Policy, 69, 529-540. https://doi.org/10.1016/j.landusepol.2017.10.009

Natho, S., Tschikof, M., Bondar-Kunze, E., & Hein, T. (2020). Modeling the effect of enhanced lateral connectivity on nutrient retention capacity in large river floodplains: how much connected floodplain do we need? Frontiers in Environmental Science, 8. https://doi.org/10.3389/fenvs.2020.00074

Neumann, A., Dong, F., Shimoda, Y., Arnillas, C., Javed, A., Yang, C., ... & Arhonditsis, G. (2021). A review of the current state of process-based and data-driven modelling: guidelines for lake erie managers and watershed modellers. Environmental Reviews, 29(4), 443-490. https://doi.org/10.1139/er-2020-0070

Nourani, V. and Kalantari, O. (2010). Integrated artificial neural network for spatiotemporal modeling of rainfall-runoff-sediment processes. Environmental Engineering Science, 27(5), 411-422. https://doi.org/10.1089/ees.2009.0353

Ouyang, W., Hao, F., & Wang, X. (2007). Regional non point source organic pollution modeling and critical area identification for watershed best environmental management. Water, Air, & Soil Pollution, 187(1-4), 251-261. https://doi.org/10.1007/s11270-007-9513-y

Peng, Z.-R., Lu, K.-F., Liu, Y., & Zhai, W. (2024). The Pathway of Urban Planning AI: From Planning Support to Plan-Making. Journal of Planning Education and Research, 44(4), 2263-2279. https://doi.org/10.1177/0739456X231180568

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review, 133(5), 1155-1174. https://doi.org/10.1175/MWR2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92(437), 179-191. https://doi.org/10.2307/2291462

Rocha, A. and Fernandes, E. (2011). Electromagnetism-like augmented lagrangian algorithm for global optimization. In Computational Science and Its Applications -- ICCSA 2011 (pp. 415-425). Springer. https://doi.org/10.1007/978-3-642-20505-7_37

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., & Tarantola, S. (2008). Global Sensitivity Analysis: The Primer. John Wiley & Sons.

Saltelli, A., Tarantola, S., Campolongo, F., & Ratto, M. (2002). Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models. John Wiley & Sons.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Mathematics and Computers in Simulation, 55(1-3), 271-280.

Solomatine, D. and Shrestha, D. (2009). A novel method to estimate model uncertainty using machine learning techniques. Water Resources Research, 45(12). https://doi.org/10.1029/2008wr006839

Solomatine, D., Maskey, M., & Shrestha, D. (2007). Instance-based learning compared to other data-driven methods in hydrological forecasting. Hydrological Processes, 22(2), 275-287. https://doi.org/10.1002/hyp.6592

Stralberg, D., Carroll, C., & Nielsen, S. E. (2020). Toward a climate-informed North American protected areas network: Incorporating climate-change refugia and corridors in conservation planning. Conservation Letters, 13(4), e12712. https://doi.org/10.1111/conl.12712

Tang, W., & Bennett, D. A. (2010). Agent-based Modeling of Animal Movement: A Review. Geography Compass, 4(7), 682-700. https://doi.org/10.1111/j.1749-8198.2010.00337.x

Tosserams, S., Etman, L., & Rooda, J. (2006). An augmented lagrangian decomposition method for quasi-separable problems in mdo. 10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference. https://doi.org/10.2514/6.2006-7112

U.S. Department of the Interior, Bureau of Reclamation. (2011). Quality of Water Colorado River Basin (Report No. 23).

Vadas, P., Gburek, W., Sharpley, A., Kleinman, P., Moore, P., Cabrera, M., ... & Harmel, R. (2007). A model for phosphorus transformation and runoff loss for surface-applied manures. Journal of Environmental Quality, 36(1), 324-332. https://doi.org/10.2134/jeq2006.0213

Velázquez, J., Anctil, F., Ramos, M., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? a study on 29 french catchments using 16 hydrological model structures. Advances in Geosciences, 29, 33-42. https://doi.org/10.5194/adgeo-29-33-2011

Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. International Journal of Applied Earth Observation and Geoinformation, 121, 103364. https://doi.org/10.1016/j.jag.2023.103364

White, M., Arnold, J. G., Bieger, K., Allen, P. M., Gao, J., Čerkasova, N., Gambone, M., Park, S., Bosch, D. D., Yen, H., & Osorio, J. M. (2022). Development of a Field Scale SWAT+ Modeling Framework for the Contiguous U.S. JAWRA Journal of the American Water Resources Association, 58(6), 1545-1560. https://doi.org/10.1111/1752-1688.13056

Wollheim, W., Peterson, B., Thomas, S., Hopkinson, C., & Vörösmarty, C. (2008). Dynamics of n removal over annual time periods in a suburban river network. Journal of Geophysical Research: Biogeosciences, 113(G3). https://doi.org/10.1029/2007jg000660

Yaagoubi, R., Lakber, C.-E., & Miky, Y. (2024). A comparative analysis on the use of a cellular automata Markov chain versus a convolutional LSTM model in forecasting urban growth using sentinel 2A images. Journal of Land Use Science, 19(1), 258-277. https://doi.org/10.1080/1747423X.2024.2403789

Yang, Y., Sun, H., Xue, J., Liu, Y., Liu, L., Yan, D., & Gui, D. (2021). Estimating evapotranspiration by coupling Bayesian model averaging methods with machine learning algorithms. Environmental Monitoring and Assessment, 193(3), 156. https://doi.org/10.1007/s10661-021-08934-1

Zhang, X., Srinivasan, R., & Liew, M. (2009). Approximating swat model using artificial neural network and support vector machine. JAWRA Journal of the American Water Resources Association, 45(2), 460-474. https://doi.org/10.1111/j.1752-1688.2009.00302.x

# CHAPTER 2 HARNESSING ENSEMBLE MACHINE LEARNING MODELS FOR IMPROVED SALINITY PREDICTION IN LARGE RIVER BASIN SCALES[1]


## ABSTRACT

This study develops a robust ensemble machine learning methodology for predicting average annual salinity by combining multiple machine learning algorithms. Salt concentration is a crucial water quality indicator, and salinity issues cost $300 million annually in the U.S. Irrigated agricultural lands in the Upper Colorado River Basin contribute excessively to dissolved solid loads despite covering less than 2% of the basin area. The economic impact and complex relationship between irrigation practices, groundwater dynamics, and salinity levels necessitate improved predictive capabilities at river basin scales. Using twenty years of data from 150 watersheds, eleven machine learning algorithms were evaluated through both random and spatial cross-validation approaches, with Extreme Gradient Boosting, Gradient Boosting, and Random Forest emerging as top performers. Bayesian Model Averaging and stacked generalization were employed to create ensemble models, demonstrating enhanced performance validity. The BMA ensemble achieved better spatial generalization compared to individual models while requiring significantly less computational resources than stacking. Model uncertainty analysis revealed that BMA provided the most stable predictions among all approaches. Soil electrical conductivity and calcium carbonate content emerged as the most important predictors, followed by river flow. The


[1] Mohamed F. Mahmoud, Mazdak Arabi, Shrideep Pallickara, "Harnessing ensemble Machine learning models for improved salinity prediction in large river basin scales," *Journal of Hydrology*, Volume 652, 2025, 132691. https://doi.org/10.1016/j.jhydrol.2025.132691

resulting spatially distributed predictions revealed distinct patterns in sulfate loads and concentrations across sub-basins, providing insights for targeted salinity management. This study demonstrates the effectiveness of ensemble machine learning approaches for robust salinity prediction while highlighting the importance of comprehensive uncertainty assessment and spatial validation in environmental modeling applications.

## 2.1 INTRODUCTION

Salinity, the concentration of dissolved salts in water, is a critical water quality parameter derived from natural geological processes such as rock and soil weathering and erosion. In the United States, salinity-related damages amounted to approximately $383 million per year based on 2009 concentrations, with this annual cost projected to escalate to $671 million by 2040 without decisive measures (Colorado River Basin Salinity Control Forum, 2023; U.S. Department of the Interior, 2011). The Upper Colorado River Basin faces particular challenges, where irrigated agricultural lands, despite occupying less than 2% of the basin area, contribute 30-45% of dissolved solid loads (Kenney, 2009; Miller et al., 2017). This disproportionate impact highlights the complex interplay between land use, hydrology, and salinity dynamics.

Understanding and predicting salinity levels in river basins has traditionally followed two main approaches: process-based modeling and statistical analysis. Process-based models simulate the underlying physical mechanisms of salt transport, including surface soil erosion and weathering processes, subsurface dissolution in groundwater systems, agricultural irrigation effects on salt mobilization, and baseflow contributions to salinity loads (Hyatt et al., 1970; Lee et al., 1993; Venkatesan et al., 2011). These models offer the advantage of representing the underlying processes and mechanisms that govern salt transport, allowing for a more comprehensive

understanding of the system. However, these physically based models face significant challenges. They require extensive data inputs that are often unavailable at large river basin scales. Model calibration becomes increasingly complex for large watersheds, while computational demands can be prohibitive for basin-wide applications. Furthermore, uncertainty in parameter estimation can significantly affect model reliability.

Given these limitations, there is growing interest in data-driven approaches that can effectively capture salinity dynamics without explicitly modeling all physical processes. Machine learning (ML) offers several advantages for large-scale salinity prediction. These methods can identify complex patterns from diverse data sources while reducing reliance on detailed physical parameters. ML approaches provide computational efficiency for large-scale applications and demonstrate a remarkable capacity to integrate multiple types of predictor variables (Tillman et al., 2019).

Recent studies have demonstrated ML's potential for water quality modeling. For example, Random Forest models have been developed to quantify salt levels (Tillman et al., 2018) and the extent of exposed bare ground derived from remote sensing data (Nauman et al., 2019). ML techniques can outperform conventional statistical methods like linear or nonlinear regressions for modeling complex, nonlinear relationships between variables (Kratzert et al., 2019). It can automatically learn and extract intricate patterns from large datasets, handle nonlinear relationships, and incorporate a wide range of predictor variables without requiring explicit assumptions about the underlying data distribution (Kratzert et al., 2019).

While ML techniques have shown promise in modeling hydrologic and water quality processes, individual ML algorithms may have limitations in capturing the full complexity of these systems. Ensemble learning, which combines predictions from multiple models, has emerged as a powerful

approach to improve performance and reduce uncertainty (Hoeting et al., 1999; Raftery et al., 1997). While previous studies have demonstrated the utility of individual ML techniques like Random Forest for modeling salinity (Tillman et al., 2018; Nauman et al., 2019), several important research gaps remain. First, no comprehensive comparison exists of multiple ML algorithms' performance for salinity prediction, making it difficult to determine the optimal modeling approach. Second, individual ML models, including Random Forest, can be prone to overfitting and instability when applied to complex environmental systems (Domingos, 2000; Pacifico, 2021). Third, the potential benefits of combining multiple ML algorithms through ensemble techniques like Bayesian Model Averaging (Hoeting et al., 1999; Raftery et al., 1997) and stacked generalization (Cao et al., 2022; Naimi & Balzer, 2018) have not been explored for salinity prediction. These advanced ensemble methods could potentially improve prediction accuracy and robustness beyond what any single algorithm can achieve by leveraging the complementary strengths of different models while mitigating their individual weaknesses (Ren et al., 2016; Zhang & Ma, 2012).

This study develops a robust ensemble ML methodology to investigate salt levels in rivers and streams using hydroclimatic, physiographic, and socioeconomic characteristics of contributing watersheds. The objectives of the study are to: 1) evaluate the performance of various ML algorithms in predicting salinity levels in the Colorado River Basin (CRB). ; 2) develop and compare two ensemble techniques, Bayesian Model Averaging (BMA) and Stacked Generalization, to improve the accuracy and robustness of salinity predictions by combining multiple ML models; and 3) identify the most important watershed characteristics driving salinity levels in the CRB using the developed ensemble models. The framework incorporates alternative ML algorithms using robust ensemble techniques to improve performance validity and reduce

modeling uncertainty. Specifically, BMA and Stacked Generalization are used to combine predictions from high-variance models like decision trees to address model input, parametric, and structural uncertainty.

The ensembles provide more reliable salinity predictions compared to individual models by optimally blending multiple models to improve stability and generalization. This approach enables deeper insights into the drivers of this complex phenomenon and the identification of the most effective modeling strategies. The enhanced predictive capabilities demonstrate ML's utility for understanding and managing salinity levels in large complex river basins like the CRB. More accurate forecasts inform intervention strategies to mitigate salinity's environmental and economic impacts in the region.

This chapter addresses three objectives: (1) evaluate and compare multiple machine learning algorithms for predicting salinity and nutrient loads from natural landscapes across diverse watershed conditions; (2) develop ensemble methods using Bayesian Model Averaging and Stacked Generalization that improve prediction robustness and quantify uncertainty compared to individual algorithms; and (3) identify dominant environmental drivers of nutrient export through systematic feature importance analysis across algorithmically diverse models.

## 2.2 MATERIALS AND METHODS

Eleven ML methods are selected for the analysis. First, the most influential input variables are identified using redundancy analysis to minimize multicollinearity and sensitivity analysis to determine influential predictors. Eleven regression algorithms were evaluated through cross-validation for model optimization and validation. Two distinct ensemble strategies were employed: BMA and stacked ensembles to integrate multiple top-performing models. The performance of

these ensemble techniques was compared to identify the most robust method for salinity prediction modeling. Figure 2- 1 is summarizing the key steps in developing the prediction model.

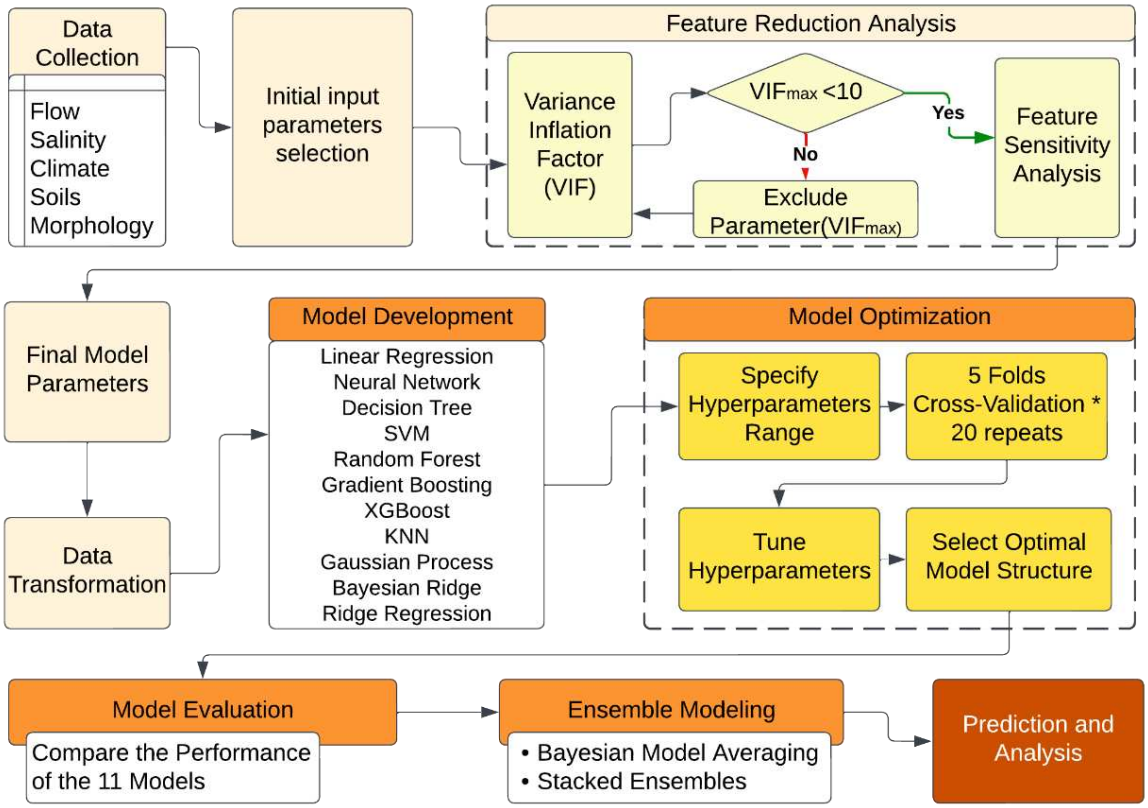


Figure 2- 1 Workflow Diagram Outlining the Key Steps in Developing the Salinity Prediction Model

### 2.2.1 INPUT VARIABLES

The modeling framework utilizes salinity, flow, discharge, meteorological, and groundwater data. Meteorological data encompassed precipitation, minimum and maximum temperature, mean temperature, dewpoint, and maximum and minimum annual vapor pressure deficit. Soil data included average curve number, surface electrical conductivity (EC), and calcium carbonate (CaCO3) content.

Twenty-four input variables were initially considered. Using variance inflation factor analysis, these were reduced to minimize redundancy. Feature sensitivity analysis was used to investigate the significance of each feature in predicting the target variable and further reduce input variables.

### 2.2.2 MODEL DEVELOPMENT

This study compared multiple modeling techniques for predicting hydrological processes in the CRB. The selection of eleven ML methods was driven by the need to represent diverse algorithmic families and capture various aspects of the underlying relationships in the data. Linear methods, including Multiple Linear Regression and Ridge Regression, were selected as baseline models. Non-linear methods such as Neural Networks, Decision Trees, and Support Vector Machines were included to handle complex relationships. Tree-based aggregation methods including Random Forest, Gradient Boosting, and XGBoost were chosen for their ability to combine multiple decision trees into robust predictive models. Instance-based methods represented by K-Nearest Neighbors were selected to capture local patterns, while probabilistic methods such as Gaussian Process and Bayesian Ridge were included to account for uncertainty in predictions.

Multiple Linear Regression (MLR) is a widely used statistical method for modeling relationships between continuous response variables and multiple predictor variables (Eberly, 2007; Uyanık & Güler, 2013). Deep Neural Networks (DNN) are artificial neural networks with multiple hidden layers that enable the model to learn complex patterns from input data, with applications in various fields such as image recognition, natural language processing, and hydrological modeling (Canziani et al., 2017; Miikkulainen et al., 2019; Samek et al., 2021). Decision Trees (DT) recursively split input data based on predictor variables (Kingsford & Salzberg, 2008; Kotsiantis, 2013; Rokach & Maimon, 2005), while SVMs aim to find the optimal hyperplane that separates

data into different classes or predicts continuous target variables (Mahesh, 2018; Marjanović et al., 2011).

Tree-based methods including Random Forest (RF) and Gradient Boosting (GB) construct multiple decision trees and combine their predictions to improve generalization and reduce overfitting (Cai et al., 2020; Cootes et al., 2012; Rodriguez-Galiano et al., 2015). Extreme Gradient Boosting (XGB), an optimized implementation of GB, offers additional features and improvements, gaining popularity due to its strong performance in ML competitions (Ibrahem Ahmed Osman et al., 2021; Yu et al., 2020). KNN, a non-parametric and instance-based learning algorithm, predicts target variables based on the values of k-nearest neighbors in the feature space (Larose & Larose, 2014; B. Sun et al., 2009; S. Sun & Huang, 2010). Finally, the Gaussian Process (GP), Bayesian Ridge (BR), and Ridge Regression are regression techniques that incorporate Bayesian non-parametric methods, prior information about model parameters, and regularization techniques to provide more robust estimates and mitigate the effects of multicollinearity (Burnaev & Vovk, 2014; Hoerl et al., 1975; Ranjan et al., 2016). A comparative analysis of the advantages and disadvantages of each method is presented in Table 2- 1.

Table 2- 1 Comparison Between the Different Machine Learning Algorithms Used in The Study.

| MID | Model | Type | Pros | Cons |
|---|---|---|---|---|
| M01 | Linear Regression | Linear | Fast, interpretable, simple | Sensitive to outliers, assumes a linear relationship |
| M02 | Neural Network Regression | Non-linear | Powerful, can capture complex relationships, handles high-dimensional data | Computationally expensive, requires careful selection of architecture and regularization parameters |
| M03 | Decision Tree Regression | Non-linear | Can capture non-linear relationships, handles interactions and categorical features | Prone to overfitting, instability, and bias |

| MID | Model | Type | Pros | Cons |
|---|---|---|---|---|
| M04 | Support Vector Regression | Non-linear | Flexible, can handle high-dimensional data, effective for smaller datasets | Slow, requires careful selection of hyperparameters, sensitive to feature scaling |
| M05 | Random Forest Regression | Tree-based | Robust, can handle complex data, does not overfit easily, can identify feature importance | Less interpretable than a single tree, and difficult to fine-tune hyperparameters |
| M06 | Gradient Boosting Regression | Tree-based | Powerful, handles complex data, can capture non-linear relationships, can identify feature importance | Sensitive to overfitting and requires careful selection of hyperparameters |
| M07 | Extreme Gradient Boosting | Tree-based | Can handle missing data, faster and more scalable than gradient boosting, can identify feature importance | Requires careful tuning of hyperparameters, can overfit easily |
| M08 | K-Nearest Neighbors | Non-parametric | Simple, works well with small datasets, non-linear, does not make assumptions about data distribution | Computationally expensive, sensitive to feature scaling, requires careful selection of hyperparameters |
| M09 | Gaussian Process Regression | Bayesian non-parametric | Handles non-linear relationships, flexible, provides a measure of uncertainty | Computationally expensive, does not scale well to larger datasets, sensitive to kernel selection |
| M10 | Bayesian Regression | Bayesian parametric | Handles uncertainty and noise well, flexible, can incorporate prior knowledge | Limited to linear relationships, assumes normality of errors |
| M11 | Ridge Regression | Linear | Regularizes coefficients to prevent overfitting, works well with multicollinear data | Cannot eliminate unimportant features, requires careful selection of regularization parameter |

The salinity training data exhibited a positively skewed, lognormal distribution with a long right tail. Transforming the data can improve model performance, as most statistical and ML techniques assume normality. Normalization reduces the influence of extreme values, strengthening linearity, homoscedasticity, and independence assumptions for enhanced accuracy (Bhanja & Das, 2018; Ha et al., 2021; More & Wolkersdorfer, 2023). To address this, a log transformation was applied to normalize the data before modeling. When transforming predictions back to the original scale, a bias correction factor (BCF) was implemented to account for the systematic underestimation that occurs during back-transformation due to Jensen's inequality (Strimbu, 2012).

$$\mathrm{BCF} = e^{\frac{\sigma^2}{2}} \tag{2-1}$$

where $\sigma^2$ represents the variance of the residuals in log space. The final predictions were obtained by first back-transforming the log-scale predictions and then applying the bias correction.

### 2.2.3 MODEL ASSESSMENT

The evaluation of ML models followed a three-stage assessment strategy. First, individual models were tested using a single 80-20% data split, mirroring the conventional calibration-validation framework commonly used in physical modeling. This approach, where a portion of available data is used for model calibration and the remaining data serves as an independent validation set, allows direct comparison with traditional modeling practices and identifies the best achievable performance under optimal conditions for each algorithm.

To assess model generalizability and provide a more robust evaluation, the models were then evaluated using 5-fold random cross-validation repeated twenty times. This approach randomly splits data into k folds, training on k-1 and testing on the holdout fold, repeated across all folds (A.

Ramezan et al., 2019; Berrar, 2019; Tougui et al., 2021). By testing models across 100 different data splits, this method provides a more comprehensive understanding of model stability and reliability. Finally, the best-performing individual models and their ensembles were subjected to spatial cross-validation to evaluate their geographic generalizability. In this approach, watersheds were clustered based on geographic coordinates using Euclidean distances, ensuring test and training sets represented distinct geographic regions (Wang et al., 2023).

For each model, initial ranges for key hyperparameters were defined, and then iteratively tuned by assessing performance over a 5-fold cross-validation repeated twenty times. The optimal hyperparameter set was selected for each model based on the best-achieved median and 10$^{th}$-percentile $R^2$ values.

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (2\text{-}2)$$

Where $y_i$ is the observed value, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean of the observed values. Performance metrics included the median, range, and 10th percentile of $R^2$ values across the repetitions, reflecting overall performance, variability across splits, and conservative estimates of worst-case behavior. Model uncertainty was assessed using the distribution of $R^2$ values across cross-validation iterations, examining median performance, robustness (10th percentile), and peak performance (90th percentile). This validation methodology identified the top model considering both predictive accuracy and reliability. The embedded hyperparameter optimization further enhanced model stability and reliability.

In addition to $R^2$ metrics, information criteria was employed to assess model performance while accounting for model complexity. Specifically, Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) (Chakrabarti & Ghosh, 2011; Hauenstein et al., 2018).

$$\text{AIC} = n * \ln\left(\frac{RSS}{n}\right) + 2k \quad (2\text{-}3)$$

$$BIC = n * \ln\left(\frac{RSS}{n}\right) + k * \ln(n) \quad (2\text{-}4)$$

where n is the number of observations, RSS is the residual sum of squares, and k is the number of model parameters. Lower AIC and BIC values indicate better model performance when accounting for complexity.

### 2.2.4 ENSEMBLE LEARNING TECHNIQUES

BMA and stacked ensembles were utilized to combine the top-performing models while accounting for uncertainty in model selection and leveraging the diversity across models. BMA assigns posterior probabilities to each model based on how well the data supports it (Ley & Steel, 2009; Raftery et al., 1997; Yang et al., 2021). The BMA approach considers all candidate models and calculates the weighted average of their predictions, with the weights given by their respective posterior probabilities. The posterior probability can be calculated based on the prior probability of the model and the data likelihood given by the model.

The prior probability $P(M_i)$ reflects any prior knowledge about the models before observing the data. It can be specified based on factors like model complexity or performance in previous studies. In the absence of strong prior knowledge, a common assumption is to set a uniform prior (Hoeting et al., 1999). This assigns an equal prior probability to each model (1/Number of models).

The likelihood $P(D|M_i)$ quantifies how well model $M_i$ fits the observed data D. For regression models, a common assumption is that the residuals follow a normal distribution (Raftery et al., 2005):

$$\ln\left(P(D|M_i)\right) = \sum_n [\ln\left(\frac{1}{\sqrt{(2*\pi)}\sigma_i}\right) + \left(-\frac{1}{2}\frac{(y_n - \hat{y}_{i,n})^2}{{\sigma_i}^2}\right)] \quad (2\text{-}5)$$

where n is the number of observations, $\sigma_i$ is the residual standard deviation for model $M_i$, $y_n$ is the $n^{th}$ observation, and $\hat{y}_{i,n}$ is the prediction from model $M_i$ for observation n.

The posterior probability of a model $M_i$ can be computed using Bayes' theorem as follows:

$$P(M_i \mid D) = (P(D \mid M_i) * P(M_i)) / \Sigma(P(D \mid M_j) * P(M_j)) \quad (2\text{-}6)$$

where $P(M_i \mid D)$ is the posterior probability of model $M_i$ given the data D, $P(D \mid M_i)$ is the likelihood of the data D given the model $M_i$, $P(M_i)$ is the prior probability of model $M_i$, $P(D \mid M_j)$ and $P(M_j)$ are the likelihood and prior probability of model $M_j$, respectively, and the summation is over all candidate models.

The BMA predictions can then be computed as:

$$\hat{y}_{BMA} = \Sigma(P(M_i \mid D) * \hat{y}_i) \quad (2\text{-}7)$$

where $\hat{y}_{BMA}$ is the BMA prediction, $P(M_i \mid D)$ is the posterior probability of model $M_i$, and $\hat{y}_i$ is the prediction of model $M_i$.

The posterior model probabilities can be normalized to sum to 1 to give model weights (Duan et al., 2007):

$$w_i = P(M_i|D) / \sum_j P\left(M_j\middle|D\right) \quad (2\text{-}8)$$

where $w_i$ is the weight for model $M_i$ in the BMA ensemble.

The prediction variance for the BMA ensemble is given by (Krishnamurti et al., 1999):

$$VAR(\hat{y}_{BMA}) = \sum_i w_i^2 * VAR(\hat{y}_i) + \sum_i w_i * (\mu_i - \sum_j w_j * \mu_j)^2 \quad (2\text{-}9)$$

where $\mu_i$ and VAR($\hat{y}_i$) are the mean and variance of model $M_i$'s predictions.

Stacked ensembles as another ensemble technique was used to integrate multiple ML models. This ensemble technique involves training a secondary "meta-model" using the predictions from several base models as features (Cao et al., 2022; Naimi & Balzer, 2018). The top three models based on cross-validated performance were selected as the base learners. These models were trained using 5-fold cross-validation on the full dataset. The out-of-fold predictions from the base models were combined into a new feature set used to train a linear regression meta-model. Linear regression was chosen to reduce overfitting risks compared to more flexible ML algorithms. The resulting stacked ensemble integrates the strengths of the base models while avoiding overfitting. Performance was also evaluated using 5-fold cross-validation.

BMA and stacked ensemble techniques provide two key advantages - enhanced predictive performance and reduced model uncertainty. By integrating multiple models, ensemble modeling leverages the strengths of each while mitigating their individual weaknesses (Domingos, 2000; Pacifico, 2021). These methods account for uncertainty in the model selection process itself, improving generalization and avoiding overfitting that can occur with a single model. The BMA and stacked ensembles developed here combined top-performing models to obtain accurate and robust salinity predictions for the CRB. Rather than relying on any individual model, the probabilistically weighted averaging provides stability and protects against overfitting by incorporating information from several candidates. Overall, BMA and stacking allowed the construction of an optimized ensemble model to maximize predictive capability while accounting for model uncertainty.

### 2.2.5 STUDY AREA

The CRB encompasses approximately 650,000 km$^2$ across seven western U.S. states: Colorado, Wyoming, Utah, New Mexico, Nevada, Arizona, and California. Elevations range from 6 m above sea level in the southwest to over 4,300 m in the Rocky Mountains (*USGS.Gov*, 2021). The climate is characterized by hot, dry summers with temperatures from the mid-20°C to over 37°C, and cold winters with lows from negative single digits to 5°C (Christensen et al., 2004). Average annual precipitation varies from 152-203 mm in lower elevations to over 2,000 mm in high peaks, depending on location and elevation (Kalra & Ahmad, 2011). Overall, the CRB exhibits diverse climatic and topographic conditions across a vast geographic area in the western United States.

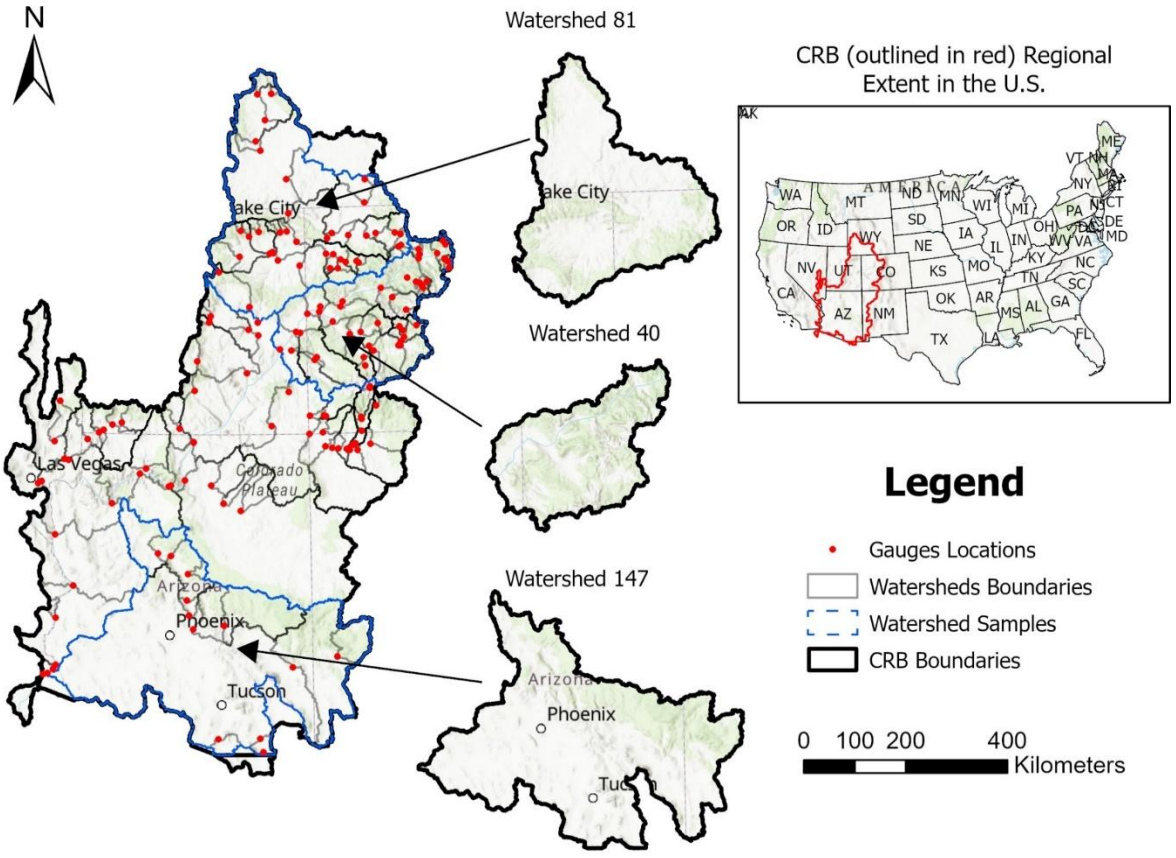


Figure 2- 2 The USGS Gauges Locations and Delineated Watershed Boundaries Within the CRB Highlighting Three Exemplary Watersheds in The Basin

#### 2.2.5.1 INPUT DATASETS

A comprehensive dataset consisting of 24 predictor variables and one target variable was compiled to develop the salinity prediction model (Table 2- 2). The dataset integrates multiple sources of information including hydrologic measurements, climate data, soil characteristics, land use patterns, and topographic features across the CRB from 2000-2020.

Table 2- 2 Summary of input and output variables used in the salinity prediction model.

| Category | Variable | Unit | Source | Resolution/Format |
|---|---|---|---|---|
| Target Variable | Sulfate Concentration | mg/L | USGS | Daily aggregated to annual |
| Watershed Characteristics | Watershed Area | $km^2$ | USGS DEM | 30m |
| | WWTF Discharge | $m^3/s$ | CDPHE Reg. 85 | Point source |
| | Average Flow | $m^3/s$ | USGS Gauges | Daily aggregated to annual |
| Soil Properties | Soil Electrical Conductivity | dS/m | SSURGO Database | Vector polygons |
| | Curve Number | - | Global CN Dataset | 250m |
| | Soil Calcium Carbonate | % | SSURGO Database | Vector polygons |
| | Groundwater Depth | m | USGS Wells | Point measurements |
| Topographic | Elevation | m | USGS DEM | 30m |
| Climate | Precipitation | m/yr | PRISM | 4km |
| | Minimum Temperature | °C | PRISM | 4km |
| | Mean Temperature | °C | PRISM | 4km |
| | Maximum Temperature | °C | PRISM | 4km |
| | Dew Point Temperature | °C | PRISM | 4km |
| | Min Vapor Pressure Deficit | hPa | PRISM | 4km |

| Category | Variable | Unit | Source | Resolution/Format |
|---|---|---|---|---|
| | Max Vapor Pressure Deficit | hPa | PRISM | 4km |
| Land Use | Forest Cover | % | NLCD | 30m |
| | Grassland | % | NLCD | 30m |
| | Shrubland | % | NLCD | 30m |
| | Developed Land | % | NLCD | 30m |
| | Cropland | % | NLCD | 30m |
| | Agricultural Area | $km^2$ | CDSS | Vector polygons |
| | Impervious Surface | % | NLCD | 30m |
| Location | Latitude | °N | USGS/EPA | Point location |
| | Longitude | °W | USGS/EPA | Point location |

Flow, salinity, and groundwater data were collected from the United States Geological Survey (USGS) monitoring network (USGS, 2021). Approximately 150 gauges providing streamflow and water quality data, along with five thousand wells recording groundwater levels, were selected to ensure comprehensive basin coverage. Daily discharge measurements and sulfate concentrations from these gauges were aggregated to calculate annual averages for each location. Sulfate concentration was selected as the target variable and indicator of total salinity due to its consistent measurement across all monitoring sites (Figure 2- 2).

Gridded meteorological data was obtained from the PRISM (Parameter-elevation Regressions on Independent Slopes Model) dataset at 4-kilometer resolution. PRISM incorporates point measurements and factors like elevation and terrain to generate high-resolution spatial climate data (Rupp et al., 2022). The climate variables include annual averages of precipitation, temperature metrics (minimum, maximum, and mean), dewpoint temperature, and vapor pressure deficit

(minimum and maximum). PRISM data were chosen for their reliability and consistent coverage across the entire basin (PRISM Climate Group, 2014).

Land use information was derived from the USGS National Land Cover Database (NLCD) at 30-meter resolution. The land cover classes were grouped into five main categories: forest (including deciduous, evergreen, and mixed forests), grassland, shrubland, developed land, and cropland. Additionally, watershed imperviousness was calculated from the NLCD impervious surface dataset. Irrigated agricultural areas were specifically obtained from the Colorado Decision Support System (CDSS) to distinguish irrigation from non-irrigated cropland.

Soil characteristics were primarily sourced from the USDA's Soil Survey Geographic (SSURGO) database (Soil Survey Staff et al., 2022), supplemented with predictive mapping where data gaps existed (Nauman & Duniway, 2020). Key soil parameters included electrical conductivity (EC) and calcium carbonate (CaCO3) content. Average curve number for each watershed was derived from global gridded maps (Jaafar et al., 2019) to estimate runoff potential.

Catchments for selected USGS gauges were delineated using a digital elevation model (DEM) to establish watershed boundaries. DEMs represent topography and are used to derive features like slope and drainage (Guth et al., 2021). Watershed characteristics such as area and average elevation were calculated using GIS software and were subsequently used as input parameters for hydrological modeling. Additionally, average discharge data from wastewater treatment facilities (WWTFs) in each watershed was obtained from the EPA Enforcement and Compliance History Online database (ECHO, 2022). Total WWTF discharge was allocated to its respective watershed for analysis. The DEM enabled watershed delineation and derivation of morphological variables influencing salinity, while WWTF data provided anthropogenic input information.

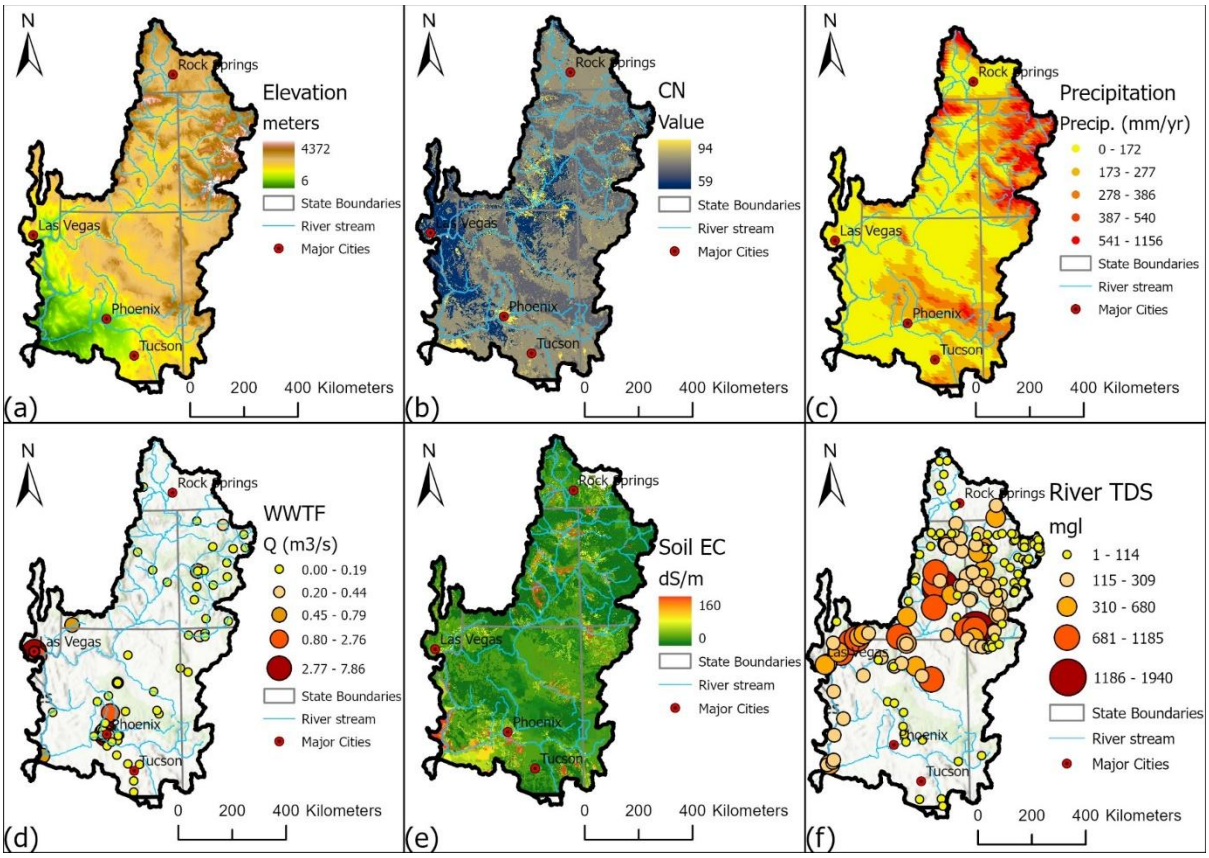


Figure 2- 3 Sample Data of The Study Area of Colorado Rive Basin: (a) Elevation Raster Map Using the Digital Elevation Model (DEM); (b) Curve Number (C.N.) Values Map; (c) Precipitation Distribution in The Basin; (d) WWTF Discharge Rates; (e) Soil E.C. Raster Map for The Basin; (f) Average Annual River Salinity Data.

### 2.2.5.2 REDUNDANCY ANALYSIS

The Variance Inflation Factor (VIF) method is a widely used statistical approach for identifying multicollinearity in multiple regression analysis (Akinwande et al., 2015; Cheng et al., 2022; Craney & Surles, 2002; Miles, 2014; Vu et al., 2015). Multicollinearity occurs when predictor variables are highly correlated, which can lead to unreliable and unstable estimates of regression coefficients (Tamura et al., 2019; Thompson et al., 2017). The VIF method reduces redundancy among input parameters by quantifying multicollinearity, retaining only the most independent, relevant variables for the salinity prediction model. The VIF for each predictor variable is

calculated based on the coefficient of determination ($R_i^2$) of the regression model ($VIF_i = \frac{1}{1-R_i^2}$) when the $i^{th}$ predictor variable is regressed against all other predictor variables. Higher VIF values indicate greater multicollinearity, as the variance of the coefficient is inflated for correlated variables. Although thresholds vary, a VIF exceeding 10 generally signifies high collinearity (O'brien, 2007; Snee, 1983). The VIF was computed for the twenty-four initial inputs. This resulted in a final set of fourteen inputs (Figure 2- 4). Applying VIF for feature reduction minimized multicollinearity effects, promoting reliable and interpretable models less prone to overfitting.

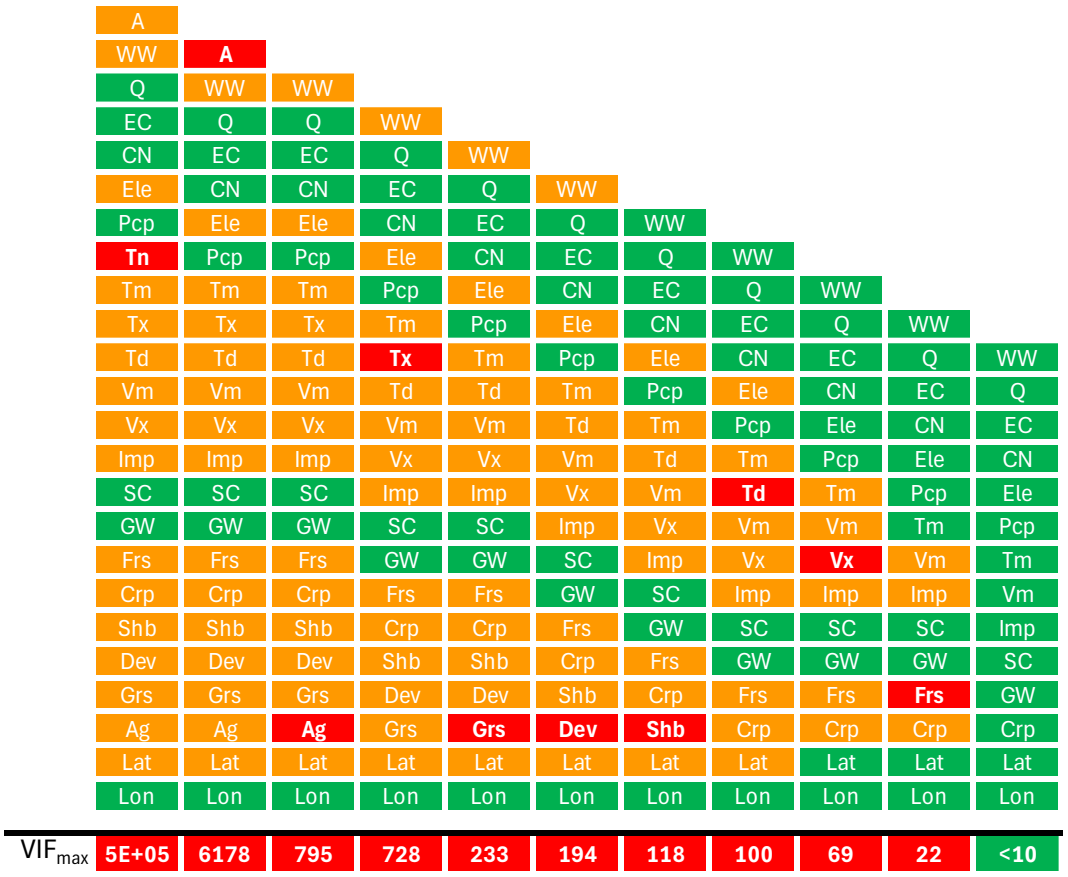


Figure 2- 4 Calculated Variance Inflation Factor (VIF) For Each of The Initial Variables in Each of The Variables Reduction Iterations. Green Color Represents a VIF Lower Than 10, Orange Color Represents a VIF higher than 10, Red Color Represents the Variable with the Maximum VIF in Each Iteration

#### 2.2.5.3 FEATURE SENSITIVITY ANALYSIS

The feature importance for predicting the target variable was calculated using the XGB, RF, and GB regression models. These models are three of the ML models that will be used in the analysis. The model calculates the feature importance by determining how often each feature appears as a split node in the constructed decision trees. To assess the importance of each feature, the "gain" importance type was utilized, which calculates the average improvement in the objective function due to each feature when used as a split node ($I_f = \frac{G_f}{N_f}$) where $I_f$ represents the importance score of a given feature f, $G_f$ represents the sum of gains associated with feature f when used as a split node, and $N_f$ represents the total number of splits involving feature f.

Based on the normalized feature scores (Figure 2- 5), Soil EC is the most critical predictor, followed by the soil CaCO3 and average flow. Conversely, the total discharge of the WWTF has the lowest importance score, suggesting that it contributes the least to the model's predictive ability. Nonetheless, given that its feature score is below the 0.1 threshold, it was excluded from the final list of parameters. These results provide valuable insights into the relative importance of each feature, helping to better understand the underlying relationships between the predictors and the target variable. This understanding can guide future research and decision-making processes by focusing on the most influential features to further improve model performance and inform policy development.

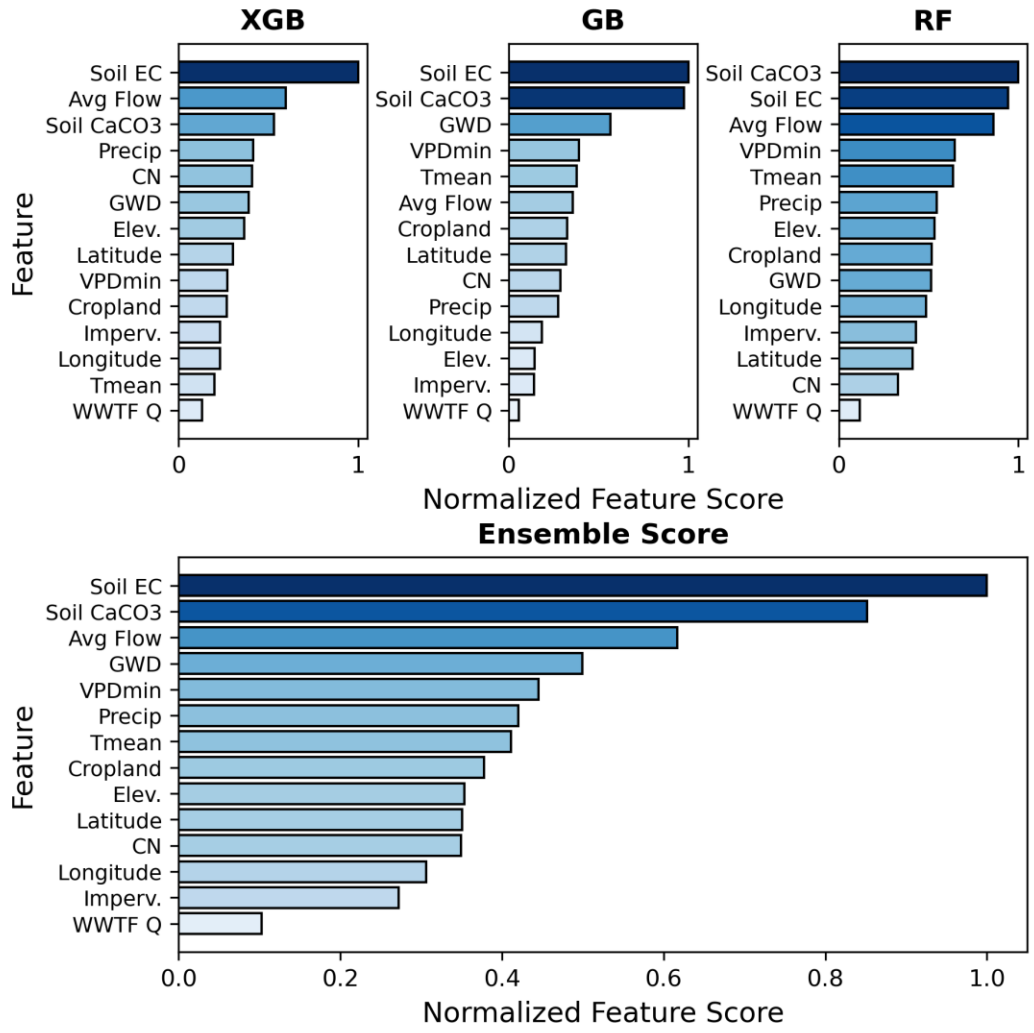


Figure 2- 5 Normalized Significance of Each Feature Used in the Top Candidate Models: Extreme Gradient Boosting (XGB), Gradient Boosting (GB), And Random Forest (RF) and the Ensemble Feature Significance Represented by Feature Score as Higher Score Implies a Higher Feature Importance.

## 2.3 RESULTS AND DISCUSSION

### 2.3.1 INDIVIDUAL MODELS' PERFORMANCE

For the first assessment strategy using a single data split, the eleven ML algorithms were evaluated for the best achievable performance mirroring the conventional calibration-validation framework commonly used in physical modeling (Figure 2- 6). Tree-based models showed superior

performance, with XGB achieving the highest test $R^2$ of 0.87, followed closely by GB ($R^2$ = 0.84). BR, RF, and KNN showed varying degrees of overfitting with a high training performance but notably lower testing performance. Linear models (Ridge, MLR) and more complex algorithms like SVM and GP showed relatively poor performance; suggesting a highly nonlinear relationship between input features and salinity levels that linear approaches or simple kernel transformations cannot adequately capture.

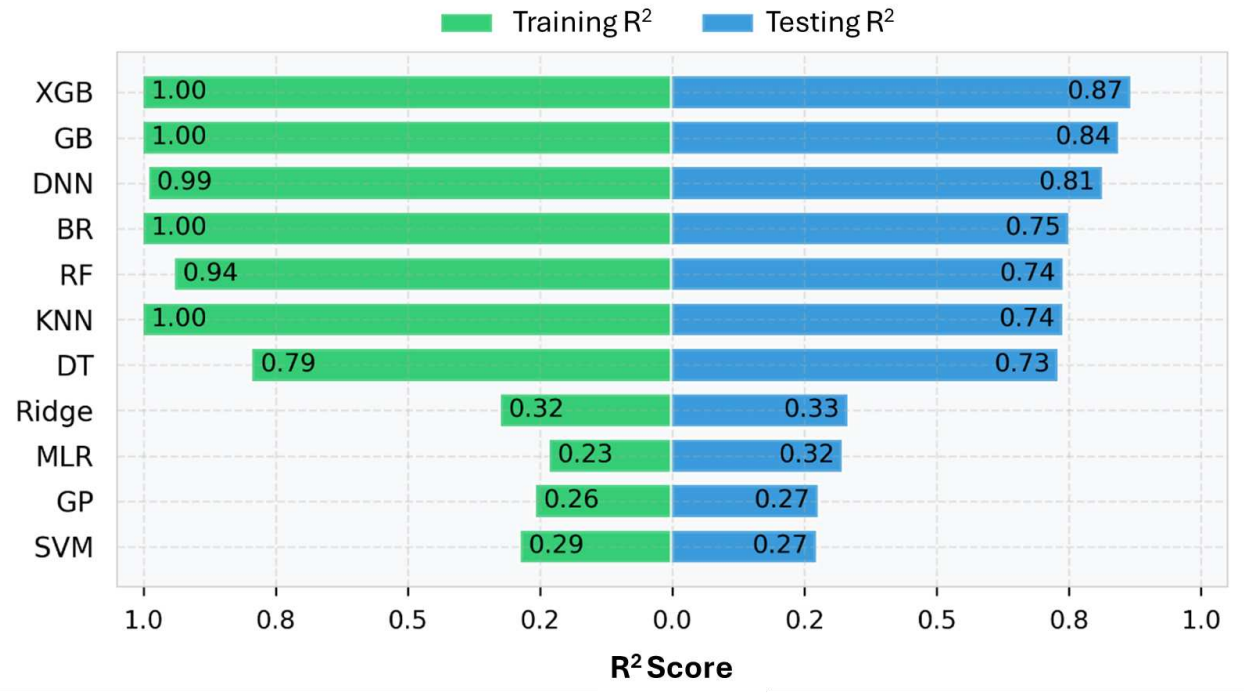


Figure 2- 6 Performance comparison of machine learning models for salinity prediction. The plot shows training (green) and testing (blue) $R^2$ scores for eleven algorithms. Models are ordered by testing performance.

To assess model generalizability and provide a more robust evaluation, the models were then evaluated using 5-fold random cross-validation repeated twenty times. A comprehensive comparison of the models' performances was made using violin plots, which effectively demonstrate the distribution of $R^2$ values, as well as key statistics such as median, maximum, minimum, and $10^{th}$ percentile $R^2$ values of the 100 data splits for each model (Figure 2- 7).

The results indicate that the XGB model outperforms the other methods in terms of median $R^2$ value (0.68), followed closely by the RF model with a median $R^2$ value of 0.62. GB also demonstrates relatively robust performance with a median $R^2$ of 0.64. Maximum and minimum $R^2$ values also give the edge for these three models. The $10^{th}$ percentile $R^2$ values provide further insight into the overall performance and robustness of the models. GB displayed the highest $10^{th}$ percentile $R^2$ value (0.49), followed by RF and XGB with values of 0.33 and 0.48, respectively. These results suggest that XGB, GB, and RF models consistently perform better than the other methods across different data scenarios.

The superior performance of the tree-based methods (XGB, RF, GB) over other approaches can be attributed to several factors. Tree-based models naturally handle non-linear relationships and interactions between variables, particularly important for salinity prediction where multiple environmental factors interact in complex ways. These methods also manage missing data effectively and are relatively robust to outliers, both common challenges in environmental monitoring data. The sequential improvement approach of XGB, combined with its regularization techniques, appears particularly well-suited to capturing the hierarchical nature of watershed processes affecting salinity.

The relatively lower performance of the DNN model compared to XGB can be attributed primarily to the limited dataset size. This aligns with findings from (Es-Sabery et al., 2021), who noted that deep learning approaches require substantial training data to achieve optimal performance. While DNNs have shown promise in hydrological modeling, their effectiveness is often constrained in situations with limited data availability.

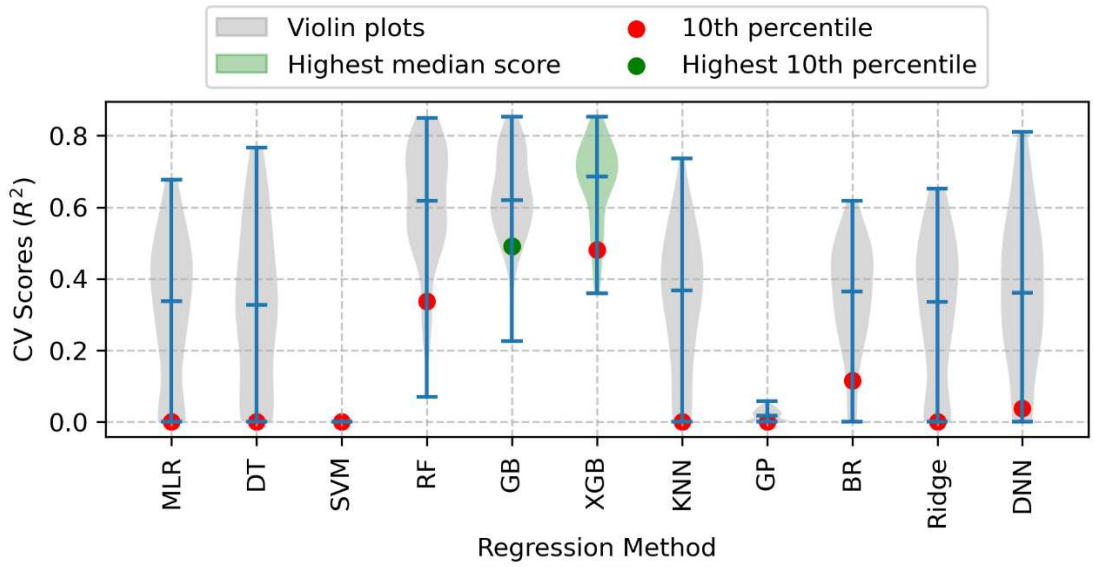


Figure 2- 7 Violin Plots Demonstrating the Performance Distribution of Each of The Tested Machine Learning Algorithms While Highlighting Max, Min, Median, And 10$^{th}$ Percentile Values for Each Model.

The log transformation and bias correction substantially improved model performance compared to the original scale modeling. The initial log transformation for the best performing model (XGB) without bias correction reduced RMSE from 87.6 to 82.6 mg/L, representing a 5.71% improvement, while $R^2$ increased from 0.605 to 0.629. The bias correction factor was calculated as 1.182, indicating moderate positive skewness in the residuals. After applying this correction, the model achieved further improvements with RMSE decreasing to 71.3 mg/L (18.61% total improvement) and $R^2$ increasing to 0.667 (10.17% total improvement). These enhancements in model performance demonstrate the effectiveness of the transformation and bias correction approach for handling the skewed distribution of salinity data.

The effectiveness of log transformation aligns with the natural behavior of environmental parameters, which often follow lognormal distributions due to multiplicative processes in natural systems. Similar transformations have proven effective in other water quality modeling studies (Ha et al., 2021; More & Wolkersdorfer, 2023). The substantial improvement after bias correction

highlights the importance of addressing systematic biases in back-transformation, a step often overlooked in environmental modeling studies.

#### 2.3.1.1 ENSEMBLE MODELS EVALUATION

BMA and stacked ensembles were employed to combine the top three performing models, namely XGB, GB, and RF, given their similar performance metrics. Using random 5-fold cross-validation, the training $R^2$ of the BMA model reached 0.99 aligning with the top three performing models as shown in Figure 2- 8, indicating that the model captures the underlying patterns in the training data exceptionally well. The stacked ensemble's training R2, while comparatively lower, cannot be directly compared with other methods due to differences in the training set. In testing performance, both the Bayesian Averaged model and the stacked model demonstrated a better performance than the best individual model (XGB). Both ensemble models have a higher median $R^2$, $10^{th}$ percentile, and $90^{th}$ percentile $R^2$ value than the XGB model, indicating that the average model performs consistently well across various scenarios (Figure 2- 9).

The superior performance of ensemble methods over individual models can be attributed to their ability to leverage complementary strengths of different algorithms. BMA's probabilistic weighting mechanism appears particularly effective in the context of salinity prediction, where different models may excel under varying hydrogeological conditions. The improvement in the 10th percentile $R^2$ values suggests that ensemble methods are especially valuable for handling challenging prediction scenarios, likely cases where input variables have relationships that are better captured by some models than others.

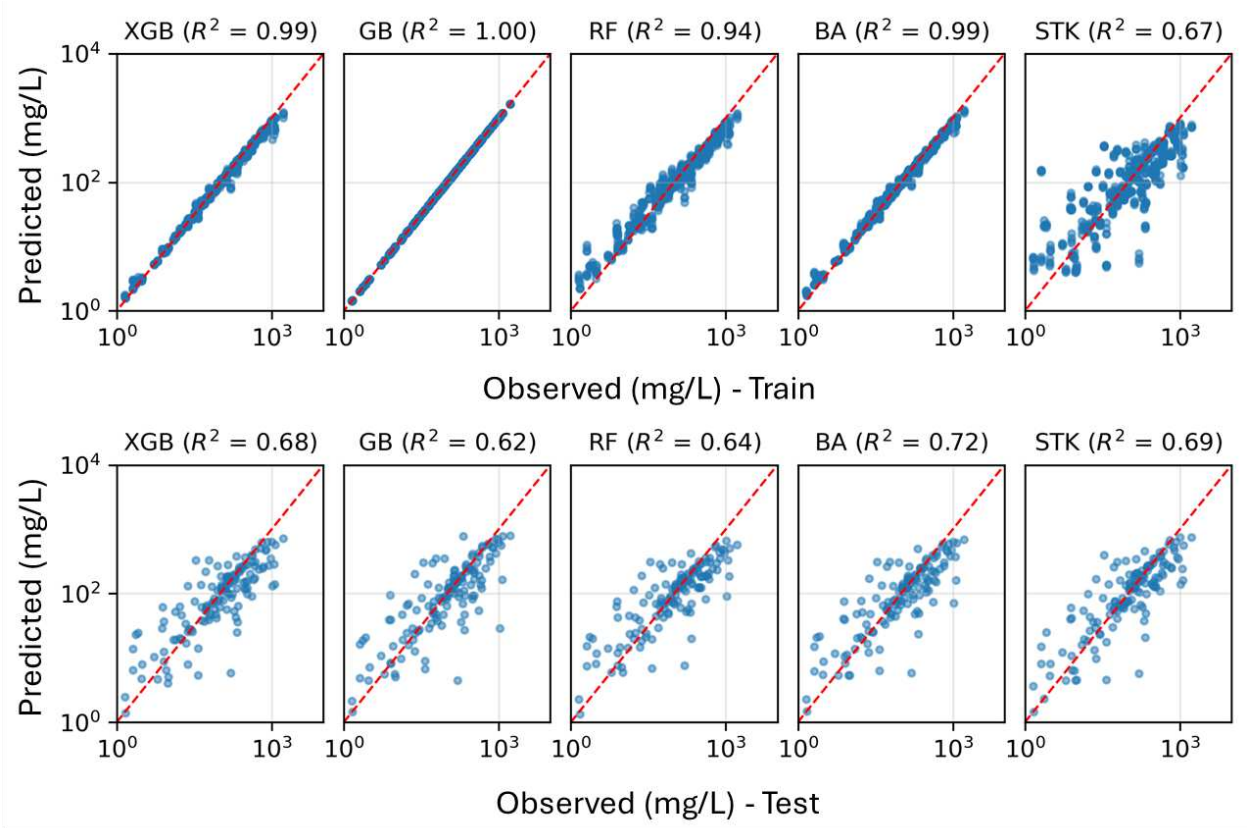


Figure 2- 8 Training Performance (Top Row) and Testing Performance (Bottom Row) of The Top Three Performing Models: Extreme Gradient Boosting (XGB), Gradient Boosting (GB), And Random Forest (RF) And the Ensemble methods: Bayesian Averaging (BA) And Stacked ensembles (STK).

The observed gap between training ($R^2$=1.0) and testing ($R^2$=0.72) performance warrants careful consideration in the context of environmental modeling. This difference can be attributed to several key factors inherent in modeling complex hydrogeological systems. Salinity transport in groundwater involves intricate, nonlinear relationships between multiple hydrogeological and meteorological variables. Recent studies modeling similar water quality parameters have reported comparable performance gaps (More & Wolkersdorfer, 2023; Sahour et al., 2020), suggesting this is a common challenge in using ML in environmental modeling. The training-testing performance differential may also reflect the spatial and temporal heterogeneity of the underlying processes. As demonstrated by (More & Wolkersdorfer, 2023; Sahour et al., 2020), model performance is notably

influenced by how well the training data captures the full range of input-output relationships. Despite this gap, the model's testing performance remains well within acceptable ranges for environmental modeling applications, with testing $R^2$ values comparable to or exceeding those reported in similar studies (0.61 to 0.89).

#### 2.3.1.2 ENSEMBLE MODELS' PERFORMANCE

When comparing the BMA and stacked ensembles, there is a clear advantage for BMA model averaging. BMA achieves a notably higher median $R^2$ (0.716) compared to stacking (0.689), demonstrating stronger overall predictive performance. This advantage extends to the robustness of predictions, with BMA showing superior performance at the 10th percentile ($R^2$ = 0.552) compared to stacking ($R^2$ = 0.540). BMA also leads in capturing peak performance, with a 90th percentile $R^2$ of 0.821 versus 0.811 for stacking.

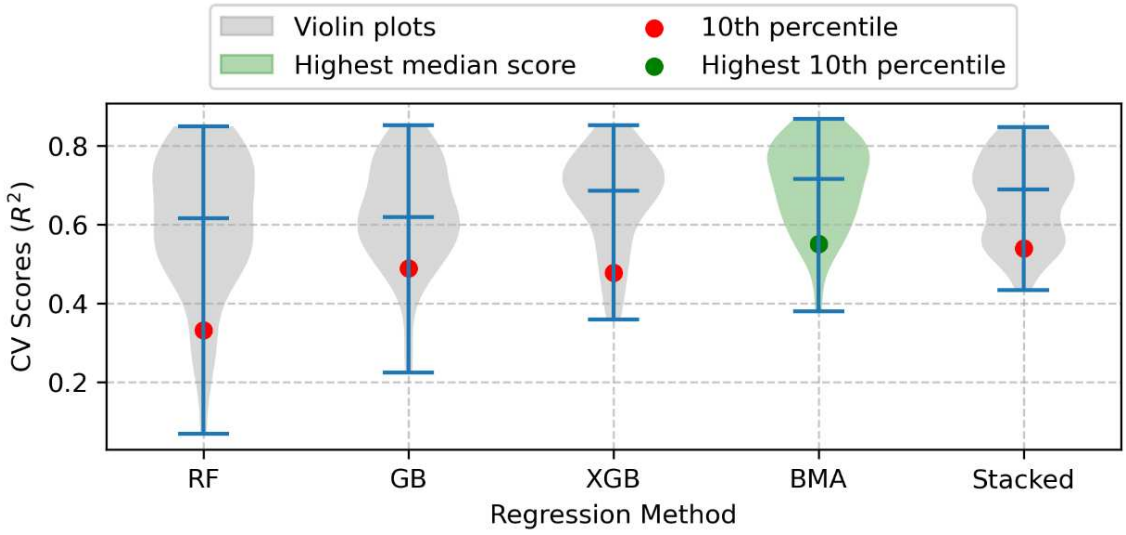


Figure 2- 9 Violin Plots Demonstrating the Performance Distribution of Top Three Performing Models and Two Ensemble Models: Extreme Gradient Boosting (XGB), Gradient Boosting (GB), Random Forest (RF), Bayesian Model Averaging (BMA), and Stacked Ensembles Performance While Highlighting Max, Min, Median, And 10$^{th}$ Percentile Values for Each Model.

Information criteria analysis provided additional insights into model performance while accounting for complexity (Table 2- 3). For individual models, k includes the 13 input variables plus the number of hyperparameters optimized during model development (k=18-19). For ensemble methods, k incorporates both model complexity and ensemble structure: BMA includes the input variables, combined number of models hyperparameters, plus ensemble weights (k=32), while the stacked model includes an additional model layer (k=33).

Table 2- 3 Performance Metrics Including Information Criteria for All Models

| Model | Median $R^2$ | Parameters (k) | RSS | AIC | BIC |
|---|---|---|---|---|---|
| XGB | 0.68 | 19 | 21.6 | 173.0 | 230.2 |
| GB | 0.62 | 18 | 25.6 | 196.7 | 250.9 |
| RF | 0.64 | 18 | 24.3 | 188.6 | 242.8 |
| BMA | 0.72 | 32 | 17.9 | 170.8 | 267.1 |
| Stacked | 0.69 | 33 | 20.9 | 196.2 | 295.5 |

BMA achieved the strongest predictive performance ($R^2 = 0.72$) and the best AIC score (170.8). While its BIC value (267.1) is higher than individual models due to increased parameters, the improvement in predictive performance and lower RSS (17.9) suggests an effective balance between accuracy and complexity. The stacked ensemble also shows good predictive performance ($R^2 = 0.69$) but with higher information criteria scores (AIC = 196.2, BIC = 295.5). These results suggest that while both ensemble methods offer improved predictions over individual models, BMA provides these improvements with relatively more efficient use of its additional complexity.

Computational efficiency analysis further supports the advantages of BMA. While model optimization took slightly longer for stacking (1885s vs 1662s), the two-level structure led to higher average training times (11.59s vs 1.43s). This gap widened during prediction, with stacking

requiring much longer than BMA (0.05s vs 0.0008s). These efficiency gains stem from BMA's simpler weighted averaging approach compared to stacking's additional model layer. For practical applications, especially in resource-constrained environments or when real-time predictions are needed, BMA's combination of superior performance and computational efficiency makes it the more attractive option.

#### 2.3.1.3 MODEL UNCERTAINTY AND PERFORMANCE DISTRIBUTION

The uncertainty of the ML models was systematically analyzed across multiple dimensions including performance variability and prediction stability. Table 2- 4 summarizes the key uncertainty metrics for each model.

Cross-validation analysis revealed varying levels of prediction uncertainty across models. The BMA demonstrated the highest mean $R^2$ (0.700 ± 0.110) with the tightest 95% confidence interval [0.479, 0.861], indicating robust performance across different data splits. XGB showed comparable but slightly lower performance ($R^2$ = 0.667 ± 0.122, CI: [0.389, 0.832]), while RF exhibited the highest uncertainty ($R^2$ = 0.602 ± 0.176, CI: [0.243, 0.841]). The stacked ensemble achieved similar mean performance to XGB ($R^2$ = 0.668 ± 0.108) with a narrower confidence interval [0.471, 0.825], suggesting more consistent predictions.

The coefficient of variation (CV) served as a key metric for assessing model stability, with lower values indicating more stable performance. BMA exhibited the highest stability (CV = 15.7%), followed closely by the stacked ensemble (CV = 16.2%). Individual models showed higher variability, with XGB (CV = 18.4%), GB (CV = 21.0%), and RF (CV = 29.2%) demonstrating progressively decreasing stability. These patterns align with theoretical expectations that ensemble methods should reduce prediction variance by averaging out individual model uncertainties. The

particularly low CV of BMA suggests that its probabilistic weighting mechanism effectively balances the strengths and weaknesses of individual models across different prediction scenarios.

The stability analysis through range metrics provides additional insights, with the stacked ensemble showing the smallest total range (0.413), followed by BMA (0.488) and XGB (0.493), while RF exhibited the largest range (0.781). This compressed range in ensemble methods indicates their ability to avoid extreme prediction errors that might occur with individual models. The slightly better range performance of stacking compared to BMA suggests that while BMA offers better average performance and stability, stacking might have advantages in preventing extreme prediction errors.

Statistical significance testing through paired t-tests against XGB as the baseline revealed meaningful performance differences. BMA showed a statistically significant improvement over XGB ($p = 0.002$) with a relative performance gain of +5.0%. Both GB and RF performed significantly worse than XGB ($p = 0.029$ and $p < 0.001$ respectively) with relative decrements of -6.2% and -9.7%. The stacked ensemble's marginal improvement of +0.3% over XGB was not statistically significant, suggesting that its primary benefit lies in stability rather than accuracy improvement.

Table 2- 4 Model Uncertainty Metrics

| Model | Mean $R^2 \pm$ SD | 95% CI | CV (%) | Range | $R^2$ Improvement vs XGB |
|---|---|---|---|---|---|
| XGB | 0.667±0.122 | [0.389,0.832] | 18.4 | 0.493 | - |
| GB | 0.625±0.132 | [0.327,0.837] | 21.0 | 0.628 | -6.2%* |
| RF | 0.602±0.176 | [0.243,0.841] | 29.2 | 0.781 | -9.7%* |
| BMA | 0.700±0.110 | [0.479,0.861] | 15.7 | 0.488 | +5.0%* |
| Stacked | 0.668±0.108 | [0.471,0.825] | 16.2 | 0.413 | +0.3% |

*Statistically significant at $p < 0.05$

The lower variability of ensemble methods, particularly BMA, suggests they provide more reliable predictions when deployed in operational settings. Comprehensive uncertainty metrics support BMA as the most robust choice for water quality modeling applications where prediction stability is as important as absolute accuracy. This aligns with previous studies in environmental modeling where prediction consistency across varying conditions is often prioritized over marginal improvements in peak performance.

#### 2.3.1.4 COMPARATIVE ANALYSIS OF ENSEMBLE METHODS

When comparing BMA and stacked ensembles, several key differences emerge in both methodology and performance. BMA demonstrates superior predictive performance with a median $R^2$ of 0.716 compared to stacking's 0.689, while offering better computational efficiency with prediction times reduced by a factor of fifty (0.0008s vs 0.05s per prediction).

The fundamental methodological differences between these approaches lie in their core architecture. BMA employs a probabilistic framework that weights individual models based on their posterior probabilities, whereas stacking uses a meta-learner approach to learn optimal combinations through an additional modeling layer. The probabilistic nature of BMA offers natural uncertainty estimates, while stacking requires additional techniques for uncertainty quantification.

From a computational perspective, BMA demonstrates notably lower overhead, primarily involving weight calculation for model combinations. In terms of robustness, BMA shows superior stability with a coefficient of variation of 15.7% compared to stacking's 16.2%, indicating more consistent performance across different prediction scenarios.

Overall, both methods improve upon individual models, BMA offers a more favorable balance of performance, efficiency, and implementation simplicity for salinity prediction applications. Its

computational efficiency and natural uncertainty quantification makes it suitable for operational deployment in water quality monitoring systems, especially in scenarios where rapid predictions and reliable uncertainty estimates are crucial for decision-making processes, such as evaluating the salinity impacts of inter-basin transfers, assessing the risks of changing irrigation patterns, and optimizing flow management strategies to meet regulatory standards.

#### 2.3.1.5 SPATIAL CROSS-VALIDATION RESULTS

The spatial cross-validation results provide important insights into the models' practical applicability across the CRB. As shown in Figure 2- 10a, the study area was divided into six spatial folds based on geographic clustering, ensuring a robust assessment of model generalization across distinct regions, ensuring a robust assessment of the methodology's performance across distinct hydrogeological regions. As expected with machine learning approaches, all models showed decreased performance metrics when moving from random to spatial cross-validation, with the median $R^2$ dropping from 0.72 to 0.59 for BMA. This performance decrease is a natural consequence of predictions in areas with potentially different underlying hydrogeological processes than those represented in the training data. This spatial performance gap is consistent with other environmental modeling studies and highlights the importance of comprehensive spatial sampling strategies (De la Fuente et al., 2023).

The violin plots in Figure 2- 10b illustrate the superior performance of BMA in spatial cross-validation (median $R^2 = 0.59$) compared to individual models (XGB: 0.47, GB: 0.41, RF: 0.37), demonstrating the advantage of ensemble methods in handling spatial heterogeneity. BMA's notably higher 10th percentile $R^2$ value (0.27 versus 0.16, 0.13, and 0.10 for individual models) indicates more robust performance in challenging prediction scenarios. This pattern suggests that

ensemble methods, particularly BMA, are better equipped to handle the varying relationships between predictors and salinity across different geographic regions.

The spatial performance patterns reveal interesting regional variations in model accuracy. Performance tends to be stronger in regions with denser monitoring networks and more consistent hydrogeological characteristics. The lower performance in certain regions likely reflects local complexities in salinity dynamics that differ from the dominant patterns in the training data.

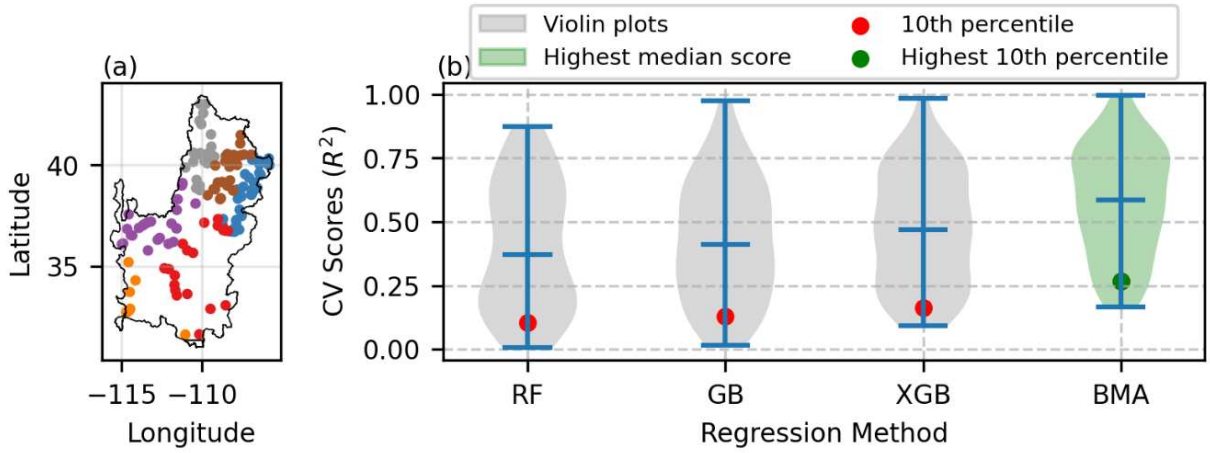


Figure 2- 10 (a) Geographic Distribution of Spatial Cross-Validation Folds Sample in Colorado River Basin; each fold represented by a different color, and (b) Comparative Performance of Machine Learning Models Under Spatial Folds Using Violin Plots, Showing $R^2$ Score Distributions. Green Violin Indicates the Highest Median Performance; Red and Green Dots Show 10$^{th}$ Percentile $R^2$ Values and the Highest 10$^{th}$ Percentile, respectively.

### 2.3.2 SULFATE LOADS AND CONCENTRATIONS

Sulfate loads and concentrations at the HUC8 sub-basin level were estimated by integrating modeled streamflow from the SWAT+ national model (White et al., 2022) with sulfate predictions from the ML model developed here. SWAT+ national model provided flow estimates throughout the basin, while the ML model leveraged water quality and watershed characteristics to generate sulfate concentrations. Loads for each HUC8 watershed were subsequently derived using the flow

volumes and sulfate concentrations. This approach coupled the predictive capabilities of a physically based hydrologic model with data-driven water quality modeling to obtain a spatially distributed characterization of sulfate across the CRB.

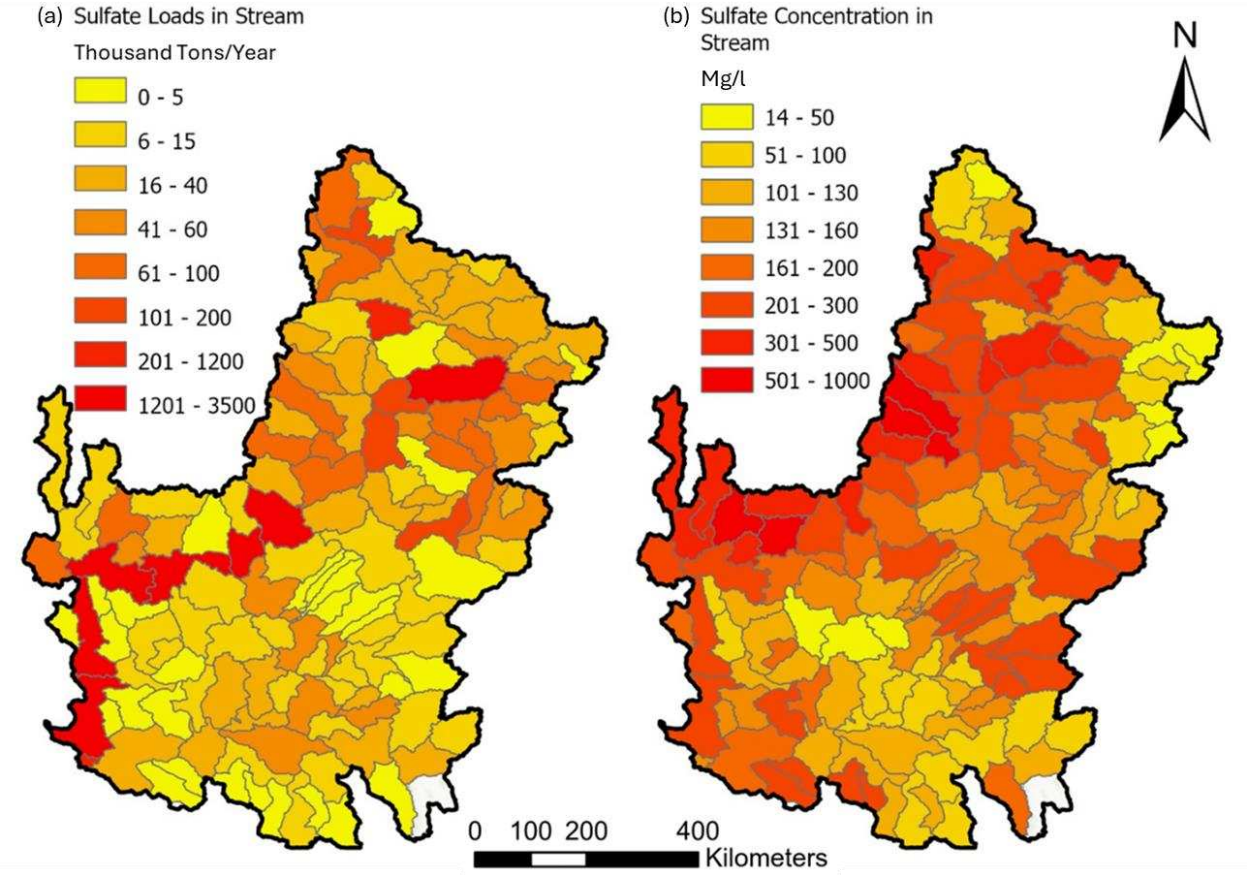


Figure 2- 11 (a) Spatial Representation of Sulfate Loads in The Stream Network at the HUC8 Sub-Basins Level. (b) Corresponding Depiction of Sulfate Concentrations in The Stream Network.

Sulfate loads and concentrations exhibited distinct spatial patterns when analyzed at the HUC8 sub-basin level across the CRB (Figure 2- 11). Sulfate loads were highest in the western watersheds, with progressive accumulation occurring downstream along the river's mainstem. This pattern aligns with the expected cumulative nature of dissolved solids transport in river systems and reflects the influence of both natural geological sources and anthropogenic inputs. However, concentrations remained relatively low in the central mainstream due to dilution from higher flows.

Certain tributaries showed elevated concentrations despite lower loads, due to reduced flows failing to dilute salt adequately. The pattern is especially pronounced in regions with significant irrigation return flows or areas of high geological salt content, where limited water volume fails to adequately dilute the salt inputs. These findings align with previous studies (Miller et al., 2017; Tillman et al., 2019) that identified similar concentration patterns in agricultural areas of the basin.

The spatial distribution of sulfate also reveals potential management implications. Areas with high concentrations but lower loads might benefit from flow management strategies, while regions with high loads but moderate concentrations might require source control measures. The identification of these distinct patterns demonstrates the value of considering both loads and concentrations in salinity management planning.

### 2.3.3 STUDY LIMITATIONS AND FUTURE RESEARCH DIRECTIONS

While this study demonstrates the effectiveness of ensemble machine learning methodologies for salinity prediction, several important limitations warrant discussion. The twenty-year study period using annual averages may not fully capture temporal variations in salinity patterns. As with all machine learning approaches, model performance is fundamentally bound by the characteristics of its training data, evidenced by the expected decrease in performance during spatial cross-validation (from $R^2 = 0.72$ to 0.59) when predicting in geographically distinct regions. This performance pattern reinforces that the study's contribution lies in demonstrating the effectiveness of the ensemble methodology rather than creating a universal salinity prediction model.

The demonstrated ensemble approach can be applied to other river basins but would require retraining with local hydrogeological data and careful consideration of available monitoring networks. Future research should focus on testing these ensemble methods with high frequency

monitoring data and expanding the methodology to handle various temporal resolutions. Future research could extend this ensemble methodology to evaluate potential salinity impacts of water management scenarios, including inter-basin transfers and changing irrigation patterns. Building on these opportunities, the superior performance of ensemble methods compared to individual models validates their utility for water quality modeling applications. The study provides a robust methodological framework that can be adapted and retrained for salinity prediction in different geographical contexts while maintaining the practical utility demonstrated in the Colorado River Basin.

## 2.4 CONCLUSION

This study demonstrates the effectiveness of ensemble machine learning approaches for salinity prediction in large river basins through systematic evaluation of multiple modeling strategies. Cross-validation analysis identified Extreme Gradient Boosting, Gradient Boosting, and Random Forest as the top performers among eleven machine learning algorithms tested. These models were integrated using Bayesian Model Averaging (BMA) and stacked generalization to develop robust ensemble frameworks.

The BMA ensemble demonstrated superior performance and stability compared to both individual models and stacked ensembles. While achieving comparable training performance ($R^2 = 0.99$), BMA showed improved testing metrics ($R^2 = 0.72$) and notably better performance in spatial cross-validation ($R^2 = 0.59$). The enhanced stability of BMA was evidenced by the lowest coefficient of variation (15.7%) and tight confidence intervals [0.479, 0.861], suggesting more reliable predictions across varying conditions. Furthermore, BMA achieved these improvements while

requiring significantly less computational resources than stacking, with prediction times reduced by a factor of fifty.

Data preprocessing through log transformation and bias correction proved crucial for model performance, improving RMSE by 18.61% and $R^2$ by 10.17%. The study revealed soil electrical conductivity and calcium carbonate content as the most influential predictors, followed by river flow. Spatial analysis highlighted distinct patterns in sulfate loads and concentrations across sub-basins, with high loads in western watersheds but varying concentrations influenced by flow dynamics.

The spatial cross-validation results underscore both the methodology's strengths and limitations, demonstrating that while ensemble methods improve prediction stability across different regions, performance naturally decreases when predicting in areas with distinct hydrogeological characteristics from the training data. This reinforces that the study's primary contribution lies in establishing an effective methodological framework rather than creating a universal prediction model.

This chapter contributes an ensemble machine learning framework using Bayesian Model Averaging that provides robust, uncertainty-quantified predictions of salinity and nutrient loads from natural landscapes. Three specific contributions emerge. First, the framework demonstrates that ensemble approaches reduce prediction variance by 46% compared to individual algorithms while maintaining cross-validation performance ($R^2$ = 0.70–0.83), establishing principled approaches for combining algorithmically diverse models in environmental applications. Second, spatial cross-validation protocols quantify transferability limits, revealing 10–20% performance reduction when predicting in ungauged watersheds—providing realistic bounds on model applicability and informing monitoring network priorities. Third, consistent ranking of soil

electrical conductivity and calcium carbonate as dominant predictors across Random Forest, Gradient Boosting, and XGBoost provides stronger evidence for geological controls on nutrient export than any single model could provide, demonstrating that cross-algorithm agreement strengthens process inference.

Future applications of this approach in different geographical contexts would require retraining with local data, while methodological improvements could focus on incorporating high frequency monitoring data and enhancing spatial generalization capabilities. The demonstrated success of ensemble methods, particularly BMA, in handling the complex relationships between environmental predictors and salinity levels provides a robust foundation for water quality modeling applications where prediction stability and computational efficiency are crucial considerations.

## 2.5 EXTENSION: ENSEMBLE MACHINE LEARNING FOR FOREST AND RANGELAND NUTRIENT PREDICTION

### 2.5.1 INTRODUCTION AND MOTIVATION

The ensemble machine learning framework presented in Sections 2.1-2.4 demonstrated robust performance for salinity prediction in the Colorado River Basin, achieving median $R^2$ values of 0.72 through Bayesian Model Averaging while maintaining computational efficiency and uncertainty quantification. This section extends the methodology to address a distinct challenge in watershed nutrient modeling: prediction of Total Nitrogen (TN) and Total Phosphorus (TP) loads from forest and rangeland areas under alternative future scenarios.

Understanding nutrient contributions from natural landscapes presents unique challenges that differ from salinity prediction. Forest and rangeland areas, while often assumed to contribute

minimal nutrient loads compared to agricultural or urban sources, can represent significant portions of watershed area and provide baseline nutrient export that influences water quality management targets (Ouyang et al., 2007). Traditional modeling approaches often struggle to capture the complex interactions between climate, soil properties, vegetation characteristics, and topography that control nutrient export from these minimally disturbed landscapes. Furthermore, integration with scenario-based planning frameworks requires predictions that respond appropriately to altered precipitation patterns and land management conditions while maintaining physical consistency.

The application to forest and rangeland nutrient prediction provides a test of the ensemble framework's transferability across water quality constituents. Unlike salinity, which represents a single parameter with relatively direct relationships to source characteristics, nutrient prediction requires simultaneous modeling of nitrogen and phosphorus species that exhibit distinct biogeochemical behaviors and transport pathways. Additionally, the sparse nature of nutrient concentration data in headwater watersheds necessitates data augmentation strategies not required for salinity applications.

The extension to forest and rangeland nutrient prediction contributes the application of ensemble uncertainty quantification to a distinct environmental prediction challenge. While the salinity application (Sections 2.1-2.4) addressed a single constituent with relatively direct source relationships, nutrient prediction requires simultaneous modeling of nitrogen and phosphorus species exhibiting distinct biogeochemical behaviors. This extension tests the generalizability of ensemble approaches across water quality constituents and establishes baseline natural landscape predictions essential for the integrated framework developed in subsequent chapters.

### 2.5.2 STUDY DESIGN AND DATA

#### 2.5.2.1 WATERSHED SELECTION AND CHARACTERISTICS

Analysis focused on 87 watersheds across Colorado selected to isolate nutrient dynamics from forest and rangeland areas while minimizing confounding influences from other land uses (Figure 2- 12). Selection criteria required watersheds containing more than 10,000 acres with forest and rangeland representing over 90% of total area. This careful selection enabled focused analysis of natural landscape nutrient export processes. The watersheds span elevations from alpine headwaters exceeding 3,500 meters to montane valleys near 2,000 meters, encompassing diverse climate regimes, soil types, and vegetation community's characteristic of Colorado's mountainous regions.

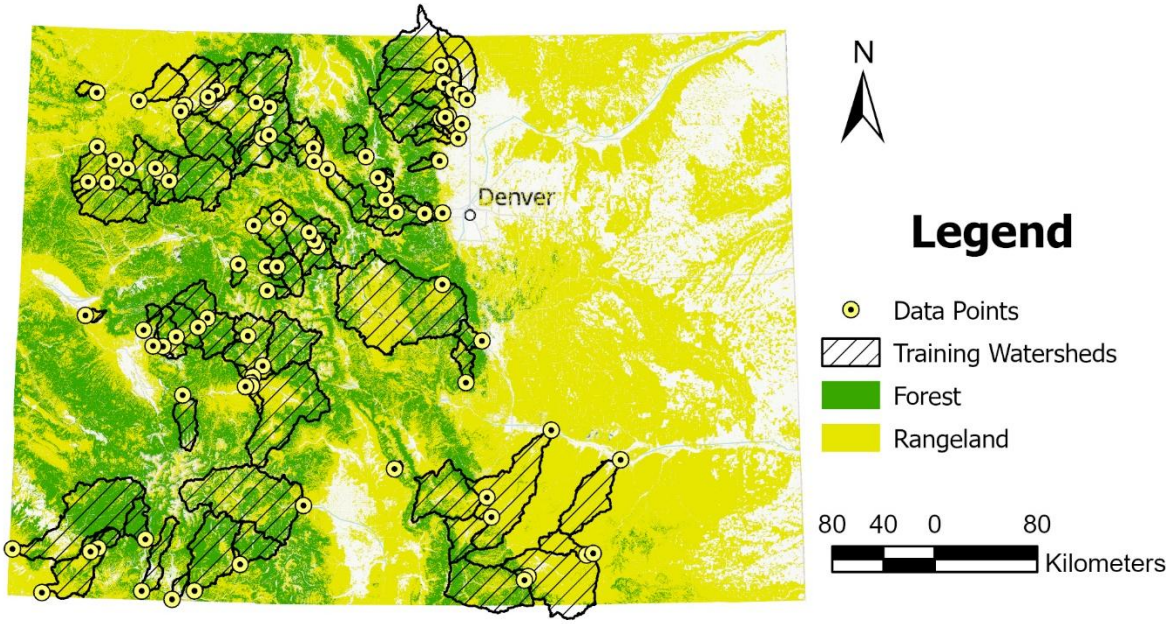


Figure 2- 12 Selected watersheds with water quality monitoring points and flow measurement stations across Colorado.

#### 2.5.2.2 INPUT VARIABLES AND DATA SOURCES

The modeling framework incorporates comprehensive environmental data organized into seven major categories (Table 2- 5). Geographic features include watershed area, latitude, longitude, and slope characteristics derived from digital elevation models. Soil physical properties encompass available water capacity at multiple depths, bulk density, and textural components (sand, silt, clay percentages) obtained from SSURGO databases. Soil chemical properties include calcium carbonate content, electrical conductivity, organic matter content, and pH, all measured at 5 cm and 15 cm depths. Hydrological properties focus on saturated hydraulic conductivity at various soil depths to characterize infiltration capacity.

Vegetation and spectral indices derived from remote sensing provide information about landscape conditions and seasonal dynamics. These indices include red-blue, red-green, red-SWIR1, SWIR1-NIR, and SWIR1-SWIR2 vegetation indices that capture differences in canopy density, phenology, and stress conditions. Climatic variables incorporate mean, maximum, and minimum temperatures, vapor pressure deficit (maximum and minimum), and precipitation. Additional parameters include curve number for runoff estimation and digital elevation metrics.

Variance inflation factor analysis was applied to reduce redundancy among the 40 initially selected variables. This procedure identified and eliminated highly collinear predictors that would contribute minimal additional information while increasing model complexity and risk of overfitting. The analysis retained 14 variables for model training (Table 2- 5), representing the minimum set capable of capturing essential environmental gradients controlling nutrient export from natural landscapes while maintaining statistical independence.

Table 2- 5 Input variables evaluated for forest and rangeland nutrient prediction. Variables retained after variance inflation factor (VIF) analysis are indicated in underlined bold.

| Category | Variables |
|---|---|
| Geographical Features | **Area (Forest & Rangeland)**, Longitude, Latitude, Slope |
| Soil Physical Properties | awc_15, awc_5, **bedrock**, bulkd_15, bulkd_5, clay_15, clay_5, sand_15, **sand_5**, silt_15, silt_5 |
| Soil Chemical Properties | **caco3_15**, **caco3_5**, ec_15, **ec_5**, **om_15**, **om_5**, ph_15, ph_5 |
| Hydrological Properties | **ksat_15**, ksat_5 |
| Vegetation/Spectral Indices | redblue, **redgreen**, **redswir1**, swir1nir, swir1swir2 |
| Climatic Variables | tdmean, tmax, **tmin**, vpdmax, vpdmin, **ppt** |
| Miscellaneous | cn, dem, solclear, soltotal |

#### 2.5.2.3 WATER QUALITY DATA AND AUGMENTATION

The study utilized an extensive water quality dataset comprising approximately 66,000 measurements for TP and 70,000 measurements for nitrogen species collected from monitoring stations within and adjacent to the selected watersheds. A challenge in nutrient modeling stems from incomplete sampling across nitrogen species, with some monitoring events recording Total Kjeldahl Nitrogen (TKN) but not nitrate-nitrite, and vice versa. This gap prevents direct calculation of Total Nitrogen (TN = TKN + nitrate-nitrite) for many observations.

A machine learning approach was developed to address this data limitation. Random Forest models were trained to predict TKN from nitrate-nitrite measurements and conversely predict nitrate-nitrite from TKN measurements, incorporating measurement date and geographic location as additional predictors. This data augmentation strategy achieved $R^2$ values of approximately 0.8 for both prediction directions (Figure 2- 13), substantially expanding the available training dataset.

Cross-validation procedures ensured that augmented data did not introduce spurious patterns that could degrade model performance on independent test data.

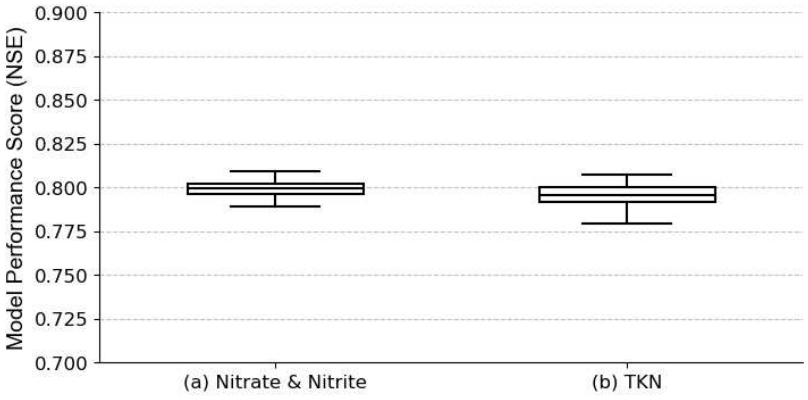


Figure 2- 13 Cross-validation model performance for predicting nitrogen species.(a) prediction of TKN from nitrate-nitrite concentration, date, and location. (b) prediction of nitrate-nitrite from Total Kjeldahl Nitrogen (TKN) concentration, measurement date, and location.

### 2.5.3 METHODOLOGY

#### 2.5.3.1 ENSEMBLE MODELING FRAMEWORK

The modeling framework employs the same three primary machine learning algorithms validated in the salinity application: Random Forest (RF), Gradient Boosting Regression (GBR), and Extreme Gradient Boosting (XGB). As demonstrated in Section 2.3, these algorithms effectively capture complex environmental relationships through ensemble tree-based approaches. The Bayesian Model Averaging (BMA) ensemble approach combines predictions from individual models while accounting for prediction uncertainty (Hoeting et al., 1999; Raftery et al., 1997).

#### 2.5.3.2 MODEL TRAINING AND VALIDATION

The training process employed a rigorous cross-validation strategy to ensure robust performance assessment and prevent overfitting to specific watershed characteristics. Five-fold cross-validation testing provided assessment of generalization capability by systematically withholding subsets of

watersheds during training and evaluating predictions for these held-out locations. This spatial cross-validation approach more rigorously tests model transferability than random data splitting because it requires prediction for watersheds with potentially distinct combinations of environmental conditions. Cross-validation results demonstrated robust generalization, with $R^2$ values of 0.70 for TP and 0.83 for TN. The reduction from training to cross-validation performance, while expected, remained within acceptable ranges for environmental modeling applications and compared favorably to the salinity application (training $R^2$=0.99, cross-validation $R^2$=0.72).

### 2.5.4 RESULTS

#### 2.5.4.1 PREDICTIVE PERFORMANCE

The performance metrics achieved in cross-validation (NSE = 0.80 for TN, 0.70 for TP) indicate that the ensemble framework successfully transfers to nutrient prediction from natural landscapes despite fundamental differences from the salinity application. Nitrogen prediction achieved higher accuracy than phosphorus, consistent with nitrogen's more conservative transport behavior through Colorado's forested watersheds. The framework's ability to maintain performance across constituents with distinct biogeochemical behaviors supports its utility as a generalizable approach for water quality prediction at regional scales.

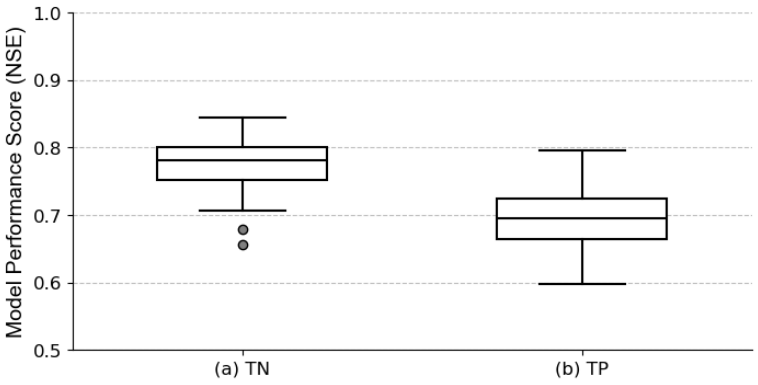


Figure 2- 14 BMA model cross-validation results for predicting TN and TP loads from forest and rangeland areas.

#### 2.5.4.2 SPATIAL DISTRIBUTION OF FOREST AND RANGELAND NUTRIENT LOADS

To characterize the baseline nutrient export regime, we modeled Total Nitrogen (TN) and Total Phosphorus (TP) loads at the HUC12 scale across Colorado's forest and rangeland sectors. Rather than focusing on aggregated statewide magnitudes, which can obscure critical local variances, Figure 2- 15 illustrates the spatial heterogeneity of nutrient source areas.

The model reveals distinct spatial patterns governed by the interplay of hydrology and land cover. High-loading "hotspots" are distinctively clustered in high-elevation headwaters, correlating with areas of higher precipitation and steeper hydraulic gradients. Conversely, vast rangeland areas in the eastern plains exhibit lower specific yields, likely driven by transport limitations in arid conditions. This spatial discretization identifies critical source areas where forest and rangeland background loads may exert disproportionate influence on downstream water quality, providing a more granular tool for watershed management than basin-scale averages.

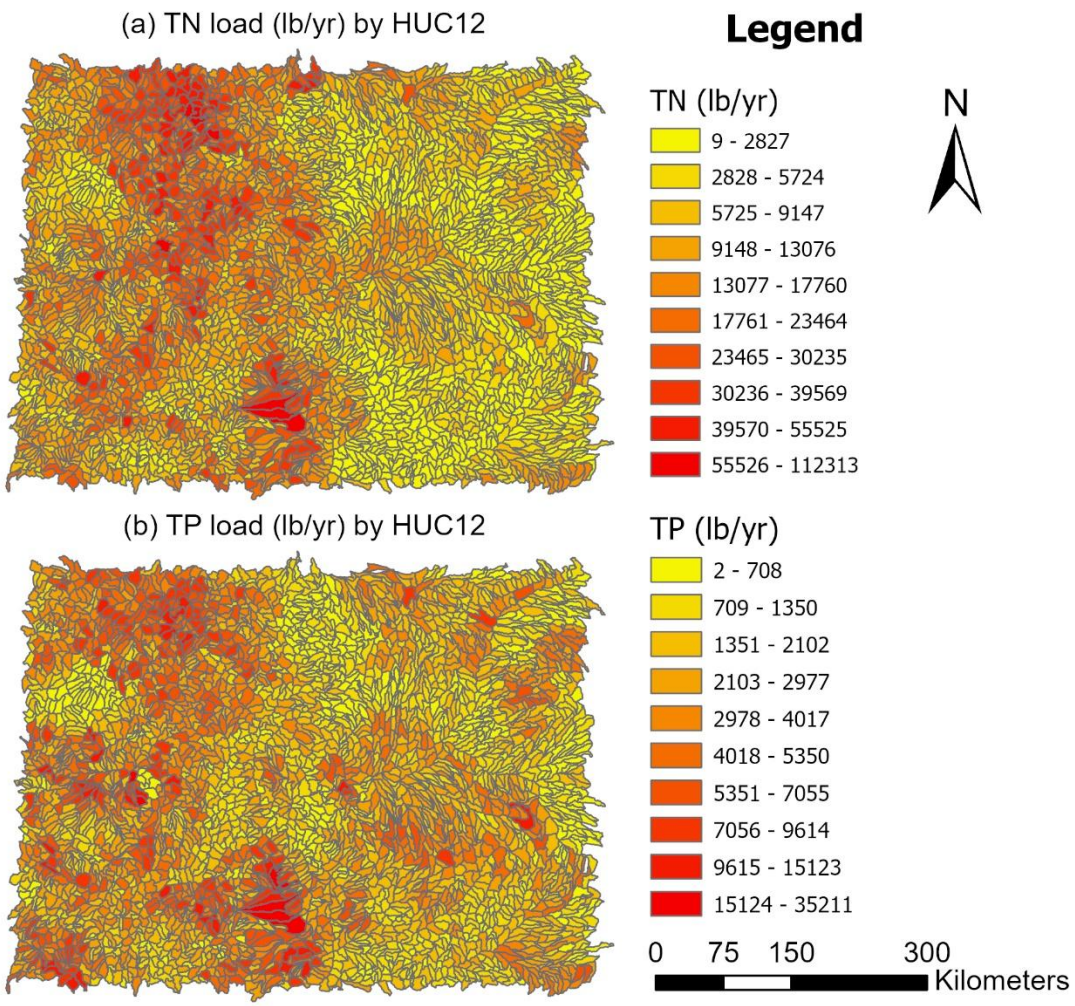


Figure 2- 15 The spatial heterogeneity of nutrient sources from Forest & Rangeland areas

## 2.5.5 DISCUSSION

### 2.5.5.1 METHODOLOGICAL INSIGHTS FROM MULTI-CONSTITUENT APPLICATION

Application of the ensemble machine learning framework to both salinity (Section 2.3) and nutrient prediction reveals consistent strengths and limitations across water quality constituents. Both applications achieved strong cross-validation performance ($R^2$ = 0.72 for salinity, 0.70-0.83 for nutrients), indicating that the framework successfully captures dominant environmental gradients controlling constituent export.

However, the applications also highlight constituent-specific challenges. Salinity prediction benefited from continuous monitoring data at high temporal frequency, enabling robust characterization of seasonal dynamics and event-based variability. In contrast, nutrient data from forest and rangeland watersheds exhibited sparser temporal coverage, necessitating data augmentation strategies and introducing additional uncertainty. The lower $R^2$ achieved for phosphorus compared to nitrogen in the nutrient application (0.70 vs 0.83) mirrors patterns observed in process-based modeling studies and likely reflects fundamental differences in constituent mobility rather than methodological limitations.

The data augmentation approach developed for nitrogen species merits further consideration. While achieving $R^2 \approx 0.8$ for TKN-nitrate relationships substantially expanded the training dataset, this procedure introduces uncertainty through two-stage modeling. Predictions generated from augmented data should be interpreted with recognition that they represent model-derived estimates rather than direct observations.

#### 2.5.5.2 LIMITATIONS AND FUTURE DIRECTIONS

Several limitations merit acknowledgment. The spatial cross-validation approach employed provides robust assessment of transferability to unmonitored watersheds within Colorado but does not test performance in regions with fundamentally different geology, climate, or vegetation types. Application beyond the Rocky Mountain region would require retraining with local data, though the methodological framework itself remains transferable.

The scenario analysis currently addresses only climate forcing variables (precipitation, temperature) and does not explicitly incorporate potential changes in disturbance regimes (fire frequency) or forest management practices that could substantially influence nutrient export under future conditions. Expanding the framework to address these factors would require additional

predictor variables and training data representing the full range of management and disturbance conditions.

## 2.6 TRANSITION TO URBAN NUTRIENT CONTRIBUTIONS

This chapter established ensemble machine learning as a robust approach for predicting nutrient loads from natural landscapes (forests and rangelands), achieving cross-validation $R^2$ values of 0.70-0.83 while maintaining computational efficiency essential for state-wide application. However, natural landscapes represent only one component of watershed nutrient budgets. Urban development -and its associated stormwater runoff- constitutes an increasingly significant source as Colorado's population grows along the Front Range corridor. The next chapter addresses this urban dimension by developing a deep learning framework to predict spatial patterns of urban expansion through 2050 under alternative growth scenarios. These land cover predictions provide the foundation for estimating scenario-specific urban stormwater nutrient loads that, combined with the forest/rangeland predictions developed here, enable comprehensive accounting of landscape-scale nutrient contributions across diverse land use types.

## 2.7 REFERENCES

A. Ramezan, C., A. Warner, T., & E. Maxwell, A. (2019). Evaluation of Sampling and Cross-Validation Tuning Strategies for Regional-Scale Machine Learning Classification. Remote Sensing, 11(2), Article 2. https://doi.org/10.3390/rs11020185

Akinwande, M. O., Dikko, H. G., & Samson, A. (2015). Variance Inflation Factor: As a Condition for the Inclusion of Suppressor Variable(s) in Regression Analysis. Open Journal of Statistics, 05(07), Article 07. https://doi.org/10.4236/ojs.2015.57075

Berrar, D. (2019). Cross-Validation. In Encyclopedia of Bioinformatics and Computational Biology (pp. 542–545). Elsevier. https://doi.org/10.1016/B978-0-12-809633-8.20349-X

Bhanja, S., & Das, A. (2018). Impact of Data Normalization on Deep Neural Network for Time Series Forecasting.

Burnaev, E., & Vovk, V. (2014). Efficiency of conformalized ridge regression. Proceedings of The 27th Conference on Learning Theory, 605–622. https://proceedings.mlr.press/v35/burnaev14.html

Cai, J., Xu, K., Zhu, Y., Hu, F., & Li, L. (2020). Prediction and analysis of net ecosystem carbon exchange based on gradient boosting regression and random forest. Applied Energy, 262, 114566. https://doi.org/10.1016/j.apenergy.2020.114566

Canziani, A., Paszke, A., & Culurciello, E. (2017). An Analysis of Deep Neural Network Models for Practical Applications (arXiv:1605.07678). arXiv. https://doi.org/10.48550/arXiv.1605.07678

Cao, H., Gu, Y., Fang, J., Hu, Y., Ding, W., He, H., & Chen, G. (2022). Application of stacking ensemble learning model in quantitative analysis of biomaterial activity. Microchemical Journal, 183, 108075. https://doi.org/10.1016/j.microc.2022.108075

Chakrabarti, A., & Ghosh, J. K. (2011). AIC, BIC and Recent Advances in Model Selection. In P. S. Bandyopadhyay & M. R. Forster (Eds.), Philosophy of Statistics (Vol. 7, pp. 583–605). North-Holland. https://doi.org/10.1016/B978-0-444-51862-0.50018-6

Cheng, J., Sun, J., Yao, K., Xu, M., & Cao, Y. (2022). A variable selection method based on mutual information and variance inflation factor. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 268, 120652. https://doi.org/10.1016/j.saa.2021.120652

Christensen, N. S., Wood, A. W., Voisin, N., Lettenmaier, D. P., & Palmer, R. N. (2004). The Effects of Climate Change on the Hydrology and Water Resources of the Colorado River Basin. Climatic Change, 62(1), 337–363. https://doi.org/10.1023/B:CLIM.0000013684.13621.1f

Colorado River Basin Salinity Control Forum. (2023). Water Quality Standards for Salinity—Colorado River System. https://barnettwater.sharepoint.com/:b:/g/ERm4dI86t1JEndfeaozZT4cBmfb8JoqmHh48Y4Vm3JA03w?e=vuThnW

Cootes, T. F., Ionita, M. C., Lindner, C., & Sauer, P. (2012). Robust and Accurate Shape Model Fitting Using Random Forest Regression Voting. In A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, & C. Schmid (Eds.), Computer Vision – ECCV 2012 (pp. 278–291). Springer. https://doi.org/10.1007/978-3-642-33786-4_21

Craney, T. A., & Surles, J. G. (2002). Model-Dependent Variance Inflation Factor Cutoff Values. Quality Engineering, 14(3), 391–403. https://doi.org/10.1081/QEN-120001878

De la Fuente, L. A., Gupta, H. V., & Condon, L. E. (2023). Toward a Multi-Representational Approach to Prediction and Understanding, in Support of Discovery in Hydrology. Water Resources Research, 59(1), e2021WR031548. https://doi.org/10.1029/2021WR031548

Domingos, P. (2000). Bayesian Averaging of Classifiers and the Overfitting Problem. Proceedings of the Seventeenth International Conference on Machine Learning, 223–230.

Duan, Q., Ajami, N. K., Gao, X., & Sorooshian, S. (2007). Multi-model ensemble hydrologic prediction using Bayesian model averaging. Advances in Water Resources, 30(5), 1371–1386. https://doi.org/10.1016/j.advwatres.2006.11.014

Eberly, L. E. (2007). Multiple Linear Regression. In W. T. Ambrosius (Ed.), Topics in Biostatistics (pp. 165–187). Humana Press. https://doi.org/10.1007/978-1-59745-530-5_9

ECHO. (2022). Enforcement and Compliance History Online | US EPA. https://echo.epa.gov/

Es-Sabery, F., Hair, A., Qadir, J., Sainz-De-Abajo, B., Garcia-Zapirain, B., & Torre-Diez, I. (2021). Sentence-Level Classification Using Parallel Fuzzy Deep Learning Classifier. IEEE Access, 9, 17943–17985. https://doi.org/10.1109/ACCESS.2021.3053917

Guth, P. L., Van Niekerk, A., Grohmann, C. H., Muller, J.-P., Hawker, L., Florinsky, I. V., Gesch, D., Reuter, H. I., Herrera-Cruz, V., Riazanoff, S., López-Vázquez, C., Carabajal, C. C., Albinet, C., & Strobl, P. (2021). Digital Elevation Models: Terminology and Definitions. Remote Sensing, 13(18), Article 18. https://doi.org/10.3390/rs13183581

Ha, T. N., Lubo-Robles, D., Marfurt, K. J., & Wallet, B. C. (2021). An in-depth analysis of logarithmic data transformation and per-class normalization in machine learning: Application to unsupervised classification of a turbidite system in the Canterbury Basin,

New Zealand, and supervised classification of salt in the Eugene Island minibasin, Gulf of Mexico. Interpretation, 9(3), T685–T710. https://doi.org/10.1190/INT-2021-0008.1

Hauenstein, S., Wood, S. N., & Dormann, C. F. (2018). Computing AIC for black-box models using generalized degrees of freedom: A comparison with cross-validation. Communications in Statistics - Simulation and Computation, 47(5), 1382–1396. https://doi.org/10.1080/03610918.2017.1315728

Hoerl, A. E., Kannard, R. W., & Baldwin, K. F. (1975). Ridge regression:some simulations. Communications in Statistics, 4(2), 105–123. https://doi.org/10.1080/03610927508827232

Hoeting, J. A., Madigan, D., Raftery, A. E., & Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. Statistical Science, 14(4), 382–401. https://www.jstor.org/stable/2676803

Hyatt, M., Riley, J., McKee, M., & Israelsen, E. (1970). Computer Simulation of the Hydrologic-Salinity Flow System Within the Upper Colorado River Basin. Reports. https://digitalcommons.usu.edu/water_rep/383

Ibrahem Ahmed Osman, A., Najah Ahmed, A., Chow, M. F., Feng Huang, Y., & El-Shafie, A. (2021). Extreme gradient boosting (Xgboost) model to predict the groundwater levels in Selangor Malaysia. Ain Shams Engineering Journal, 12(2), 1545–1556. https://doi.org/10.1016/j.asej.2020.11.011

Jaafar, H. H., Ahmad, F. A., & El Beyrouthy, N. (2019). GCN250, new global gridded curve numbers for hydrologic modeling and design. Scientific Data, 6(1), Article 1. https://doi.org/10.1038/s41597-019-0155-x

Kalra, A., & Ahmad, S. (2011). Evaluating changes and estimating seasonal precipitation for the Colorado River Basin using a stochastic nonparametric disaggregation technique. Water Resources Research, 47(5). https://doi.org/10.1029/2010WR009118

Kenney, T. (2009). Scientific Investigations Report (Scientific Investigations Report) [Scientific Investigations Report].

Kingsford, C., & Salzberg, S. L. (2008). What are decision trees? Nature Biotechnology, 26(9), Article 9. https://doi.org/10.1038/nbt0908-1011

Kotsiantis, S. B. (2013). Decision trees: A recent overview. Artificial Intelligence Review, 39(4), 261–283. https://doi.org/10.1007/s10462-011-9272-4

Kratzert, F., Klotz, D., Herrnegger, M., Sampson, A. K., Hochreiter, S., & Nearing, G. S. (2019). Toward Improved Predictions in Ungauged Basins: Exploiting the Power of Machine Learning. Water Resources Research, 55(12), 11344–11354. https://doi.org/10.1029/2019WR026065

Krishnamurti, T. N., Kishtawal, C. M., LaRow, T. E., Bachiochi, D. R., Zhang, Z., Williford, C. E., Gadgil, S., & Surendran, S. (1999). Improved Weather and Seasonal Climate Forecasts from Multimodel Superensemble. Science, 285(5433), 1548–1550. https://doi.org/10.1126/science.285.5433.1548

Larose, D. T., & Larose, C. D. (2014). k-Nearest Neighbor Algorithm. In Discovering Knowledge in Data: An Introduction to Data Mining (pp. 149–164). Discovering Knowledge in Data: An Introduction to Data Mining. Wiley. https://doi.org/10.1002/9781118874059.ch7

Lee, D. J., Howitt, R. E., & Mariño, M. A. (1993). A stochastic model of river water quality: Application to salinity in the Colorado River. Water Resources Research, 29(12), 3917–3923. https://doi.org/10.1029/93WR02464

Ley, E., & Steel, M. F. J. (2009). On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. Journal of Applied Econometrics, 24(4), 651–674. https://doi.org/10.1002/jae.1057

Mahesh, B. (2018). Machine Learning Algorithms—A Review. International Journal of Science and Research (IJSR). 9. 10.21275/ART20203995.

Marjanović, M., Kovačević, M., Bajat, B., & Voženílek, V. (2011). Landslide susceptibility assessment using SVM machine learning algorithm. Engineering Geology, 123(3), 225–234. https://doi.org/10.1016/j.enggeo.2011.09.006

Miikkulainen, R., Liang, J., Meyerson, E., Rawal, A., Fink, D., Francon, O., Raju, B., Shahrzad, H., Navruzyan, A., Duffy, N., & Hodjat, B. (2019). Chapter 15—Evolving Deep Neural Networks. In R. Kozma, C. Alippi, Y. Choe, & F. C. Morabito (Eds.), Artificial Intelligence in the Age of Neural Networks and Brain Computing (pp. 293–312). Academic Press. https://doi.org/10.1016/B978-0-12-815480-9.00015-3

Miles, J. (2014). Tolerance and Variance Inflation Factor. In Wiley StatsRef: Statistics Reference Online. John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118445112.stat06593

Miller, M. P., Buto, S. G., Lambert, P. M., & Rumsey, C. A. (2017). Enhanced and updated spatially referenced statistical assessment of dissolved-solids load sources and transport in streams of the Upper Colorado River Basin. In Enhanced and updated spatially referenced statistical assessment of dissolved-solids load sources and transport in streams of the Upper Colorado River Basin (USGS Numbered Series 2017–5009; Scientific Investigations Report, Vols. 2017–5009, p. 34). U.S. Geological Survey. https://doi.org/10.3133/sir20175009

More, K. S., & Wolkersdorfer, C. (2023). Application of machine learning algorithms for nonlinear system forecasting through analytics—A case study with mining influenced water data. Water Resources and Industry, 29, 100209. https://doi.org/10.1016/j.wri.2023.100209

Naimi, A. I., & Balzer, L. B. (2018). Stacked generalization: An introduction to super learning. European Journal of Epidemiology, 33(5), 459–464. https://doi.org/10.1007/s10654-018-0390-z

Nauman, T. W., & Duniway, M. C. (2020). A hybrid approach for predictive soil property mapping using conventional soil survey data. Soil Science Society of America Journal, 84(4), 1170–1194. https://doi.org/10.1002/saj2.20080

Nauman, T. W., Ely, C. P., Miller, M. P., & Duniway, M. C. (2019). Salinity Yield Modeling of the Upper Colorado River Basin Using 30-m Resolution Soil Maps and Random Forests. Water Resources Research, 55(6), 4954–4973. https://doi.org/10.1029/2018WR024054

O'brien, R. M. (2007). A Caution Regarding Rules of Thumb for Variance Inflation Factors. Quality & Quantity, 41(5), 673–690. https://doi.org/10.1007/s11135-006-9018-6

Pacifico, A. (2021). Robust open Bayesian analysis: Overfitting, model uncertainty, and endogeneity issues in multiple regression models. Econometric Reviews, 40(2), 148–176. https://doi.org/10.1080/07474938.2020.1770996

PRISM Climate Group, O. S. U. (2014, February 4). PRISM Gridded Climate Data. PRISM Gridded Climate Data. https://prism.oregonstate.edu

Raftery, A. E., Gneiting, T., Balabdaoui, F., & Polakowski, M. (2005). Using Bayesian Model Averaging to Calibrate Forecast Ensembles. Monthly Weather Review, 133(5), 1155–1174. https://doi.org/10.1175/MWR2906.1

Raftery, A. E., Madigan, D., & Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. Journal of the American Statistical Association, 92(437), 179–191. https://doi.org/10.2307/2291462

Ranjan, R., Huang, B., & Fatehi, A. (2016). Robust Gaussian process modeling using EM algorithm. Journal of Process Control, 42, 125–136. https://doi.org/10.1016/j.jprocont.2016.04.003

Rodriguez-Galiano, V., Sanchez-Castillo, M., Chica-Olmo, M., & Chica-Rivas, M. (2015). Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews, 71, 804–818. https://doi.org/10.1016/j.oregeorev.2015.01.001

Rokach, L., & Maimon, O. (2005). Decision Trees. In O. Maimon & L. Rokach (Eds.), Data Mining and Knowledge Discovery Handbook (pp. 165–192). Springer US. https://doi.org/10.1007/0-387-25465-X_9

Rupp, D. E., Daly, C., Doggett, M. K., Smith, J. I., & Steinberg, B. (2022). Mapping an Observation-Based Global Solar Irradiance Climatology across the Conterminous United States. Journal of Applied Meteorology and Climatology, 61(7), 857–876. https://doi.org/10.1175/JAMC-D-21-0236.1

Sahour, H., Gholami, V., & Vazifedan, M. (2020). A comparative analysis of statistical and machine learning techniques for mapping the spatial distribution of groundwater salinity in a coastal aquifer. Journal of Hydrology, 591, 125321. https://doi.org/10.1016/j.jhydrol.2020.125321

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining Deep Neural Networks and Beyond: A Review of Methods and Applications. Proceedings of the IEEE, 109(3), 247–278. Proceedings of the IEEE. https://doi.org/10.1109/JPROC.2021.3060483

Snee, R. D. (1983). Regression Diagnostics: Identifying Influential Data and Sources of Collinearity. Journal of Quality Technology, 15(3), 149–153. https://doi.org/10.1080/00224065.1983.11978865

Soil Survey Staff, Natural Resources Conservation Service, & United States Department of Agriculture. (2022). Soil Survey Geographic (SSURGO) Database. Soil Survey Geographic (SSURGO) Database. https://sdmdataaccess.sc.egov.usda.gov

Strimbu, B. (2012). Correction for bias of models with lognormal distributed variables in absence of original data. Annals of Forest Research, 55(2), Article 2. https://doi.org/10.15287/afr.2012.66

Sun, B., Du, J., & Gao, T. (2009). Study on the Improvement of K-Nearest-Neighbor Algorithm. 2009 International Conference on Artificial Intelligence and Computational Intelligence, 4, 390–393. https://doi.org/10.1109/AICI.2009.312

Sun, S., & Huang, R. (2010). An adaptive k-nearest neighbor algorithm. 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery, 1, 91–94. https://doi.org/10.1109/FSKD.2010.5569740

Tamura, R., Kobayashi, K., Takano, Y., Miyashiro, R., Nakata, K., & Matsui, T. (2019). Mixed integer quadratic optimization formulations for eliminating multicollinearity based on variance inflation factor. Journal of Global Optimization, 73(2), 431–446. https://doi.org/10.1007/s10898-018-0713-3

Thompson, C. G., Kim, R. S., Aloe, A. M., & Becker, B. J. (2017). Extracting the Variance Inflation Factor and Other Multicollinearity Diagnostics from Typical Regression Results. Basic and Applied Social Psychology, 39(2), 81–90. https://doi.org/10.1080/01973533.2016.1277529

Tillman, F. D., Anning, D. W., Heilman, J. A., Buto, S. G., & Miller, M. P. (2018). Managing Salinity in Upper Colorado River Basin Streams: Selecting Catchments for Sediment Control Efforts Using Watershed Characteristics and Random Forests Models. Water, 10(6), Article 6. https://doi.org/10.3390/w10060676

Tillman, F. D., Coes, A. L., Anning, D. W., Mason, J. P., & Coplen, T. B. (2019). Investigation of recent decadal-scale cyclical fluctuations in salinity in the lower Colorado River. Journal

of Environmental Management, 235, 442–452. https://doi.org/10.1016/j.jenvman.2019.01.072

Tougui, I., Jilbab, A., & Mhamdi, J. E. (2021). Impact of the Choice of Cross-Validation Techniques on the Results of Machine Learning-Based Diagnostic Applications. Healthcare Informatics Research, 27(3), 189–199. https://doi.org/10.4258/hir.2021.27.3.189

U.S. Department of the Interior, B. of R. (2011). QUALITY OF WATER COLORADO RIVER BASIN (23).

USGS. (2021). USGS Streamgaging Network | U.S. Geological Survey. https://www.usgs.gov/mission-areas/water-resources/science/usgs-streamgaging-network

USGS. (2021). USGS Streamgaging Network | U.S. Geological Survey. https://www.usgs.gov/mission-areas/water-resources/science/usgs-streamgaging-network

Uyanık, G. K., & Güler, N. (2013). A Study on Multiple Linear Regression Analysis. Procedia - Social and Behavioral Sciences, 106, 234–240. https://doi.org/10.1016/j.sbspro.2013.12.027

Venkatesan, A. K., Ahmad, S., Johnson, W., & Batista, J. R. (2011). Systems dynamic model to forecast salinity load to the Colorado River due to urbanization within the Las Vegas Valley. Science of The Total Environment, 409(13), 2616–2625. https://doi.org/10.1016/j.scitotenv.2011.03.018

Vu, D. H., Muttaqi, K. M., & Agalgaonkar, A. P. (2015). A variance inflation factor and backward elimination based robust regression model for forecasting monthly electricity demand using climatic variables. Applied Energy, 140, 385–394. https://doi.org/10.1016/j.apenergy.2014.12.011

Wang, Y., Khodadadzadeh, M., & Zurita-Milla, R. (2023). Spatial+: A new cross-validation method to evaluate geospatial machine learning models. International Journal of Applied Earth Observation and Geoinformation, 121, 103364. https://doi.org/10.1016/j.jag.2023.103364

White, M. J., Arnold, J. G., Bieger, K., Allen, P. M., Gao, J., Čerkasova, N., Gambone, M., Park, S., Bosch, D. D., Yen, H., & Osorio, J. M. (2022). Development of a Field Scale SWAT+ Modeling Framework for the Contiguous U.S. JAWRA Journal of the American Water Resources Association, 58(6), 1545–1560. https://doi.org/10.1111/1752-1688.13056

Yang, Y., Sun, H., Xue, J., Liu, Y., Liu, L., Yan, D., & Gui, D. (2021). Estimating evapotranspiration by coupling Bayesian model averaging methods with machine learning algorithms. Environmental Monitoring and Assessment, 193(3), 156. https://doi.org/10.1007/s10661-021-08934-1

Yu, X., Wang, Y., Wu, L., Chen, G., Wang, L., & Qin, H. (2020). Comparison of support vector regression and extreme gradient boosting for decomposition-based data-driven 10-day streamflow forecasting. Journal of Hydrology, 582, 124293. https://doi.org/10.1016/j.jhydrol.2019.124293

# CHAPTER 3 AI-ENHANCED URBAN COGNITION: CONVLSTM NETWORKS FOR MULTI-SCENARIO LAND COVER PREDICTION IN METROPOLITAN REGIONS[2]


## ABSTRACT

Traditional urban growth models, such as Cellular Automata (CA), often struggle to capture complex, non-linear temporal dependencies over long time scales due to their reliance on static transition probabilities. This study develops a Convolutional Long Short-Term Memory (ConvLSTM) framework applied to a uniquely dense 39-year annual land cover dataset (1985-2023) across Colorado's Metropolitan Planning Organization areas to address limitations in long-term spatiotemporal urban growth prediction. The framework integrates stakeholder-driven scenario planning to evaluate five alternative development trajectories ranging from business-as-usual to high-growth futures. Analysis reveals a counterintuitive "Temporal Depth Paradox" where prediction accuracy (F1 score) increases substantially from 0.27 for 1-year forecasts to 0.87 for 20-year forecasts, challenging fundamental assumptions in urban forecasting. The model effectively distinguishes between different development processes, showing higher predictive accuracy for greenfield expansion versus complex urban redevelopment. By embedding ecological constraints and evaluating alternative growth scenarios, the framework demonstrates how AI can support sustainable development goals while maintaining human oversight in planning processes. Quantitative scenario analysis reveals that compact development patterns (Adaptive Innovation


[2] Mohamed F. Mahmoud, Mazdak Arabi. "AI-Enhanced Urban Cognition: ConvLSTM Networks for Multi-Scenario Land Cover Prediction in Metropolitan Regions," *Research Square* (Preprint), Version 1, 04 September 2025. https://doi.org/10.21203/rs.3.rs-7348736/v1

scenario) produces 20.5% infrastructure efficiency improvements and preserve approximately 1,000 additional hectares of natural areas by 2050 compared to unregulated growth scenarios. The findings provide planners with evidence-based tools to evaluate how different policy interventions might reshape urban morphology, enhance infrastructure efficiency, and preserve natural resources in rapidly growing regions.

## 3.1 INTRODUCTION

The integration of artificial intelligence into urban planning represents a fundamental shift of how we understand, design, and manage cities. As urban landscapes evolve under mounting pressures from population growth and urbanization, climate variability, and limited resources, AI technologies offer advanced capabilities to envision and shape sustainable urban futures (Anwar & Sakti, 2024; Kamrowska-Załuska, 2021). This paper develops a deep learning methodology that integrates spatiotemporal prediction with scenario-based planning to create a predictive framework for understanding and guiding urban development in alignment with Sustainable Development Goal 11-Sustainable Cities and Communities.

Land cover and land use change prediction has traditionally relied on statistical models, cellular automata, and agent-based simulations; approaches that, while valuable, often fail to capture the complex spatiotemporal dynamics inherent in urban growth processes (Tang & Bennett, 2010). Statistical models typically fail to adequately represent non-linear relationships that characterize land cover transformations (Li & and Yeh, 2002), while cellular automata models, though promising for simulating spatial dynamics through transition rules (Mustafa et al., 2017), frequently overlook critical temporal dependencies. Similarly, agent-based models can represent

behavioral aspects of land use decisions (Bakker et al., 2015) but face computational scalability challenges when applied to regional contexts.

The integration of scenario planning with urban modeling has matured significantly, establishing robust frameworks that our research builds upon. Established approaches now encompass participatory model co-creation with local stakeholders (Kim et al., 2022), game-theoretic modeling of multi-actor cooperation dynamics (Chu et al., 2020), holistic climate risk methodologies integrating physical and socio-economic vulnerabilities (Gandini et al., 2021), systematic frameworks for optimal nature-based solution placement (Cong et al., 2023), and interactive platforms for translating scientific outputs into stakeholder narratives (Miles et al., 2024). These advances establish that stakeholder-engaged, sustainability-oriented urban planning frameworks are well-developed. This research contribution addresses a different dimension: enhancing the predictive engines that generate spatially explicit forecasts within these frameworks, specifically for contexts requiring long-term (20-30 year) projections based on dense temporal data where traditional Cellular Automata models face architectural limitations.

The predictive engines underlying these scenario frameworks typically rely on Cellular Automata (CA) models, which have emerged as the dominant benchmark for spatial urban simulation. CA models simulate land use transitions through local rules applied to grid cells, often coupled with Markov chains (CA-Markov) to define transition probabilities based on historical patterns (Gui et al., 2025). Recent comprehensive reviews demonstrate that CA-based approaches excel in several key areas: computational efficiency, interpretability of transition rules, and strong performance for short-term (1-5 year) forecasting when temporal data is sparse (Gui et al., 2025; Yaagoubi et al., 2024). Furthermore, CA models can be enhanced with machine learning components -such as

neural networks for transition rule learning (MLP-CA) or genetic algorithms for parameter optimization- to improve predictive accuracy while maintaining computational tractability.

However, the fundamental architecture of CA-Markov models imposes important limitations for certain forecasting contexts. The Markovian assumption that future states depend only on the current state, means that CA models rely on static transition probability matrices that cannot capture complex, non-linear temporal dependencies spanning multiple decades. While this simplification is advantageous for data-sparse environments and enables fast computation, it becomes limited when: (1) dense, multi-decadal time series are available, (2) long-term (20-30 year) forecasts are needed, and (3) the goal is to learn deep temporal patterns (cycles, trends, path dependencies) rather than immediate transitions. In such data-rich, long-horizon contexts, models explicitly designed for temporal sequence learning may offer advantages.

The emergence of deep learning approaches offers new opportunities by leveraging the capability to learn complex multidimensional patterns from large datasets (Mahmoud et al., 2025; Zhang et al., 2018). Specifically, Convolutional Long Short-Term Memory (ConvLSTM) networks have demonstrated remarkable potential for capturing both spatial relationships and temporal sequences simultaneously (Kim et al., 2022). By integrating convolutional operations with recurrent network structures, ConvLSTM architectures effectively model the spatiotemporal dependencies inherent in land cover dynamics, enabling more accurate predictions of future development patterns (Cresson, 2019; Mc Cutchan & Giannopoulos, 2022).

This research addresses the gap by pairing a Convolutional Long Short-Term Memory (ConvLSTM) architecture -a model explicitly designed for spatiotemporal sequences- with a uniquely long and dense 39-year land use annual dataset. This combination enables the discovery of deep temporal patterns in urban growth that simpler models, by design, cannot detect. Research

objectives include: (1) developing a ConvLSTM neural network architecture optimized for long-term urban land cover prediction using a 39-year annual dataset; (2) integrating this model with the Colorado Water Plan scenario framework to generate spatially explicit land cover predictions under five alternative futures; (3) quantifying the implications of different growth trajectories for urban form, infrastructure efficiency, and environmental conservation; and (4) evaluating how policy interventions (conservation constraints, transit-oriented development) modify predicted spatial patterns. This approach demonstrates how deep learning models can support evidence-based planning by revealing long-term spatiotemporal patterns while maintaining human control over planning objectives (Kurniawan et al., 2024; Peng et al., 2024).

The research presented here contributes to both methodological innovation in land cover prediction and AI's role in urban sustainability planning. By bridging sophisticated deep learning techniques with scenario-based planning, it demonstrates how computational approaches can enhance our understanding of potential future landscapes while accommodating the uncertainties inherent in complex social-ecological systems. This work directly supports SDG 11 targets by providing a framework that identifies patterns of urban expansion that optimize infrastructure efficiency while preserving ecological systems. The findings offer valuable insights for urban planners, water resource managers, and policymakers working to guide urban growth toward more sustainable, resilient, and equitable development patterns while maintaining human-centered planning priorities. Such applications include transit-oriented design that concentrates development around public transportation nodes, green infrastructure planning that integrates natural systems with urban development, affordable housing allocation strategies that promote social equity, and climate-resilient development patterns that minimize vulnerability to environmental hazards.

This research is conducted across Colorado's Metropolitan Planning Organization (MPO) areas, which represent major urban regions facing significant growth and sustainability challenges (Figure 3- 1). These federally designated planning entities provide an ideal spatial framework for analyzing AI-enhanced urban growth prediction and scenario-based planning approaches that can inform sustainable development strategies across multiple jurisdictions.

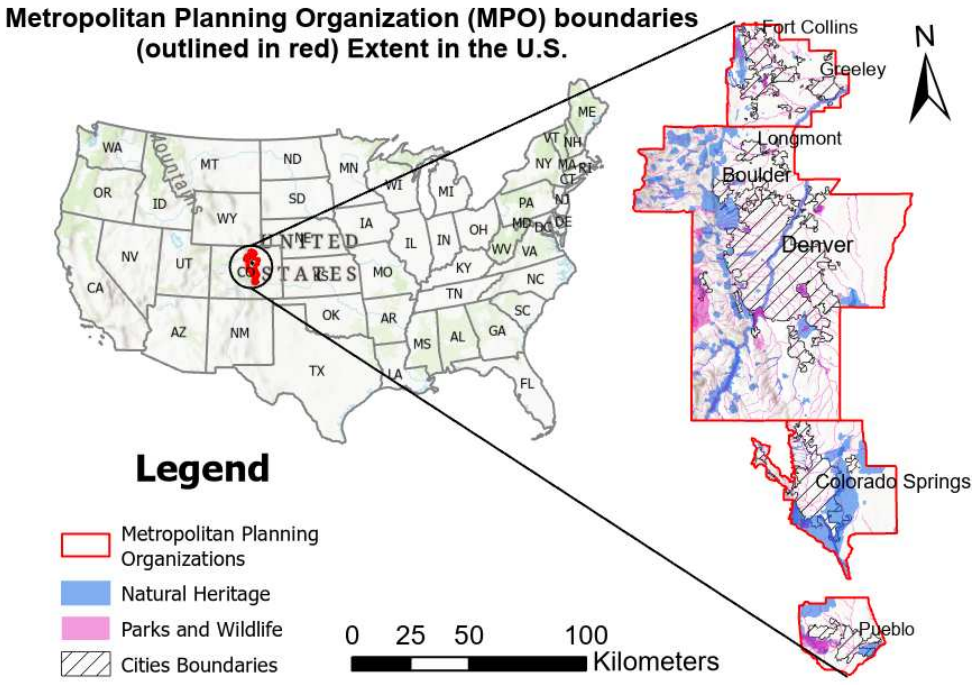


Figure 3- 1 Metropolitan Planning Organization (MPO) areas in Colorado serving as AI testing grounds for sustainable urban development prediction.

This chapter addresses three objectives: (1) develop a deep learning framework using ConvLSTM neural networks capable of learning complex spatiotemporal patterns from a 39-year annual land cover dataset for long-term urban growth prediction; (2) integrate scenario-based planning with deep learning to generate spatially-explicit land use maps under five Colorado Water Plan scenarios through 2050; and (3) embed policy constraints (conservation areas, transit-oriented

development) into the prediction framework to ensure outputs respect regulatory requirements and planning priorities.

## 3.2 METHODOLOGY

This research develops a ConvLSTM-based methodology for long-term land cover prediction, integrated with multi-scenario analysis to explore alternative urban development trajectories. The methodology operates through three sequential phases (Figure 3- 2). (1) training a deep learning model on 39 years of historical data to learn spatiotemporal patterns, (2) translating scenario-based population projections into land area demands through logistic regression, and (3) integrating learned probability distributions with scenario demands through a constrained rank-allocation algorithm that respects conservation priorities and optional policy interventions. This structure separates pattern learning from quantity specification, enabling systematic evaluation of how different growth assumptions and policy choices might reshape urban landscapes through 2050. The following subsections detail each component's technical implementation and explain how they interconnect to produce spatially explicit, policy-relevant predictions.

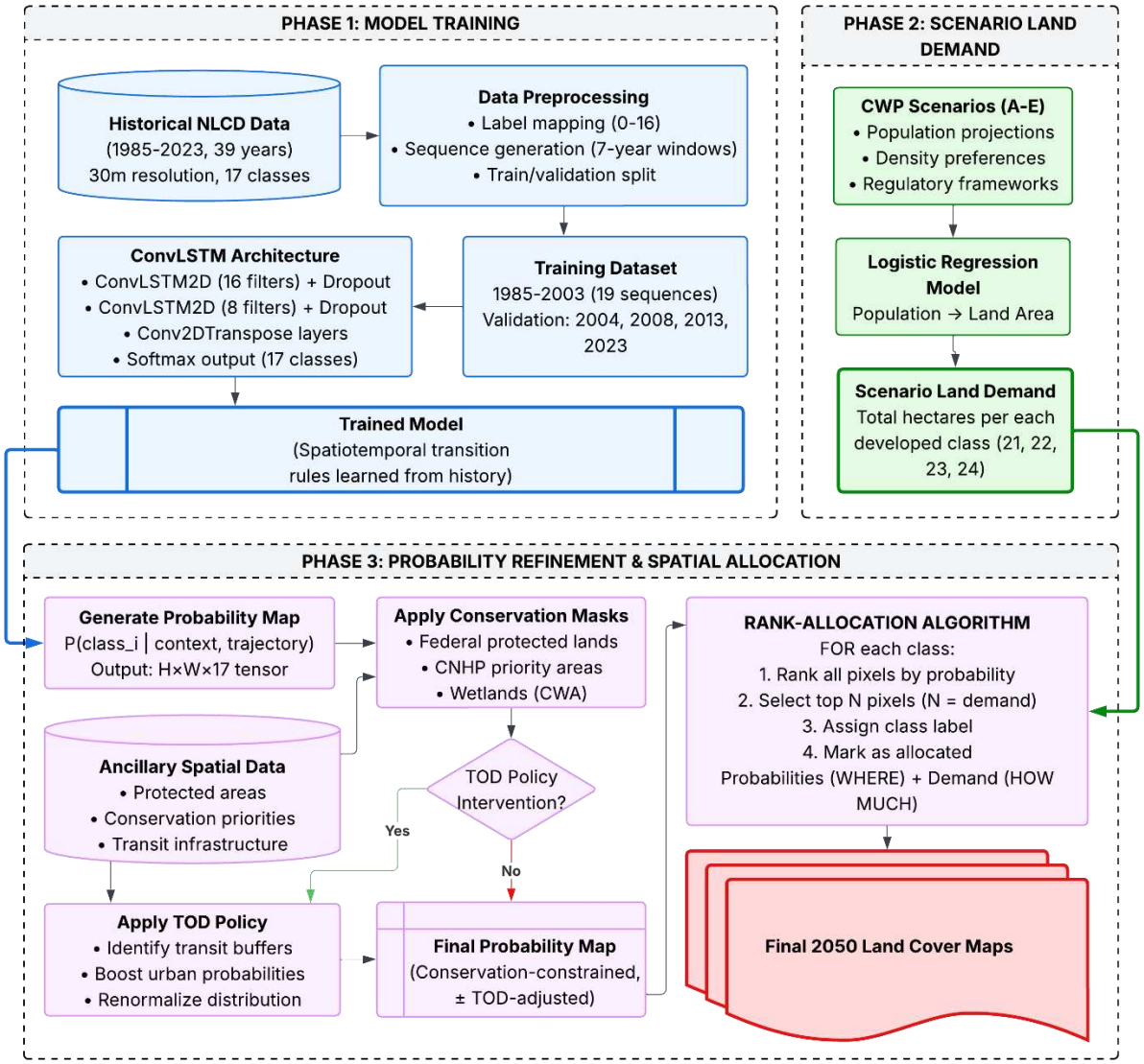


Figure 3- 2 Three-phase methodology framework integrating deep learning, scenario planning, and constrained spatial allocation.

### 3.2.1 DEEP LEARNING-BASED SPATIOTEMPORAL MODELING WITH CONVLSTM

The Convolutional Long Short-Term Memory (ConvLSTM) neural network architecture (Figure 3- 2, Phase 1) addresses a critical limitation in urban growth prediction: the need to simultaneously model spatial autocorrelation and temporal dependencies. Unlike traditional neural networks that treat either spatial or temporal dimensions in isolation, ConvLSTM processes both dimensions

through a unified architecture, enabling detection of complex spatiotemporal patterns that emerge over multi-decadal timescales (Al-Raeei, 2025; Peng et al., 2024).

The selection of ConvLSTM was driven by our specific research context and recent comparative literature. Cellular Automata (CA) models, particularly CA-Markov, represent the established benchmark for urban simulation (Gui et al., 2025). Recent comparative studies demonstrate that in data-sparse, short-term forecasting contexts, CA-Markov outperforms ConvLSTM: (Yaagoubi et al., 2024) found CA-Markov achieved 97-98% accuracy versus 94.7% for ConvLSTM using 13 images over 6 years, attributing the difference to insufficient training data for deep learning.

However, this research addresses a fundamentally different problem: long-term (20+ year) scenario-oriented forecast with a dense 39-year annual dataset. In this data-rich context, CA-Markov's core limitation becomes critical. The Markovian assumption -that future states depend only on the current state- means CA models use static transition probability matrices that cannot capture multi-year development cycles, decadal policy trends, or path-dependent trajectories. ConvLSTM's architecture addresses this through: (1) convolutional layers that capture spatial autocorrelation where neighboring pixels influence transitions (Wang & Yorke-Smith, 2025); (2) LSTM gates that maintain temporal memory across multiple years, learning which long-term patterns to retain; and (3) end-to-end learning of non-linear spatiotemporal relationships from data rather than predefined rules (Peng et al., 2024). This demonstrates that model selection should be context-dependent: CA-Markov excels for short-term prediction with sparse data; ConvLSTM excels for long-term prediction with dense multi-decadal sequences.

The implemented ConvLSTM architecture creates a deep learning architecture through a sequence of layers designed to progressively extract spatiotemporal features from input data (Figure 3- 2). The network begins with two ConvLSTM2D layers that process input sequences of land cover

data. The first ConvLSTM2D layer contains 16 filters with a 3×3 kernel size, processing the input with padding to maintain spatial dimensions. This layer returns the full sequence, passing it to a second ConvLSTM2D layer with 8 filters that returns only the final state, effectively compressing the temporal information into a spatial representation. This design enables the network to identify both short-term transitions and long-term development trajectories that would remain hidden to human analysts or traditional models (Kamrowska-Załuska, 2021).

To mitigate overfitting and enhance the model's generalizability across diverse urban contexts, Dropout layers (rate=0.2) are inserted after each ConvLSTM2D layer, randomly deactivating neurons during training. Batch Normalization layers standardize activations, accelerating training and improving stability. The network then expands the spatial representation using a Conv2DTranspose layer with 32 filters, followed by another Batch Normalization layer. The final Conv2DTranspose layer produces outputs with the same spatial dimensions as the input, with the number of filters matching the number of land cover categories (17). A softmax activation function ensures outputs can be interpreted as probabilities across different land cover classes (Kouretas & Paliouras, 2019).

Land cover data typically exhibits significant class imbalance, with certain categories occupying much larger areas than others (Lalenis, 2025). This imbalance can bias models toward majority classes at the expense of accurately predicting transitions in minority classes, which are often of particular interest in urban growth studies and critical for inclusive planning. To counter this algorithmic bias, a custom weighted categorical cross-entropy loss function is implemented to dynamically adjust the importance of each class:

$$L(y, \hat{y}) = -\sum_{i=1}^{N} \sum_{j=1}^{C} w_i y_{ij} \log(\hat{y}_{ij}) \quad (3\text{-}1)$$

where $y_{ij}$ is the true value (0 or 1) for pixel $i$ and class $j$, $\hat{y}_{ij}$ is the predicted probability, and $w_j$ is the weight for class $j$.

The class weights are calculated dynamically during training using an inverse frequency approach:

$$w_j = \left(\frac{f_j + \in}{\sum_{k=1}^{C} f_k}\right)^{-\alpha} \quad (3\text{-}2)$$

where $f_j$ is the frequency of class $j$ in the batch, $\in$ is a small constant ($1e^{-8}$) to prevent division by zero, and $\alpha$ is a weighting factor (set to 0.5 based on experimentation). These weights are normalized to sum to 1, effectively increasing the importance of underrepresented classes during training and ensuring the model's predictions don't reinforce existing biases in urban development patterns (Michailidis et al., 2024).

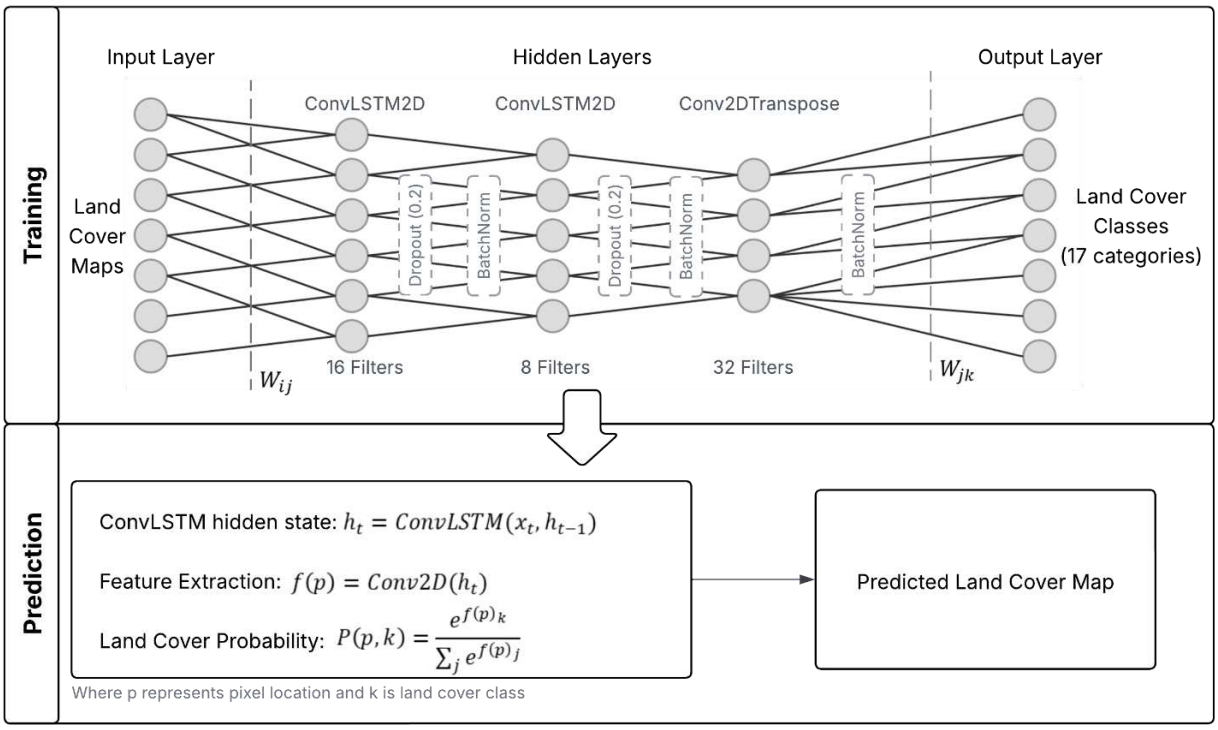


Figure 3- 3 ConvLSTM Model Architecture for Land Cover Prediction

A two-phase training strategy is employed to enhance model robustness and validation. In the first phase, the model was trained on data from 1985 to 2003 and then used to generate predictions for selected validation years (2004, 2008, 2013, and 2023). These predictions are compared against actual land cover maps to assess model performance across different temporal horizons (1, 5, 10, and 20 years). This validation phase provided insights into how prediction accuracy varied with the length of the forecast period and across different landscape contexts, essential for understanding the model's capabilities for long-term urban planning applications (Tékouabou et al., 2022).

In the second phase, the model is retrained using the complete dataset (1985-2023) to incorporate all available historical information for generating future predictions for 2050. This full model benefited from the expanded temporal coverage, allowing it to learn from more recent land use transitions that might better reflect contemporary development patterns. The result is not merely a predictive tool but an analytical framework that transforms how planners can understand the complex spatiotemporal dynamics of urban growth, enabling more informed and sustainable planning decisions (Peng et al., 2024; Souza et al., 2022). The trained ConvLSTM model outputs probability distributions that feed into Phase 3 of the framework (Figure 3- 2).

### 3.2.2 MULTI-SCENARIO ANALYSIS FRAMEWORK

This scenario-based approach (Figure 3- 2, Phase 2) creates an analytical framework that enhances urban planning in three keyways. First, it provides a structured method for incorporating different social, economic, and environmental narratives into the AI prediction system, addressing the complexity of urban development that cannot be captured in single forecasts (Tékouabou et al., 2022). Second, it enables the exploration of different policy interventions and their spatial

implications, facilitating a more representative planning process that considers multiple stakeholder perspectives (Chauncey & McKenna, 2024). Third, it transforms AI from a prescriptive to an exploratory tool that supports human decision-making while acknowledging fundamental uncertainties in urban development trajectories (Kurniawan et al., 2024).

This study incorporated the five scenario frameworks from the Colorado Water Plan (CWP) to create a predictive modeling system that explores alternative futures for land cover change. While the CWP scenarios are defined by nine high-impact drivers (CWCB, 2015), this study focuses on the three most directly relevant to spatial urban development patterns: population growth, urban land use density, and regulatory frameworks (Figure 3- 4). Rather than producing singular predictions, this approach transforms AI from a deterministic tool into an analytical framework that explores multiple plausible futures, addressing concerns about over-reliance on AI predictions and maintaining human agency in planning processes (Michailidis et al., 2024).

Using 2023 as baseline (MPO area population: 4,478,161), scenarios with varying growth trajectories for the study area are implemented: Business as Usual (Scenario A: 6,484,488, +45%) represents continuation of recent trends with no change in density or regulations; Weak Economy (Scenario B: 6,006,735, +34%) considers slower economic growth with unchanged urban patterns; Cooperative Growth (Scenario C: 6,457,129, +44%) emphasizes higher density development with increased regulations; Adaptive Innovation (Scenario D: 7,041,719, +57%) combines accelerated growth with higher density and increased regulatory frameworks; and Hot Growth (Scenario E: 6,997,952, +56%) features rapid growth with lower density preferences and decreased regulations. These scenarios represent diverse potential futures that might emerge from different combinations of demographic, planning, and policy conditions.

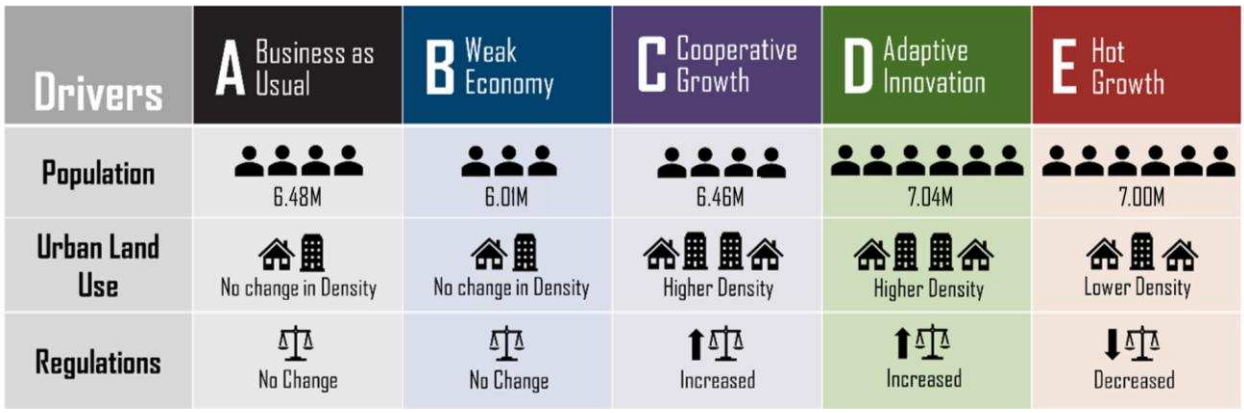


Figure 3- 4 Colorado Water Plan Scenario Characteristics for the Three Key Drivers Implemented in the ConvLSTM Urban Growth Modeling Framework (CWCB, 2015).

A logistic regression model is developed to translate population projections into future land use patterns for the ConvLSTM model (Figure 3- 5). Using 39 years of annual NLCD data (1985-2023), a relationships between population growth and land area changes are established for each urban category (Open Space, Low Intensity, Medium Intensity, and High Intensity) using the following logistic model equation:

$$Y_i = \frac{Y_{max_i}}{1+(\frac{Y_{max_i}}{Y_{0_i}}-1)\, e^{(-k_i t_i)}} \tag{3-3}$$

where, $Y_i$ denotes the area of a land use category $i$ in acres, $Y_{max_i}$ is the maximum possible area for land use category $i$, $Y_{0_i}$ represents the land use area $i$ at the first available data point, $t_i$ is the change in population between two steps, $k_i$ is the growth/decline rate for land use category $i$.

The logistic regression model was applied to translate population projections into land use changes across the diverse future conditions outlined in the Colorado Water Plan. The specific characteristics of these planning scenarios are illustrated in Figure 3-4. The logistic regression models demonstrate strong explanatory power with $R^2$ values of 0.97 for high-intensity

development, 0.86 for medium-intensity development, 0.99 for low-intensity development, 0.76 for open space, and 0.97 for other categories, indicating robust relationships between population growth and land use changes across the historical period.

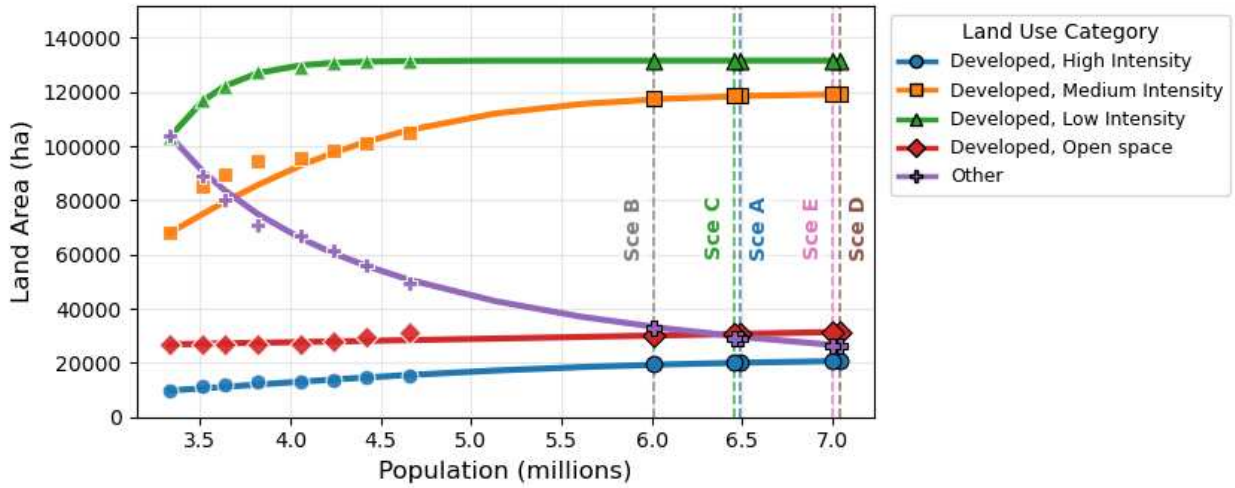


Figure 3- 5 Logistic regression model translating population to land use changes across Colorado Water Plan scenarios.

Logistic regression relationships are used exclusively to establish total regional land demand for each urban category under a given scenario. They answer the question "how much" total development will occur. The spatial allocation of this demand -answering "where" development occurs- is determined by the ConvLSTM-learned spatiotemporal patterns through the rank-allocation algorithm illustrated in Figure 3- 6. The ConvLSTM model learns purely from historical data (1985-2023) and generates probability maps without knowledge of scenarios. The rank-allocation algorithm then integrates WHERE (high-probability pixels from ConvLSTM) with HOW MUCH (scenario-derived demand) by selecting the top N pixels for each class where N equals the demand quantity. This separation enables scenario exploration while maintaining historical pattern fidelity (Figure 3- 2, Phase 3).

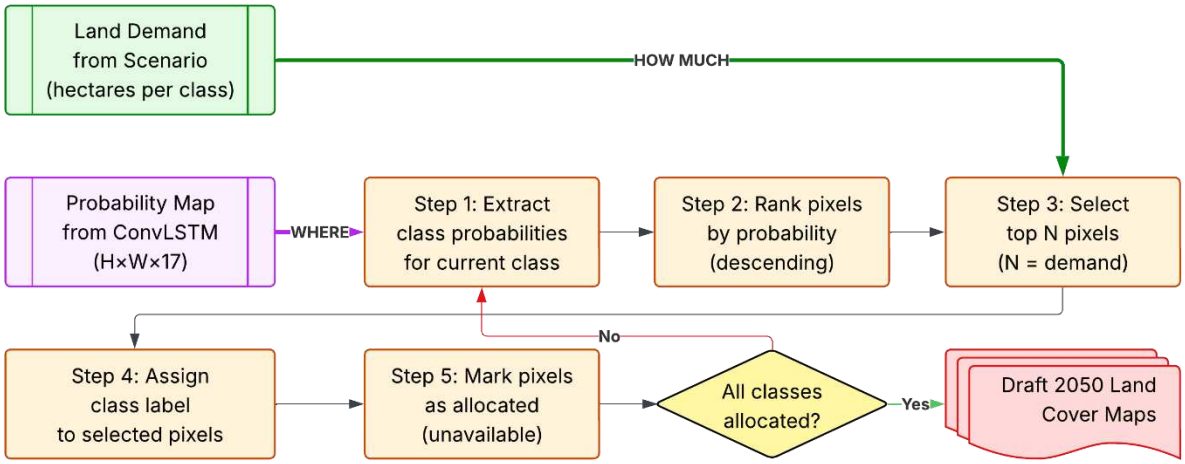


Figure 3- 6 Rank-allocation algorithm integrating scenario-based land demand with ConvLSTM probability predictions.

### 3.2.3 EMBEDDING REGULATORY AND CONSERVATION CONSTRAINTS

Purely data-driven models may predict development in legally protected areas or ecologically sensitive zones, rendering predictions unrealistic for planning applications. To ensure predictions are policy-relevant and legally compliant, we integrate human-defined ecological and regulatory constraints into the spatial allocation process (Figure 3- 2, Phase 3). This approach ensures predictions respect legal frameworks while maintaining the model's ability to identify optimal development patterns within established boundaries (Abbass et al., 2024; He & Chen, 2024).

The system embeds human-defined ecological values and legal constraints directly into the modeling process, providing transparent human oversight while allowing the AI to identify optimal development patterns within established boundaries (Wang & Yorke-Smith, 2025). The model incorporates legally protected lands where development is prohibited, including areas under Title 16, NEPA, and ESA protections, as well as wetlands protected by the Clean Water Act. Colorado Natural Heritage Program (CNHP) data is integrated to identify high conservation priority areas

based on biodiversity significance (CNHP, 2023). The constraints are applied during the probability refinement stage (Figure 3- 2), after the ConvLSTM generates raw transition probabilities but before the rank-allocation algorithm assigns final land uses. That allows the spatial allocation to naturally avoid protected areas because they have near-zero probability of urban conversion.

The implementation uses binary masks for each protected category, modifying the model-generated probability distributions within protected areas. This ensures these areas maintain their natural land cover in all future predictions, demonstrating how AI systems can respect legal frameworks while providing planning insights (Michailidis et al., 2024). This transforms the AI from a purely predictive tool to a system supporting sustainable development goals (Leal filho et al., 2024).

Conservation constraints are applied during probability refinement in Phase 3 (Figure 3- 2), after the trained ConvLSTM generates raw probability predictions but before the rank-allocation algorithm assigns final land uses. For each pixel within protected areas (federal lands, CNHP B1-B3 priority areas, wetlands), the probability distribution is modified by multiplying the current land cover class probability by large weights and renormalizing to maintain a valid probability distribution. This results in protected areas having near-zero probability of transitioning to urban uses. The rank-allocation algorithm then operates on these refined probabilities, naturally avoiding protected areas because they rank extremely low for any urban transition. This approach differs fundamentally from hard masking, which would create spatial discontinuities and waste allocated demand on pixels that are subsequently erased. By embedding constraints within the probability distribution itself, we ensure the spatial allocation naturally respects legal and ecological boundaries while maintaining the probabilistic framework throughout.

### 3.2.4 TRANSIT-ORIENTED DEVELOPMENT AS SUSTAINABILITY INTERVENTION

The modeling framework incorporated transit-oriented development (TOD) policies to demonstrate how different policy interventions can be evaluated within the ConvLSTM-scenario integration structure. This approach in the study region reflects Colorado's recent emphasis on transportation-land use coordination through (House Bill 24-1313, 2024), aligning with sustainability principles advocated by (Allam et al., 2022). TOD integration demonstrates a core capability of the methodology, the ability to modify probability distributions based on policy assumptions, enabling systematic comparison of alternative planning strategies.

TOD policy is implemented through the same probability refinement mechanism used for conservation constraints (Section 3.2.3), but with opposite effect, boosting rather than suppressing urban development probabilities. After conservation masks are applied but before rank-allocation, pixels within transit-oriented development zones have their urban class probabilities (classes 21, 22, 23, 24) multiplied by location-specific weights (×1.5-2.0 based on distance from transit) and renormalized, creating spatial gradients that favor transit-adjacent development during rank-allocation.

To demonstrate the framework's policy evaluation capability, we apply this comparative analysis to Scenario A, generating two versions: (1) a regulated condition incorporating both conservation constraints (Section 3.2.2) and TOD probability boosts, and (2) an unregulated condition with no probability modifications. Both versions allocate identical total development but produce different spatial distributions. This paired comparison enables stakeholders to quantify how integrated sustainability frameworks (conservation protections + transit coordination) might reshape development patterns compared to unregulated growth while accommodating the same population

projection, providing insights into spatial trade-offs between development accommodation and conservation objectives. Here, we demonstrate how AI can serve as a policy exploration tool that enhances human decision-making rather than replacing it (Michailidis et al., 2024; Wang & Yorke-Smith, 2025).

### 3.2.5 PERFORMANCE EVALUATION THROUGH MULTI-SCALE VALIDATION

Traditional pixel-by-pixel accuracy metrics often fail to capture the spatial context of land use change predictions, particularly for long-term forecasts where exact location accuracy is less critical than capturing broader change patterns. To address this limitation and create a more robust evaluation framework, we implemented distance-based accuracy metrics using buffer zones of 150m (5 pixels) and 300m (10 pixels) to evaluate the model's spatial prediction performance (Brenning, 2023).

This multi-scale validation approach acknowledges that predictions falling within a reasonable distance of actual changes still provide valuable planning information, even if the exact pixel location differs slightly. It represents a more nuanced way to evaluate AI performance in spatiotemporal domains, moving beyond simplistic accuracy metrics to consider the practical utility of predictions for planning applications (Tékouabou et al., 2022). For each validation year (2004, 2008, 2013, and 2023), three key performance metrics are calculated within the specified buffer distances (Table 3- 1). Precision measures the reliability of predictions by calculating the proportion of predicted changes that match actual changes. Recall assesses completeness by measuring the proportion of actual changes that are correctly predicted. The F1 score provides a balanced measure combining both precision and recall through their harmonic mean. All metrics range from 0 to 1, with values closer to 1 indicating better performance.

Table 3- 1 Performance Metrics for Spatial Prediction Validation

| Metric | Formula | Interpretation |
|---|---|---|
| Precision | $\frac{TP}{TP + FP}$ | Proportion of predicted changes that were correct; higher values indicate fewer false alarms. |
| Recall | $\frac{TP}{TP + FN}$ | Proportion of actual changes that were detected; higher values indicate better completeness. |
| F1 Score | $\frac{2 * (Precision * Recall)}{(Precision + Recall)}$ | Harmonic mean balancing precision and recall; 1.0 indicates perfect performance |

Where TP = True Positives, FP = False Positives, FN = False Negatives

To evaluate the model's performance across different time frames, a multi-year validation approach using historical data from several distinct time points is considered (Koko et al., 2020). The model is initially trained on data from 1985-2003 and then validated against observed changes in 2004, 2008, 2013, and 2023, corresponding to 1, 5, 10, and 20 years beyond the training period. This strategy allow to assess how prediction accuracy varies with temporal distance, providing insights into the model's reliability for both short-term and long-term forecasts.

### 3.2.6 URBAN FORM ANALYSIS

To characterize how different scenarios and policies reshape urban morphology, a suite of landscape metrics adapted from landscape ecology but tailored specifically for urban contexts are employed. This approach transforms AI from a predictive tool to an analytical framework that reveals emergent patterns in urban form across scenarios, creating new cognitive capabilities for understanding urban complexity (Kamrowska-Załuska, 2021; Peng et al., 2024).

For analyzing development trajectories across different scenarios and regulatory conditions, a Development Intensity index is calculated following established methodologies for quantifying

urban development patterns through weighted spatial analysis (Terzi & Kaya, 2011). The index assigns weighted values to each NLCD urban land cover category based on their development density: Open Space (NLCD 21) = 0.2, Low Intensity (NLCD 22) = 0.4, Medium Intensity (NLCD 23) = 0.7, and High Intensity (NLCD 24) = 1.0, with all non-urban categories assigned a value of 0.0. For each analysis location, the Development Intensity index is computed as the mean of weighted values within a 30×30 pixel moving window (900 m × 900 m at 30-meter resolution), this technique captures neighborhood effects on urban intensity patterns (Wu et al., 2006). This approach enables tracking of development intensity changes over the historical period (1985-2023) and comparison of projected intensities under different 2050 scenarios, with values ranging from 0.0 (no development) to 1.0 (maximum high-intensity development), providing a quantitative framework for assessing how regulatory constraints and policy scenarios influence local development trajectories.

Following the approach outlined by Fang et al. (2016), patch-based metrics including Patch Density (PD), Edge Density (ED), Effective Mesh Size (EMS), and Largest Patch Index (LPI) are calculated to quantify the fragmentation and concentration patterns of new urban development across scenarios. These metrics provide insights into how different growth trajectories might influence urban consolidation versus sprawl, key considerations for sustainable urban form. Edge and shape metrics provide further insights into the complexity and compactness of projected urban forms. Edge Density were calculated to assess the irregularity and complexity of new development patterns (Tv et al., 2012). These metrics are particularly relevant for infrastructure planning, as more aggregated development patterns typically allow for more efficient service delivery, a key consideration for sustainable urban forms that minimize resource consumption and environmental impacts (Al-Raeei, 2025; Tékouabou et al., 2022).

To evaluate the infrastructure efficiency implications of ConvLSTM-predicted spatial development patterns, this analysis integrates empirically-derived relationships from Colorado county data that establish power-law scaling between population and road infrastructure requirements, consistent with universal urban scaling principles demonstrated across multiple urban systems (Bettencourt et al., 2007). Local road infrastructure follows the relationship $L = 0.2P^{0.72}$ ($R^2 = 0.95$), while major roads scale as $L = 0.04P^{0.72}$ ($R^2 = 0.93$), where $L$ represents road length in kilometers and $P$ represents population (Figure 3- 7). These sub-linear scaling relationships ($\beta < 1$) demonstrate economies of scale in infrastructure provision and align with theoretical predictions for material infrastructure networks (Bettencourt et al., 2007). The key innovation lies in applying these relationships to ConvLSTM-identified urban patches rather than aggregate populations. The ConvLSTM model generates spatially explicit predictions of where different urban development intensities will occur, which are then segmented into contiguous development patches using connected component analysis. Each patch receives population allocation based on its land use composition through an MLR impervious area model, and Colorado relationships are applied at the patch level. This approach captures how spatial fragmentation versus consolidation—as revealed by the ConvLSTM model—affects total infrastructure requirements through both individual patch economies of scale and inter-patch connectivity costs.

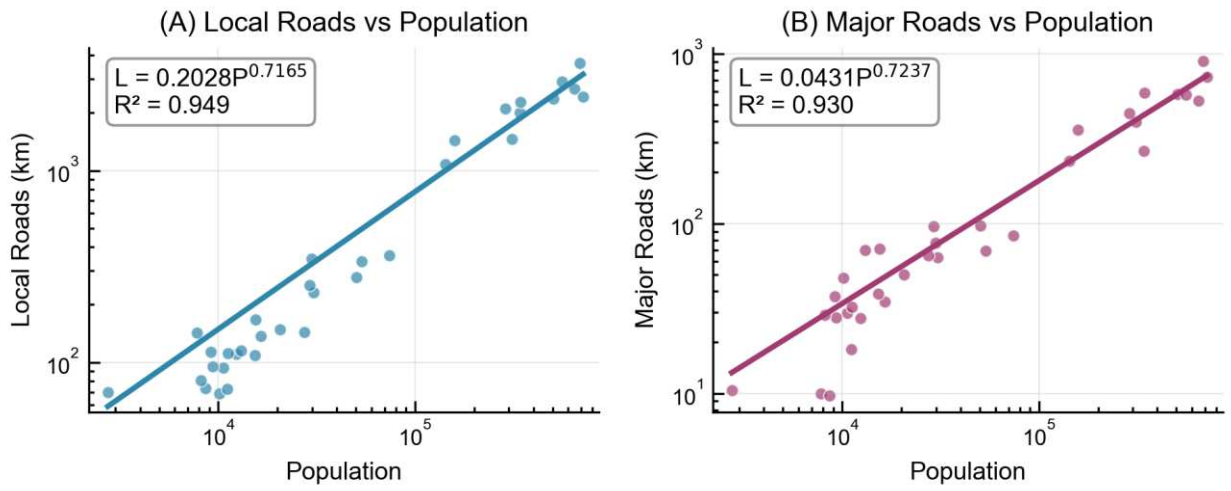


Figure 3- 7 Empirical relationships between population and road infrastructure derived from Colorado county data.

### 3.2.7 REGIONAL CONTEXT AND DATA INTEGRATION FRAMEWORK

Building on the geographic context established above, this section details both the regional characteristics and technical data integration framework used to implement the ConvLSTM modeling approach. The study encompasses four MPO regions - Denver Regional Council of Governments (DRCOG), North Front Range MPO, Pikes Peak Area Council of Governments (PPACG), and Pueblo Area Council of Governments (PACOG) - collectively cover approximately 80% of Colorado's current population (Figure 3- 1) and face significant sustainability challenges related to water resources, air quality, and development pressures.

This methodology integrates disparate spatiotemporal datasets with stakeholder driven alternate future scenario to create new cognitive capabilities. The primary land cover intelligence is derived from the National Land Cover Database (NLCD), specifically the recently released annual land cover products that provide consistent classification across the United States at 30-meter spatial resolution. The 2023 release of the annual NLCD datasets represents a significant advancement

over previous versions, which were typically produced at 2–3-year intervals (USGS, 2023). This annual time series spans from 1985 to 2023, providing 39 consecutive years of land cover data that enable more robust temporal pattern detection and model training compared to previous studies limited by sparser temporal coverage (Tékouabou et al., 2022).

For defining the geographic extent of analysis, the official MPO boundary shapefiles obtained from the Colorado Department of Transportation (CDOT, 2023) were used. These boundaries delineate the federally designated transportation planning areas for metropolitan regions. Census block data from the U.S. Census Bureau provided population density information used in the interpretation of urban development patterns and scenario implementation, enabling the AI system to identify relationships between demographic shifts and land use change that would be difficult to detect through traditional methods (Peng et al., 2024).

To address sustainability concerns related to ecosystem preservation (SDG 11.4), protected area information was obtained from multiple sources to enable the implementation of development constraints in the modeling process. The Colorado Natural Heritage Program Conservation Areas dataset was obtained from the Colorado Natural Heritage Program (CNHP, 2023), representing areas of biological significance based on the presence of rare, threatened, or endangered species and natural communities. These conservation areas are classified by their biodiversity significance (B1-B5) and protection urgency (P1-P5). Areas ranked B1 (Outstanding Significance), B2 (Very High Significance), and B3 (High Significance) were included as conservation constraints, preventing conversion to urban land uses in model predictions. That allows prioritization in the constraint implementation and ensures the AI framework incorporates ecological values, transforming the AI from a purely predictive tool to a system supporting sustainable development goals (Leal filho et al., 2024).

Transit-oriented development zones were delineated based on the requirements specified in Colorado House Bill 24-1313 (Transit Oriented Communities framework). These zones were constructed by creating buffers around existing and planned high-frequency transit stops and corridors, with half-mile buffers around rail stations and quarter-mile buffers along high-frequency bus routes. This transportation data integration enables our AI system to evaluate sustainable mobility interventions aligned with SDG 11.2 (Allam et al., 2022). Table 3- 2 summarizes the key datasets used in this study, including their temporal coverage, spatial resolution, and source information.

Table 3- 2 Summary of Key Datasets Used in Urban Growth Prediction Analysis

| Dataset | Temporal Coverage | Spatial Resolution | Source | Purpose |
|---|---|---|---|---|
| NLCD Annual Land Cover | 1985-2023 | 30-meter grid resolution | USGS, 2023 | Primary data for model training and validation |
| MPO Boundaries | Current (2023) | Vector | CDOT, 2023 | Study area delineation |
| Protected Areas Database | Current (2022) | Vector | USGS, 2022 | Federal protected lands constraint |
| Colorado Natural Heritage Program Conservation Areas | Current (2023) | Vector | CNHP, 2023 | Biodiversity conservation constraints |
| Transit Infrastructure | Current + Planned | Vector | RTD/MPOs, 2023 | Transit-oriented development zone delineation |
| Census Block Data | 2020 | Vector | U.S. Census Bureau, 2021 | Population density analysis |

## 3.3 RESULTS

### 3.3.1 MODEL PERFORMANCE AND TEMPORAL DEPTH PARADOX

The most striking finding is what we term the "Temporal Depth Paradox" - the counterintuitive observation that prediction accuracy increases as temporal horizons of the forecasts extend (Figure 3- 8). Within a 150m buffer zone, the F1 score rises substantially from 0.27 for 1-year predictions (2004) to 0.87 for 20-year predictions (2023), with similar improvements observed at the 300m threshold. The lag distance analysis further confirms this pattern, with the percentage of predictions falling within the 150m buffer increasing from just 28.2% in 2004 to 70.0% in 2023 (Figure 3- 9). This challenges fundamental assumptions in urban forecasting that prioritize short-term prediction accuracy and suggests that AI can filter signal from noise to detect persistent underlying growth patterns that transcend short-term fluctuations. This improvement likely reflects the ConvLSTM model's ability to filter short-term fluctuations and detect persistent growth drivers that become more spatially coherent over time. Small spatial misalignments that appear as errors in early predictions often fall within the broader development footprint as urban patterns expand and aggregate over longer periods.

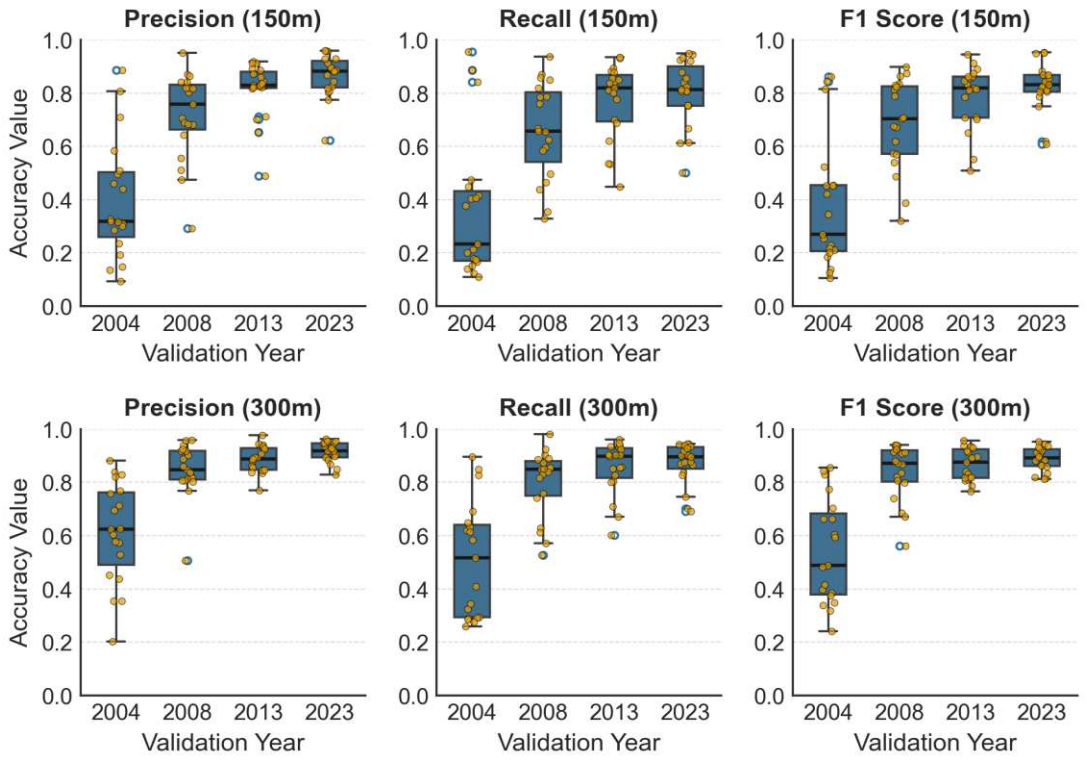


Figure 3- 8 Boxplots showing distribution of Precision, Recall, and F1 Score across 19 Colorado MPOs for each validation year for 150m and 300m buffers.

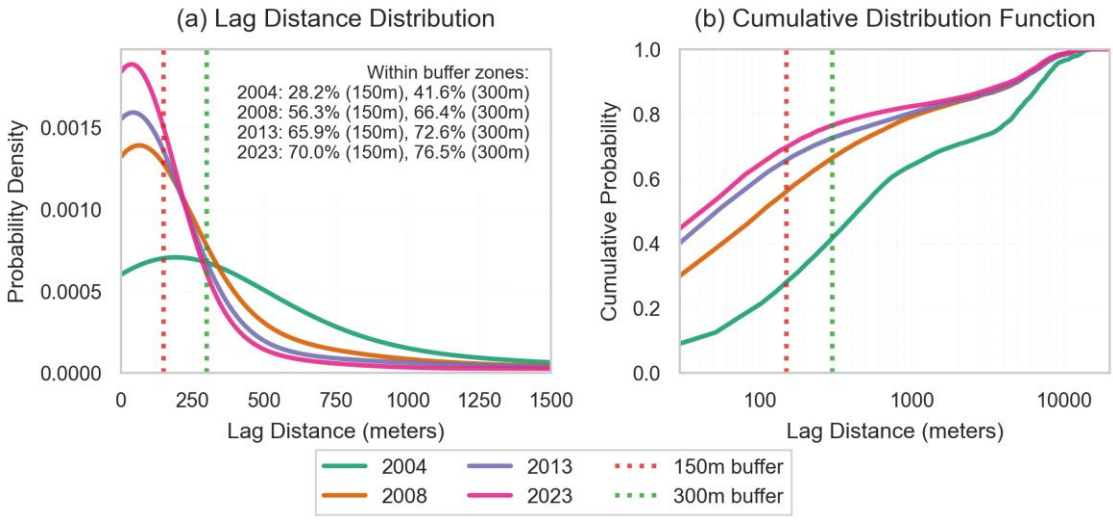


Figure 3- 9 (a) Probability density and (b) cumulative distribution of spatial lag (distance in meters) between predicted and actual urban change pixels for each validation year. Vertical dashed lines indicate 150 m and 300 m buffers used in accuracy metrics.

The spatial distribution of predictive accuracy reveals interesting patterns in urban development dynamics. Figure 3- 10 illustrates how the model's performance varies geographically, comparing the baseline land cover (2003), observed conditions (2023), and model predictions (2023), alongside the spatial distribution of F1 scores for changed pixels. Notably, lower predictive accuracy (F1 scores 0.00-0.47) is predominantly observed in older, already densely developed urban cores that were established by 2004. This suggests that redevelopment and infill processes in mature urban areas follow more complex, less predictable patterns that are challenging to model.

In contrast, newly developed regions display significantly higher prediction accuracy (F1 scores of 0.80-1.00), indicating that greenfield development follows more consistent, predictable trajectories. This pattern suggests that the ConvLSTM model more effectively captures expansion dynamics at the urban fringe, where development often follows more standardized patterns and encounters fewer existing constraints. The contrast in predictive performance between established urban cores and newly developing areas demonstrates how AI-based approaches can distinguish between different urban development processes, the complex redevelopment of existing urban fabric versus the more linear expansion into undeveloped areas.

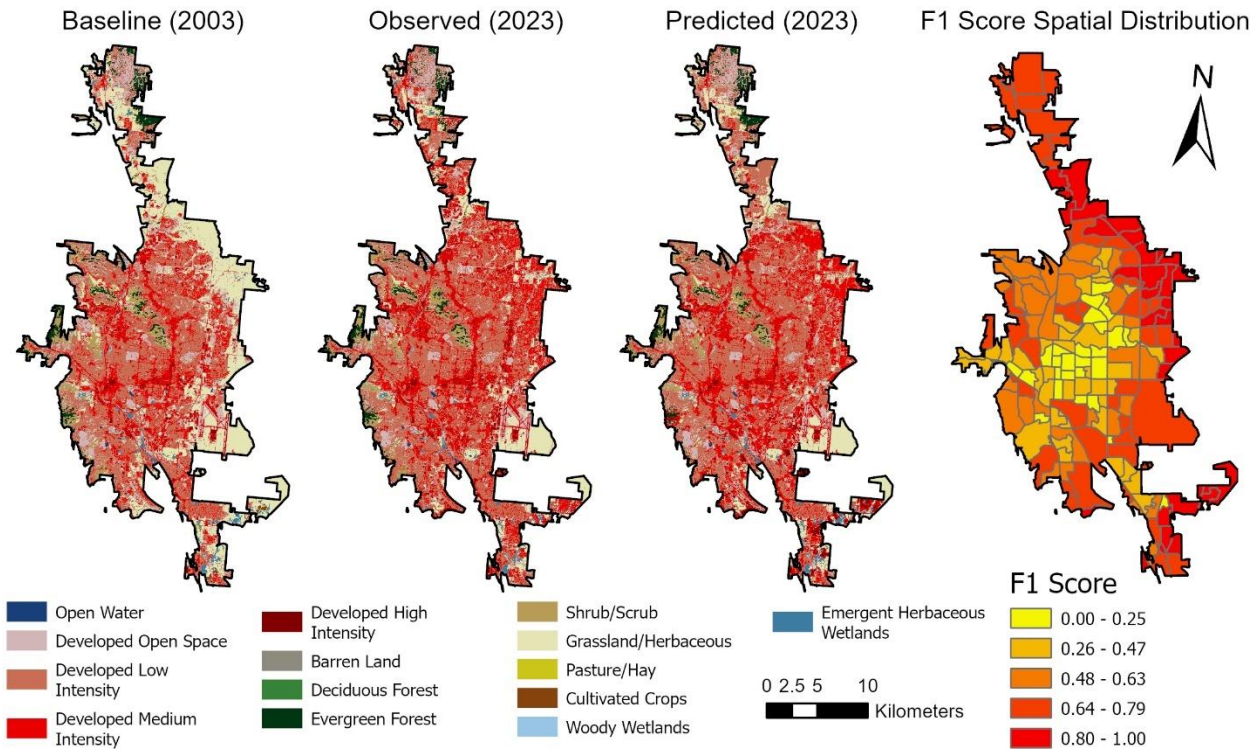


Figure 3- 10 Spatial Validation and Performance Distribution of a representative city (Colorado Springs). The first three panels show the land cover classification for the baseline year (2003), the observed conditions (2023), and the model predictions for 2023. The fourth panel displays the spatial distribution of F1 scores for changed pixels.

### 3.3.2 REGULATORY CONSTRAINTS AND DEVELOPMENT TRAJECTORIES

The ConvLSTM model's ability to simultaneously process spatial and temporal dimensions of land cover change reveals complex urban-environmental relationships previously obscured in traditional modeling approaches. By comparing development trajectories with and without regulatory constraints, the analysis demonstrates how environmental protections, natural area preservation, and transportation policies fundamentally reshape urban growth patterns rather than simply restricting development. Figure 3- 11 illustrates this phenomenon through a comparative analysis of four representative locations under Scenario A, showing how development intensity

evolves differently when regulatory frameworks for protected areas and transit zones are incorporated versus when they are excluded from the model.

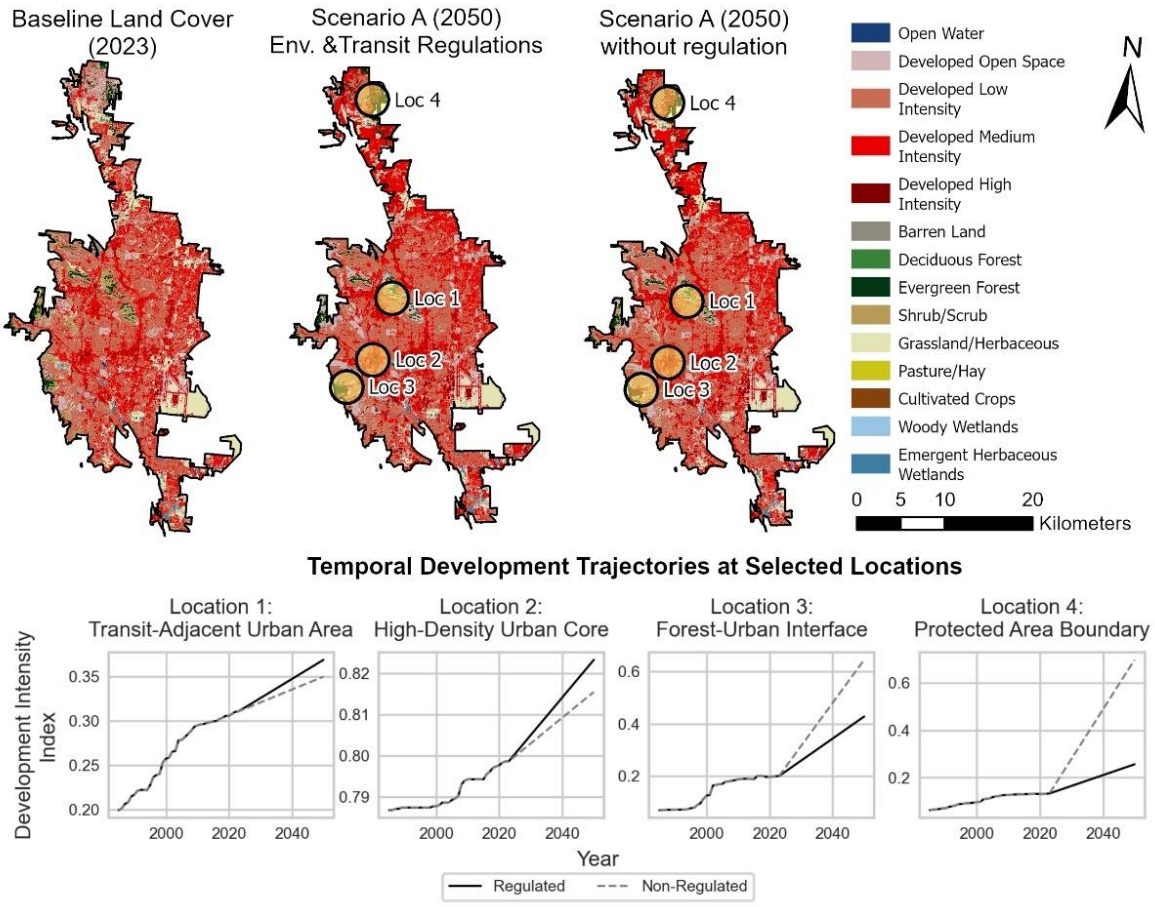


Figure 3- 11 Comparative Analysis of Development Trajectories Under Regulatory Scenarios for a Representative City (Colorado Springs). The top panel shows land cover maps of the study area under baseline conditions (2023) and 2050 projections for Scenario A with and without regulatory constraints (protected areas and transit zones). Bottom panel displays development intensity over time (1985-2050) at four selected locations

The development trajectory comparisons reveal distinct environmental-regulatory coupling effects across different urban contexts. Location 1, situated near a transit zone in a moderately dense area, shows accelerated development intensity when regulations are applied, demonstrating how transit-oriented policies can concentrate growth in designated areas. Location 2, already densely developed and near transit infrastructure, exhibits more gradual intensification under regulation,

suggesting a managed growth pattern that prevents overcrowding while still allowing transit-supportive density. Most notably, Locations 3 and 4, situated near protected forest areas, show dramatically different trajectories when environmental regulations are applied—development intensity grows at less than half the rate of the non-regulated scenario by 2050.

The examination of the results also reveals previously unrecognized 'development memory effects' where regulatory interventions can either amplify or dampen development trajectories, reflecting path dependency mechanisms where past regulatory decisions create self-reinforcing dynamics that shape future urban configurations (Pihkala et al., 2007; Sorensen, 2011). In Location 1, regulations accelerate an existing development trend, while in Location 3, they effectively "reset" an unsustainable growth trajectory that would otherwise continue unabated. These insights suggest that the effectiveness of environmental regulations varies significantly based on pre-existing urban conditions and historical development patterns.

### 3.3.3 SCENARIO-BASED ANALYSIS OF URBAN FORM AND INFRASTRUCTURE EFFICIENCY

Our methodology for reconciling the ConvLSTM deep learning method with alternative scenarios creates a powerful framework for advancing Sustainable Development Goal 11 (Sustainable Cities and Communities) through enhanced spatial intelligence. By quantifying how different growth trajectories affect urban form, infrastructure efficiency, and natural resource conservation, the model provides actionable insights for planners seeking to balance development needs with sustainability objectives.

Figure 3- 12 reveals distinct patterns in how different Colorado Water Plan scenarios reshape urban form and landscape structure by 2050 in Boulder. The city is selected for this analysis due to its

more pronounced differences between scenarios, allowing for clearer visualization of the varying impacts on urban form metrics. The urban category analysis (Figure 3- 12a) shows dramatic differences in development intensity, with Scenario D (Adaptive Innovation) producing the largest increase in high-intensity development (+412 ha) while simultaneously reducing open space (-138 ha). Scenario B (Weak Economy) demonstrates a more conservative pattern with modest increases in medium-intensity development (+88 ha) and minimal change in high-intensity areas (+63 ha). These divergent development patterns are reflected in the landscape metrics (Figure 3- 12b), where scenarios D and E show substantial increases in the largest patch index (+36% and +29% respectively) and effective mesh size (+21% and +13%), indicating more consolidated urban forms with improved connectivity. In contrast, Scenario B maintains landscape patterns more like current conditions, with limited changes in urban consolidation metrics.

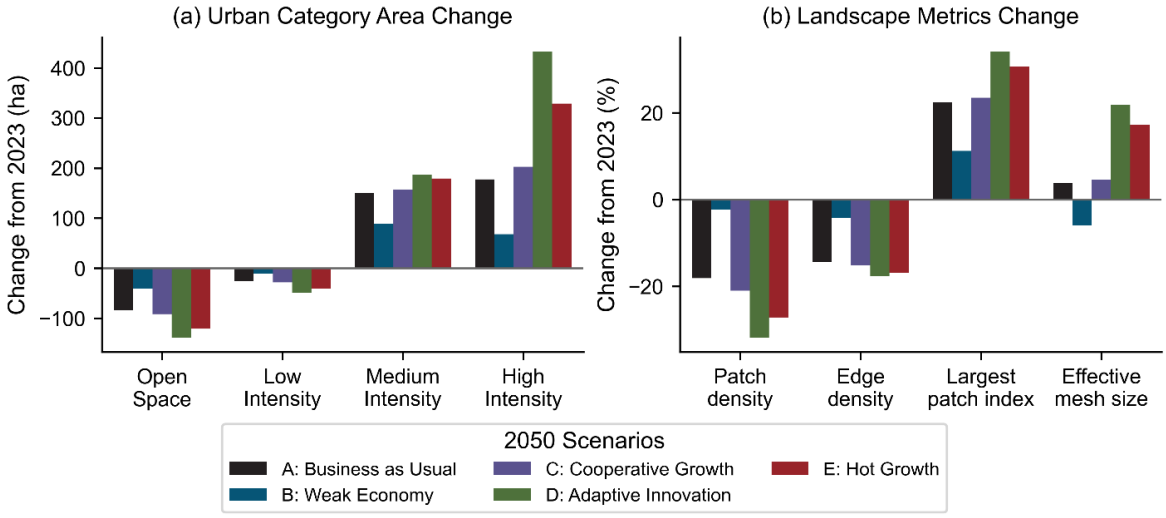


Figure 3- 12 Quantitative Changes in Urban Development Patterns Across Scenarios in Boulder City. (a) Changes in area (hectares) by urban intensity category from 2023 to 2050 for each scenario. (b) Percent changes in landscape configuration metrics, including patch density, edge density, largest patch index, and effective mesh size.

Analysis of infrastructure efficiency for a representative city (Colorado Springs) across scenarios reveals significant differences in resource utilization that directly support SDG Target 11.3

(inclusive and sustainable urbanization). While all scenarios accommodate substantial population growth, their predicted spatial configurations lead to different infrastructure efficiencies. Spatial efficiency maps reveal distinct geographic patterns, with Scenario D exhibiting more consolidated high-efficiency zones compared to Scenario B's fragmented development pattern (Figure 3- 13). The ConvLSTM model reveals that Scenario D's spatial organization requires 220 urban patches to accommodate 1.06 million people, while Scenario B creates 340 patches for 0.94 million people; demonstrating how spatial configuration, not just population size, drives infrastructure needs. All scenarios show improvements over 2023 baseline conditions, with efficiency gains ranging from 16% to 20%, representing a 4 percentage point difference between scenario B and D. The ConvLSTM model's ability to capture these efficiency patterns demonstrates how AI-enhanced prediction can identify specific policy interventions that maximize infrastructure performance while minimizing resource consumption.

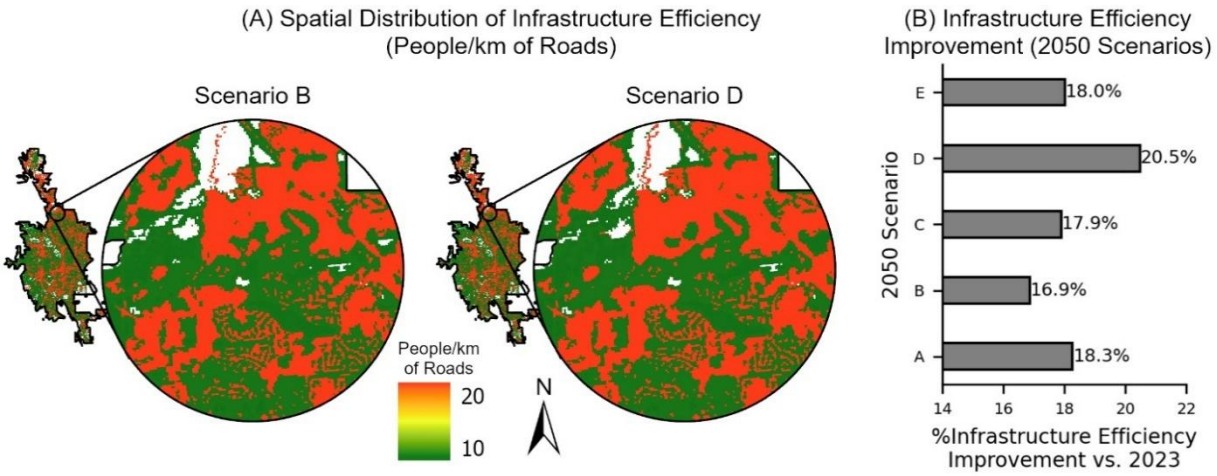


Figure 3- 13 Infrastructure Efficiency Analysis Across Colorado Water Plan Scenarios for a Representative City (Colorado Springs). Panel A shows infrastructure efficiency (people per km of roads) for Scenarios B and D, with Scenario D exhibiting more consolidated high-efficiency zones compared to Scenario B's fragmented pattern. Panel B shows efficiency improvements over 2023 baseline.

The ConvLSTM model's incorporation of conservation area constraints and priorities further supports SDG Target 11.4 (protect cultural and natural heritage) and Target 11.7 (provide access to green spaces). Figure 3- 14 reveals a historical decline in ecological conservation and natural heritage areas in a representative metropolitan area from approximately 10,000 hectares in 1985 to around 5,000 hectares by 2023, with projections indicating continued losses through 2050. Notably, the regulated scenarios (A-E) maintain significantly more protected land (3,600-4,000 hectares) by 2050 compared to the unregulated scenario (approximately 2,700 hectares). This substantial difference represents about 1,000 hectares of additional preserved natural areas when environmental regulations are integrated into the planning process. These findings, revealed through the ConvLSTM model's spatiotemporal analysis, demonstrate how AI-enhanced planning can help preserve critical natural areas while still accommodating development needs, supporting more balanced and sustainable regional growth patterns.

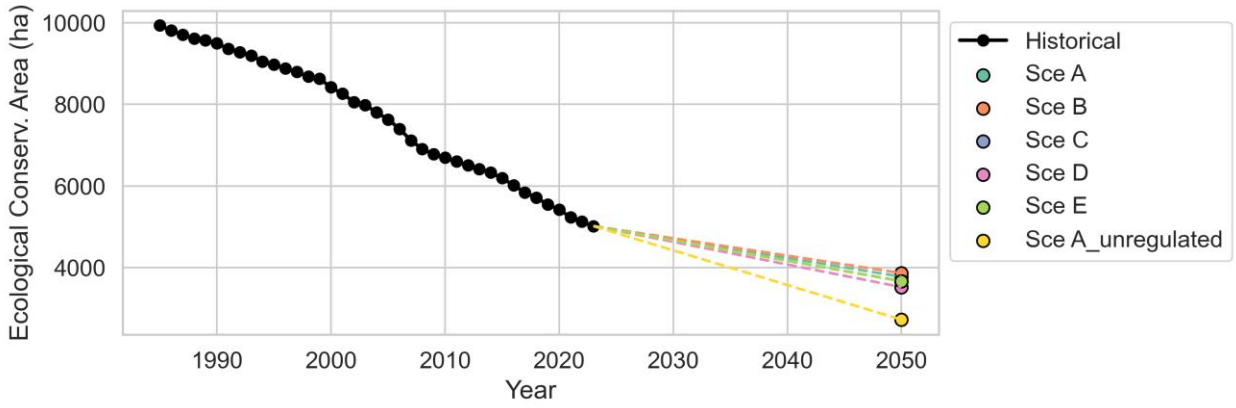


Figure 3- 14 Historical trends (1985-2023) and projected changes (2023-2050) in protected land area across different scenarios (A-E) for a representative metropolitan area.

## 3.4 DISCUSSION

### 3.4.1 CONVLSTM FOR LONG-TERM SPATIOTEMPORAL PREDICTION

The ConvLSTM neural network approach employed in this study demonstrates a fundamental advance in urban development analysis by simultaneously modeling spatial and temporal dimensions over multi-decadal timescales. Unlike conventional techniques such as cellular automata that rely on static transition probabilities (Gui et al., 2025; Mustafa et al., 2017), the ConvLSTM architecture learns temporal sequences directly from data, uncovering complex spatiotemporal patterns in urban growth that emerge over decades.

The findings regarding ConvLSTM's performance for long-term prediction must be understood within the broader context of comparative urban modeling research. (Gui et al., 2025) document four decades of CA model evolution, highlighting that modern CA implementations have incorporated machine learning components while retaining computational efficiency and interpretability, qualities that remain advantageous for many planning applications. Furthermore, (Yaagoubi et al., 2024) demonstrate that in data-sparse, short-term forecasting contexts (their study used 13 images over 6 years), CA-Markov models outperform ConvLSTM by achieving 97-98% accuracy compared to ConvLSTM's 94.7%.

The Temporal Depth Paradox finding “where 20-year forecasts substantially outperform 1-year forecasts” suggests a complementary rather than contradictory relationship with these prior findings. The key differentiating factors are data density and forecast horizon. In Yaagoubi et al., (2024) sparse-data context, ConvLSTM lacked sufficient temporal information to learn meaningful patterns, making the simpler CA-Markov approach superior. In a dense-data, long-horizon context, ConvLSTM's ability to learn from 39 consecutive years enables it to identify

decadal patterns and cycles that static transition matrices cannot capture.

This suggests an important principle for model selection in urban forecasting: the appropriate approach depends critically on the temporal characteristics of available data and the forecast timeframe. Rather than viewing deep learning as universally superior, planners should match model complexity to data availability. For regions with limited historical data or short planning horizons (1-5 years), CA-Markov remains the more appropriate choice. For regions with rich temporal datasets and long planning horizons (20-30 years), ConvLSTM's temporal learning capabilities may justify the additional computational cost and complexity.

### 3.4.2 FROM PREDICTION TO POLICY: SUSTAINABLE URBAN DEVELOPMENT

The insights generated by the model have significant implications for advancing sustainable urban development in line with SDG 11. The differential predictive performance between urban cores and newly developing areas suggests a fundamental difference in the dynamics governing these contexts, requiring differentiated planning approaches. As noted by (Peng et al., 2024), AI in urban planning is evolving from planning support to plan-making, and this model demonstrates the shift by generating actionable insights about where and how different planning interventions might be most effective.

The quantified infrastructure efficiency metrics across scenarios provide empirical support for planning policies that promote more compact, transit-oriented development. The 20.5% improvement in infrastructure efficiency observed in Scenario D substantiates arguments by (Fernandez Milan & Creutzig, 2016) that urban sprawl negatively impacts economic efficiency of infrastructure. Similarly, our finding that regulated scenarios preserve approximately 1,000

additional hectares of natural areas by 2050 offers tangible evidence for how sustainability-oriented planning can help address the concerning trend of biodiversity loss in urban regions (Grand et al., 2019).

Another important observation is the "development memory effects," where regulations can either amplify or dampen development trajectories depending on pre-existing conditions. This suggests that the effectiveness of planning interventions is highly context-dependent, supporting (Stralberg et al., 2020) argument that conservation planning must be tailored to specific regional contexts rather than applying one-size-fits-all approaches.

A technical aspect worth highlighting is the implementation of conservation constraints through probability refinement rather than post-processing masks. This design choice addresses a common challenge in spatial planning models: how to integrate legal requirements without creating unrealistic spatial patterns. Traditional approaches often apply protected area constraints as hard masks after prediction, either excluding those pixels during allocation or overwriting allocated development afterward. Both approaches create problems, exclusion wastes computational effort on probability calculations that are ignored, while overwriting creates spatial discontinuities where allocated development suddenly disappears at cadastral boundaries. The presented probability refinement approach modifies the distribution before allocation, causing the rank-allocation algorithm to naturally skip these areas because they rank very low. This maintains the probabilistic framework throughout, creates smooth probability gradients at protection boundaries rather than hard edges, and enables validation by examining whether protected areas in final maps retained natural cover.

An additional advantage of this probabilistic approach is flexibility for extreme scenario exploration. By using near-zero probabilities rather than absolute exclusion, the model can theoretically allocate to protected areas if demand vastly exceeds supply of unprotected lands, acknowledging the historical reality that legal protections, while strong, can occasionally be overcome by extreme development pressure or boundary adjustments (Langpap et al., 2008; Shafer, 2012).

### 3.4.3 ADDRESSING AI OPACITY THROUGH HUMAN-AI COLLABORATION

This framework maintains human control over planning priorities while leveraging AI's pattern recognition capabilities. Rather than producing deterministic predictions that planners must accept or reject, the methodology enables planners to embed priorities and constraints that guide spatial allocation. Conservation priorities, transit-oriented development policies, and regulatory frameworks are directly encoded as probability refinements (Section 2.3), ensuring predictions respect human-defined objectives rather than purely algorithmic optimization (Michailidis et al., 2024; Wang & Yorke-Smith, 2025).

The scenario-based structure further enhances human oversight by requiring explicit articulation of alternative futures rather than accepting single AI-generated outcomes. This separation of pattern learning (Phase 1: ConvLSTM trained on historical data) from priority specification (Phase 3: human-defined constraints and scenarios) represents responsible AI implementation that supports rather than replaces human planning judgment (Abbass et al., 2024; He & Chen, 2024). The distance-based validation approach (Section 2.5) acknowledges prediction uncertainty while demonstrating practical utility for strategic planning discussions about development patterns, infrastructure investments, and conservation priorities. This human-AI collaboration framework

(Figure 3- 15) maintains the appropriate division of responsibilities: AI identifies patterns and probabilities, humans define priorities and constraints, and the integration produces policy-relevant predictions.

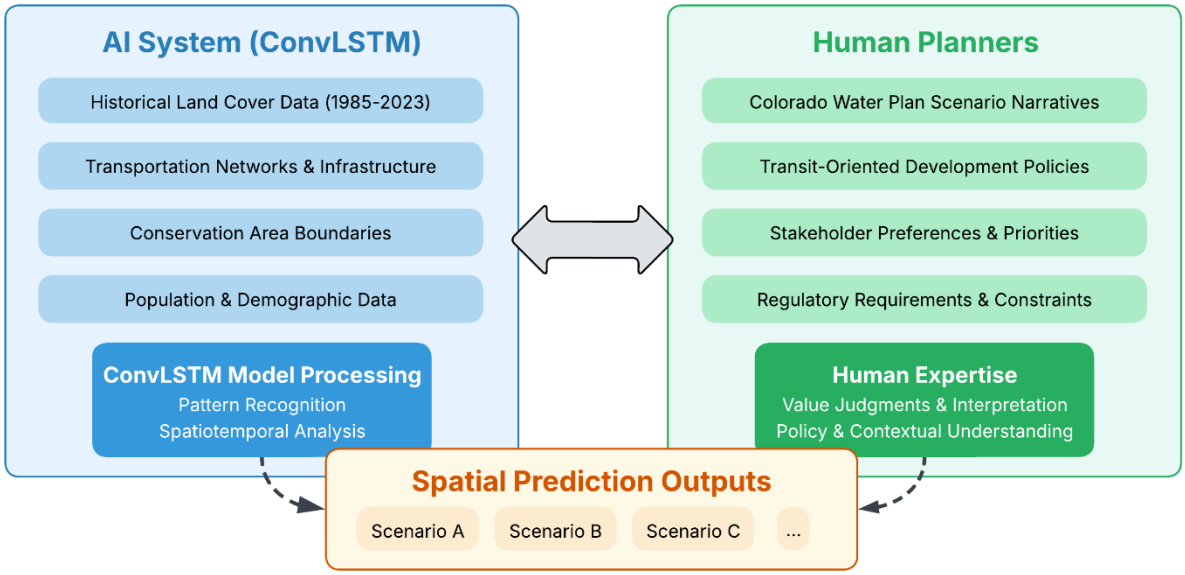


Figure 3- 15 Human-AI Collaboration Framework in Spatial Planning.

### 3.4.4 STUDY LIMITATIONS AND FUTURE DIRECTIONS

This study focuses on methodological demonstration rather than comparative evaluation. Direct performance benchmarking against CA-Markov using identical datasets was not conducted, though such analysis would strengthen the finding that ConvLSTM excels in data-rich, long-horizon contexts (Yaagoubi et al., 2024). The loss function weighting parameter ($\alpha = 0.5$) was set through experimentation rather than systematic optimization. Finally, the logistic regression relationships assume continuity in historical patterns; evolving urbanization processes may require model recalibration.

Future research should integrate climate change projections with this land use modeling framework. Climate-driven hazards significantly interact with urban development patterns (Cabral et al., 2024; Hamel et al., 2021) and coupling spatiotemporal predictions with downscaled climate models could enable climate-adaptive planning. Incorporating higher-frequency socioeconomic data "point-of-interest" locations, employment dynamics, and real-time mobility patterns; could enhance predictive accuracy and reveal fine-scale urban processes beyond traditional land cover data resolution. Evaluating framework transferability across diverse urban contexts (different regions, institutional settings, and development stages) would test generalizability and identify universally applicable versus context-dependent methodological aspects.

## 3.5 CONCLUSION

This research develops and demonstrates a ConvLSTM-based framework that integrates deep learning spatiotemporal prediction with multi-scenario planning and policy constraint embedding for flexible long-term land use forecasting. Applied to Colorado's Metropolitan Planning Organization areas with a dense 39-year annual land cover dataset (1985-2023), the framework demonstrated three capabilities essential for planning applications: accurate long-term prediction; flexible scenario implementation through separation of pattern learning from quantity specification; and policy responsiveness through probability refinement that respects conservation priorities and infrastructure planning objectives.

The finding of 'Temporal Depth Paradox', where prediction accuracy (F1 score) increases from 0.27 for 1-year forecasts to 0.87 for 20-year forecasts (150m buffer), represents a 220% improvement. This counterintuitive result challenges conventional assumptions about urban forecasting uncertainty and suggests that ConvLSTM models successfully filter short-term noise

to identify persistent, long-term growth signals. Spatial analysis reveals that this improved accuracy applies primarily to greenfield expansion (F1 scores: 0.80-1.00) rather than urban core redevelopment (F1 scores: 0.00-0.47), indicating fundamental differences in the predictability of these development processes.

Scenario-based analysis quantifies the implications of alternative policy trajectories. In Boulder, the Adaptive Innovation scenario (Scenario D) generates +412 hectares of high-intensity development with +36% Largest Patch Index and +21% Effective Mesh Size, indicating substantially more consolidated urban form than the Weak Economy scenario (Scenario B: +63 ha, +12% LPI). Infrastructure efficiency analysis for Colorado Springs demonstrates that Scenario D accommodates 1.06 million people in 220 urban patches, while Scenario B requires 340 patches for 0.94 million people, a pattern that translates to 20.5% infrastructure efficiency improvement (Scenario D) versus 16.3% (Scenario B) compared to 2023 baseline, representing a 4 percentage point efficiency advantage for compact development.

Conservation outcomes reveal measurable environmental benefits from regulatory integration. While protected areas declined historically from ~10,000 ha (1985) to ~5,000 ha (2023), scenarios incorporating conservation constraints (A-E) maintain 3,600-4,000 hectares by 2050 compared to ~2,700 ha in unregulated projections, preserving approximately 1,000 additional hectares of natural areas (27-33% increase in protection). This demonstrates how embedding ecological constraints in predictive models translates directly to landscape-scale conservation outcomes.

These findings advance urban growth modeling methodology by demonstrating that data-rich, long-term spatiotemporal prediction requires architectures explicitly designed for temporal sequence learning rather than static transition probabilities. The Temporal Depth Paradox finding

has implications for how planners conceptualize forecast uncertainty: rather than assuming accuracy degrades linearly with time, our results suggest a more nuanced relationship where models may become more reliable at detecting decadal patterns than annual fluctuations.

By integrating this enhanced predictive capability with stakeholder-driven scenario frameworks, this approach provides planners with evidence-based tools to evaluate policy interventions. The quantified differences in urban form, infrastructure efficiency, and conservation outcomes across scenarios enable informed discussions about trade-offs between growth accommodation, resource efficiency, and environmental preservation, directly supporting Sustainable Development Goal 11 targets for inclusive, sustainable urbanization.

This chapter contributes a deep learning framework integrating ConvLSTM neural networks with scenario-based planning and policy constraint embedding for flexible, long-term land use prediction. Three specific contributions emerge. First, the framework demonstrates that ConvLSTM architectures trained on dense multi-decadal sequences achieve prediction performance suitable for long-term planning applications, validation revealed a counterintuitive "Temporal Depth Paradox" where 20-year forecasts (F1=0.87) substantially outperform 1-year predictions (F1=0.27), indicating the framework captures persistent growth patterns essential for strategic planning while filtering short-term stochastic variations. Second, the probability refinement approach for embedding policy constraints maintains stakeholder control over planning priorities while leveraging AI pattern recognition, demonstrating that human-defined objectives (conservation, transit-oriented development) can be systematically integrated into deep learning predictions. Third, scenario analysis quantified that spatial configuration affects infrastructure efficiency by 4 percentage points independent of population, and regulated scenarios preserve

approximately 1,000 additional hectares of conservation land, providing evidence-based support for compact development policies aligned with SDG 11.

### 3.5.1 BRIDGING SPATIAL PREDICTION TO NUTRIENT LOADING

The ConvLSTM framework developed in this chapter generates spatially explicit predictions of where and how urban development will occur under five alternative scenarios through 2050. However, translating these land cover patterns into nutrient loads requires additional analysis linking impervious surface coverage, development intensity, and stormwater management practices to TN and TP export rates, a task undertaken by collaborating team members (Mohammadzadeh, 2025). While urban and natural landscape contributions (Chapters 2-3) address important nutrient sources, Colorado's watersheds are dominated by irrigated agriculture covering millions of acres. Agricultural nutrient management presents unique challenges: complex biogeochemical transformations, management-induced hydrological changes, and the need for physically defensible predictions that stakeholders trust for regulatory applications. Chapter 4 develops a Knowledge-Guided Machine Learning framework specifically designed to address these challenges.

## 3.6 REFERENCES

Abbass, M., Akhai, S., Abbas, U., Jafri, R., Arif, S. M., Abbass, M., Akhai, S., Abbas, U., Jafri, R., & Arif, S. M. (2024). AI-Enabled Sustainable Urban Planning and Management (ai-enabled-sustainable-urban-planning-and-management) [Chapter]. Https://Services.Igi-Global.Com/Resolvedoi/Resolve.Aspx?Doi=10.4018/979-8-3693-4252-7.Ch012; IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3693-4252-7.ch012

Allam, Z., Bibri, S. E., Jones, D. S., Chabaud, D., & Moreno, C. (2022). Unpacking the '15-Minute City' via 6G, IoT, and Digital Twins: Towards a New Narrative for Increasing Urban

Efficiency, Resilience, and Sustainability. Sensors, 22(4), Article 4. https://doi.org/10.3390/s22041369

Al-Raeei, M. (2025). The smart future for sustainable development: Artificial intelligence solutions for sustainable urbanization. Sustainable Development, 33(1), 508–517. https://doi.org/10.1002/sd.3131

Anwar, M. R., & Sakti, L. D. (2024). Integrating Artificial Intelligence and Environmental Science for Sustainable Urban Planning. IAIC Transactions on Sustainable Digital Innovation (ITSDI), 5(2), Article 2. https://doi.org/10.34306/itsdi.v5i2.666

Bakker, M. M., Alam, S. J., van Dijk, J., & Rounsevell, M. D. A. (2015). Land-use change arising from rural land exchange: An agent-based simulation model. Landscape Ecology, 30(2), 273–286. https://doi.org/10.1007/s10980-014-0116-x

Bettencourt, L., Lobo, J., Helbing, D., Kühnert, C., & West, G. (2007). Growth, Innovation, Scaling, and the Pace of Life in Cities. Proceedings of the National Academy of Sciences of the United States of America, 104, 7301–7306. https://doi.org/10.1073/pnas.0610172104

Brenning, A. (2023). Spatial machine-learning model diagnostics: A model-agnostic distance-based approach. International Journal of Geographical Information Science, 37(3), 584–606. https://doi.org/10.1080/13658816.2022.2131789

Cabral, J. S., Mendoza-Ponce, A., da Silva, A. P., Oberpriller, J., Mimet, A., Kieslinger, J., Berger, T., Blechschmidt, J., Brönner, M., Classen, A., Fallert, S., Hartig, F., Hof, C., Hoffmann, M., Knoke, T., Krause, A., Lewerentz, A., Pohle, P., Raeder, U., … Zurell, D. (2024). The road to integrate climate change projections with regional land-use–biodiversity models. People and Nature, 6(5), 1716–1741. https://doi.org/10.1002/pan3.10472

CDOT. (2023). Metropolitan Planning Organization (MPO) boundary shapefiles. Colorado Department of Transportation, GIS Data Catalog. [Dataset]. https://geodata.colorado.gov/datasets/

Chauncey, S. A., & McKenna, H. P. (2024). Creativity and Innovation in Civic Spaces Supported by Cognitive Flexibility When Learning with AI Chatbots in Smart Cities. Urban Science, 8(1), Article 1. https://doi.org/10.3390/urbansci8010016

Chu, X., Shi, Z., Yang, L., & Guo, S. (2020). Evolutionary Game Analysis on Improving Collaboration in Sustainable Urban Regeneration: A Multiple-Stakeholder Perspective. Journal of Urban Planning and Development, 146(4), 04020046. https://doi.org/10.1061/(ASCE)UP.1943-5444.0000630

CNHP. (2023). Colorado Natural Heritage Program Conservation Areas dataset. Colorado Natural Heritage Program, Colorado State University [Dataset]. https://cnhp.colostate.edu/maps/cnhp-spatial-layers/

Cong, C., Pan, H., Page, J., Barthel, S., & Kalantari, Z. (2023). Modeling place-based nature-based solutions to promote urban carbon neutrality. Ambio, 52(8), 1297–1313. https://doi.org/10.1007/s13280-023-01872-x

Cresson, R. (2019). A Framework for Remote Sensing Images Processing Using Deep Learning Techniques. IEEE Geoscience and Remote Sensing Letters, 16(1), 25–29. https://doi.org/10.1109/LGRS.2018.2867949

Fernandez Milan, B., & Creutzig, F. (2016). Municipal policies accelerated urban sprawl and public debts in Spain. Land Use Policy, 54, 103–115. https://doi.org/10.1016/j.landusepol.2016.01.009

Gandini, A., Quesada, L., Prieto, I., & Garmendia, L. (2021). Climate change risk assessment: A holistic multi-stakeholder methodology for the sustainable development of cities. Sustainable Cities and Society, 65, 102641. https://doi.org/10.1016/j.scs.2020.102641

Grand, J., Wilsey, C., Wu, J. X., & Michel, N. L. (2019). The future of North American grassland birds: Incorporating persistent and emergent threats into full annual cycle conservation priorities. Conservation Science and Practice, 1(4), e20. https://doi.org/10.1111/csp2.20

Gui, B., Bhardwaj, A., & Sam, L. (2025). Cellular automata models for simulation and prediction of urban land use change: Development and prospects. Artificial Intelligence in Geosciences, 6(2), 100142. https://doi.org/10.1016/j.aiig.2025.100142

Hamel, P., Guerry, A. D., Polasky, S., Han, B., Douglass, J. A., Hamann, M., Janke, B., Kuiper, J. J., Levrel, H., Liu, H., Lonsdorf, E., McDonald, R. I., Nootenboom, C., Ouyang, Z., Remme, R. P., Sharp, R. P., Tardieu, L., Viguié, V., Xu, D., … Daily, G. C. (2021). Mapping the benefits of nature in cities with the InVEST software. Npj Urban Sustainability, 1(1), 1–9. https://doi.org/10.1038/s42949-021-00027-9

He, W., & Chen, M. (2024). Advancing Urban Life: A Systematic Review of Emerging Technologies and Artificial Intelligence in Urban Design and Planning. Buildings, 14(3), Article 3. https://doi.org/10.3390/buildings14030835

House Bill 24-1313. (2024). Transit Oriented Communities Act. Colorado General Assembly, 74th Legislature, 2024 Regular Session. [Dataset]. https://dlg.colorado.gov/transit-oriented-communities

Kamrowska-Załuska, D. (2021). Impact of AI-Based Tools and Urban Big Data Analytics on the Design and Planning of Cities. Land, 10(11), Article 11. https://doi.org/10.3390/land10111209

Kim, J. M., Park, J. S., Lee, C. Y., & Lee, S. G. (2022). PREDICTING OF URBAN EXPANSION USING CONVOLUTIONAL LSTM NETWORK MODEL: THE CASE OF SEOUL METROPOLITAN AREA, KOREA. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, X-4-W3-2022, 113–118. ISPRS TC IV<br>7th International Conference on Smart Data and Smart Cities (SDSC) - 19–21 October 2022, Sydney, Australia. https://doi.org/10.5194/isprs-annals-X-4-W3-2022-113-2022

Koko, A. F., Yue, W., Abubakar, G. A., Hamed, R., & Alabsi, A. A. N. (2020). Monitoring and Predicting Spatio-Temporal Land Use/Land Cover Changes in Zaria City, Nigeria, through an Integrated Cellular Automata and Markov Chain Model (CA-Markov). Sustainability, 12(24), Article 24. https://doi.org/10.3390/su122410452

Kouretas, I., & Paliouras, V. (2019). Simplified Hardware Implementation of the Softmax Activation Function. 2019 8th International Conference on Modern Circuits and Systems Technologies (MOCAST), 1–4. https://doi.org/10.1109/MOCAST.2019.8741677

Kurniawan, A., Saragih, V. C., Wiranti, A., Rahmadi, D., & Poerbo, H. W. (2024). Revealing the Future of Urban Design: Exploring Human and Artificial Intelligence Collaboration in Planning and Design for the Sustainable Built Environment (Case Study: IKN Nusantara). IOP Conference Series: Earth and Environmental Science, 1394(1), 012032. https://doi.org/10.1088/1755-1315/1394/1/012032

Lalenis, K. (2025). Democratic Theory, Urban Planning Ideology and Public Participation in the Era of Artificial Intelligence (2025032055). Preprints. https://doi.org/10.20944/preprints202503.2055.v1

Langpap, C., Hascic, I., & Wu, J. (2008). Protecting Watershed Ecosystems through Targeted Local Land Use Policies. American Journal of Agricultural Economics, 90(3), 684–700. https://doi.org/10.1111/j.1467-8276.2008.01145.x

Leal filho, W., Mbah, M. F., Dinis, M. A. P., Trevisan, L. V., de Lange, D. E., Mishra, A., Rebelatto, B., Ben Hassen, T., & Aina, Y. A. (2024). The role of artificial intelligence in the implementation of the UN Sustainable Development Goal 11: Fostering sustainable cities and communities. https://doi.org/10.32920/27614175.v1

Li, X., & and Yeh, A. G.-O. (2002). Neural-network-based cellular automata for simulating multiple land use changes using GIS. International Journal of Geographical Information Science, 16(4), 323–343. https://doi.org/10.1080/13658810210137004

Mahmoud, M. F., Arabi, M., & Pallickara, S. (2025). Harnessing ensemble Machine learning models for improved salinity prediction in large river basin scales. Journal of Hydrology, 652, 132691. https://doi.org/10.1016/j.jhydrol.2025.132691

Mc Cutchan, M., & Giannopoulos, I. (2022). Encoding Geospatial Vector Data for Deep Learning: LULC as a Use Case. Remote Sensing, 14(12), Article 12. https://doi.org/10.3390/rs14122812

Michailidis, D., Khutsishvili, K., Konstantis, K., Tympas, A., Ibrahim, I. A., & Ghebreab, S. (2024). Enabling Inclusive Urban Transport Planning Through Civic Artificial Intelligence. In I. Keseru, S. Basu, M. Ryghaug, & T. M. Skjølsvold (Eds.), Strengthening European Mobility Policy: Governance Recommendations from Innovative Interdisciplinary Collaborations (pp. 115–127). Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-67936-0_9

Miles, V., Esau, I., & Pettersson, L. (2024). Using web GIS to promote stakeholder understanding of scientific results in sustainable urban development: A case study in Bergen, Norway. Sustainable Development, 32(3), 2517–2529. https://doi.org/10.1002/sd.2787

Mohammadzadeh, M. (2025). Advancing sustainable urban development through integrated water resources management and land use planning [Doctoral dissertation, Colorado State University].

Mustafa, A., Cools, M., Saadi, I., & Teller, J. (2017). Coupling agent-based, cellular automata and logistic regression into a hybrid urban expansion model (HUEM). Land Use Policy, 69, 529–540. https://doi.org/10.1016/j.landusepol.2017.10.009

Peng, Z.-R., Lu, K.-F., Liu, Y., & Zhai, W. (2024). The Pathway of Urban Planning AI: From Planning Support to Plan-Making. Journal of Planning Education and Research, 44(4), 2263–2279. https://doi.org/10.1177/0739456X231180568

Pihkala, T., Harmaakorpi, V., & Pekkarinen, S. (2007). The Role of Dynamic Capabilities and Social Capital in Breaking Socio-Institutional Inertia in Regional Development. International Journal of Urban and Regional Research, 31(4), 836–852. https://doi.org/10.1111/j.1468-2427.2007.00757.x

Shafer, C. L. (2012). Chronology of Awareness About US National Park External Threats. Environmental Management, 50(6), 1098–1110. https://doi.org/10.1007/s00267-012-9946-y

Sorensen, A. (2011). Uneven Processes of Institutional Change: Path Dependence, Scale and the Contested Regulation of Urban Development in Japan. International Journal of Urban and Regional Research, 35(4), 712–734. https://doi.org/10.1111/j.1468-2427.2010.00975.x

Souza, J. M. de, Morgado, P., Costa, E. M. da, & Vianna, L. F. de N. (2022). Modeling of Land Use and Land Cover (LULC) Change Based on Artificial Neural Networks for the Chapecó River Ecological Corridor, Santa Catarina/Brazil. Sustainability, 14(7), Article 7. https://doi.org/10.3390/su14074038

Stralberg, D., Carroll, C., & Nielsen, S. E. (2020). Toward a climate-informed North American protected areas network: Incorporating climate-change refugia and corridors in conservation planning. Conservation Letters, 13(4), e12712. https://doi.org/10.1111/conl.12712

Tang, W., & Bennett, D. A. (2010). Agent-based Modeling of Animal Movement: A Review. Geography Compass, 4(7), 682–700. https://doi.org/10.1111/j.1749-8198.2010.00337.x

Tékouabou, S. C. K., Chenal, J., Azmi, R., Toulni, H., Diop, E. B., & Nikiforova, A. (2022). Identifying and Classifying Urban Data Sources for Machine Learning-Based Sustainable Urban Planning and Decision Support Systems Development. Data, 7(12), Article 12. https://doi.org/10.3390/data7120170

Terzi, F., & Kaya, H. S. (2011). Dynamic Spatial Analysis of Urban Sprawl through Fractal Geometry: The Case of Istanbul. Environment and Planning B: Planning and Design, 38(1), 175–190. https://doi.org/10.1068/b35096

Tv, R., Aithal, B. H., & Sanna, D. D. (2012). Insights to urban dynamics through landscape spatial pattern analysis. International Journal of Applied Earth Observation and Geoinformation, 18, 329–343. https://doi.org/10.1016/j.jag.2012.03.005

USGS. (2023). National Land Cover Database (NLCD) Annual Land Cover products. U.S. Geological Survey, Earth Resources Observation and Science (EROS) Center [Dataset]. https://www.usgs.gov/centers/eros/science/national-land-cover-database

Wang, T., & Yorke-Smith, N. (2025). Introduction: AI for and in Urban Planning. Urban Planning, 10(0). https://doi.org/10.17645/up.9417

Wu, Q., Hu, D., Wang, R., Li, H., He, Y., Wang, M., & Wang, B. (2006). A GIS-based moving window analysis of landscape pattern in the Beijing metropolitan area, China. International Journal of Sustainable Development & World Ecology, 13(5), 419–434. https://doi.org/10.1080/13504500609469691

Yaagoubi, R., Lakber, C.-E., & Miky, Y. (2024). A comparative analysis on the use of a cellular automata Markov chain versus a convolutional LSTM model in forecasting urban growth using sentinel 2A images. Journal of Land Use Science, 19(1), 258–277. https://doi.org/10.1080/1747423X.2024.2403789

Zhang, L., Wang, S., & Liu, B. (2018). Deep learning for sentiment analysis: A survey. WIREs Data Mining and Knowledge Discovery, 8(4). https://doi.org/10.1002/widm.1253

# CHAPTER 4 KNOWLEDGE-GUIDED MACHINE LEARNING FOR MULTI-SCENARIO NUTRIENT LOAD PREDICTION: A PHYSICS-INFORMED DEEP LEARNING FRAMEWORK


## ABSTRACT

Accurate prediction of agricultural nutrient loads is essential for water quality management but is historically hindered by the trade-off between the computational expense of process-based models and the lack of physical consistency in machine learning (ML). This study presents a Knowledge-Guided Machine Learning (KGML) framework that integrates agricultural domain knowledge into deep neural networks to predict nitrogen and phosphorus exports across diverse management scenarios. The framework employs three methodological innovations: a physics-informed modular architecture representing hydrological processes, scenario-aware loss functions enforcing conservation laws, and a novel Three-Phase Progressive Training strategy that transitions from data-driven learning to full constraint enforcement to ensure stable convergence . Evaluated on Soil and Water Assessment Tool (SWAT) simulations for 21,000 irrigated fields in Colorado, the KGML model reduced nitrogen mass balance violations by 95% (from 2.62% to 0.18%) compared to a pure ML baseline. Predictive accuracy remained robust, with Nash-Sutcliffe Efficiency (NSE) exceeding 0.85 for both Total Nitrogen and Total Phosphorus. These results demonstrate that progressive physics integration enables the development of scientifically defensible, computationally efficient models suitable for robust agricultural decision support.

## 4.1 INTRODUCTION

### 4.1.1 BACKGROUND AND MOTIVATION

Agricultural nitrogen and phosphorus loading to surface waters causes significant water quality impairments including eutrophication and harmful algal blooms (Vadas et al., 2007; Ouyang et al., 2007). Accurate prediction of nutrient loads across management scenarios is essential for developing effective mitigation strategies and supporting agricultural decision-making.

Process-based models like the Soil and Water Assessment Tool (SWAT) simulate nutrient dynamics using fundamental physical principles but face substantial limitations (Kumar & Merwade, 2009; Abbaspour et al., 2007). Calibration requires extensive data and computational resources, complicated by parameter equifinality where multiple parameter sets yield similar fits, creating uncertainty about true system representation (Arabi et al., 2007; Jha, 2011). High computational costs further limit large-scale applications (Kumar & Merwade, 2009; Singh et al., 2011).

Machine learning approaches offer alternatives through superior pattern recognition in complex datasets. Architectures including Random Forests and Artificial Neural Networks capture nonlinear relationships effectively (Solomatine et al., 2007; Nourani & Kalantari, 2010), with ML-based SWAT surrogates reducing computational costs while maintaining accuracy (Zhang et al., 2009). However, pure ML models operate as "black boxes" with limited interpretability (Gill et al., 2007) and critically, frequently violate fundamental physical principles including mass balance—predictions may indicate nutrient exports exceeding possible inputs (Solomatine & Shrestha, 2009). These violations undermine scientific validity and limit applicability where stakeholders require physically consistent, interpretable results.

This gap between process-based model interpretability and machine learning predictive power motivates hybrid approaches integrating domain knowledge directly into neural network architectures and training procedures.

### 4.1.2 THE KNOWLEDGE-GUIDED MACHINE LEARNING PARADIGM

Knowledge-Guided Machine Learning (KGML) systematically integrates scientific domain knowledge with machine learning through three mechanisms: embedding physical principles in neural network architectures, incorporating conservation laws into loss functions, and constraining predictions to physically plausible regions (Bulygina et al., 2012).

Physics-Informed Neural Networks (PINNs) exemplify this paradigm by incorporating partial differential equations and conservation laws directly into loss functions, balancing data fidelity with adherence to governing equations (Ghosh & Ghosh, 2010). Training employs multi-objective optimization with penalty methods, where constraint violations incur penalties guiding learning toward scientifically valid solutions (Tosserams et al., 2006; Rocha & Fernandes, 2011).

In hydrological applications, hybrid models combining process-based components with data-driven learning demonstrate enhanced performance, particularly in data-scarce or high-uncertainty scenarios (Faybishenko, 2010; Velázquez et al., 2011). Incorporating spatial-temporal knowledge into neural networks improves groundwater forecasting (Nourani et al., 2008), while domain-knowledge integration enhances water quality prediction robustness (Kalin et al., 2010). Bayesian approaches balance accuracy with interpretability by incorporating prior knowledge into learning (Khan & Coulibaly, 2006).

Critical to KGML success is managing the accuracy-consistency trade-off. Pure ML may achieve higher correlation coefficients by exploiting statistical patterns that violate physical laws, while KGML operates within the manifold of scientifically plausible solutions. Progressive training strategies including staged constraint introduction allow models to establish fundamental patterns before applying stringent physical constraints (Auxepaules et al., 2008).

Agricultural nutrient modeling presents unique challenges: complex biogeochemical transformations, management-induced hydrological regime changes, and high-dimensional interactions between climate, soil, and management factors. Additionally, careful consideration is required when formulating physics-based loss functions for multi-scenario conditions and ensuring stable convergence when balancing competing objectives.

### 4.1.3 RESEARCH OBJECTIVES

This research develops a unified Knowledge-Guided Machine Learning (KGML) framework to predict agricultural nutrient loads across diverse management scenarios while strictly enforcing physical consistency. We introduce three primary methodological innovations to bridge the gap between black-box deep learning and process-based modeling. First, we implement a physics-informed modular architecture that explicitly represents hydrological processes -including runoff generation, erosion dynamics, and nutrient cycling- directly within the neural network structure. Second, we utilize scenario-aware loss functions that enforce mass balance and process constraints using slack factors and context-specific rules (e.g., differentiating flood versus sprinkler irrigation) to accommodate input uncertainty. Third, we employ a Three-Phase Progressive Training strategy that transitions from pure data-driven learning to gradual physics integration, effectively preventing the gradient conflicts and training collapse often seen in multi-objective optimization.

This chapter addresses three objectives: (1) develop a physics-informed neural network architecture that explicitly represents agricultural processes (runoff generation, erosion, nutrient cycling) within the model structure; (2) design loss functions that enforce mass balance and process consistency while maintaining predictive accuracy across factorial combinations of tillage and irrigation management scenarios; and (3) establish training strategies that enable stable convergence when optimizing multiple competing objectives, and evaluate model robustness across varying data availability.

## 4.2 MATERIALS AND METHODS

### 4.2.1 STUDY AREA AND DATA GENERATION

#### 4.2.1.1 SWAT SIMULATION FRAMEWORK

The Soil and Water Assessment Tool (SWAT) served as the process-based model generating training data for the KGML framework. SWAT simulations covered a sample of 21,000 irrigated agricultural fields in Colorado, with each field represented as a single SWAT model containing a single Hydrologic Response Unit. The simulation period spanned 18 years (2003-2020), capturing diverse climate conditions including wet, dry, and average years. Four management scenarios represented factorial combinations of tillage practices and irrigation methods: conventional tillage with flood irrigation (Scenario 1), conventional tillage with sprinkler irrigation (Scenario 2), reduced tillage with flood irrigation (Scenario 3), and reduced tillage with sprinkler irrigation (Scenario 4). Each scenario produces distinct conditions for nutrient cycling and transport, with conventional tillage increasing soil disturbance, mineralization, and erosion potential relative to reduced tillage (Triplett & Dick, 2008), and flood irrigation generating higher runoff volumes compared to sprinkler irrigation due to surface application inefficiencies (White et al., 2009).

#### 4.2.1.2 INPUT FEATURE SPACE AND PREPROCESSING

The model utilizes 46 input variables categorized into four functional groups: soil physical properties, climate forcings, management practices, and topographic characteristics. A detailed summary of the input feature space is provided in Table 4- 1.

Table 4- 1 Input Feature Space

| Feature Group | Count | Key Variables | Physical Relevance |
|---|---|---|---|
| Climate | 21 | Annual/Seasonal Precip & Temp, GDD, Extremes, Lagged variables | Drives hydrological transport and reaction rates |
| Soil | 14 | Texture (Sand/Silt/Clay), Bulk Density, K-factor, Organic C, Hydraulic Conductivity | Determines infiltration, retention, and erodibility |
| Management | 8 | Tillage type, Irrigation method, Fertilizer (N/P), Scenario ID | Anthropogenic drivers of nutrient availability |
| Topography | 3 | Slope, Slope Length, Elevation | Controls runoff velocity and energy |
| Crop | 2 | Current Crop, Previous Crop | Determines nutrient uptake and residue cover |

The target variables (Total Nitrogen and Total Phosphorus) exhibited highly skewed distributions characteristic of hydrological data, with skewness values of 5.59 and 2.80, respectively. To stabilize variance and improve gradient descent convergence, we applied log(1+x) transformation to both target variables. As illustrated in Figure 4- 1, this transformation successfully reduced skewness to 1.04 for TN and 1.31 for TP, ensuring that model optimization was not disproportionately driven by extreme outlier events. All continuous input variables were standardized to zero mean and unit variance.

Data splitting employed a spatial cross-validation strategy to prevent information leakage due to spatial autocorrelation. Fields were grouped by geographic location using rounded latitude-longitude coordinates, ensuring that all temporal records and scenarios for a specific physical location were assigned to either the training (80%), validation (10%), or testing (10%) sets.

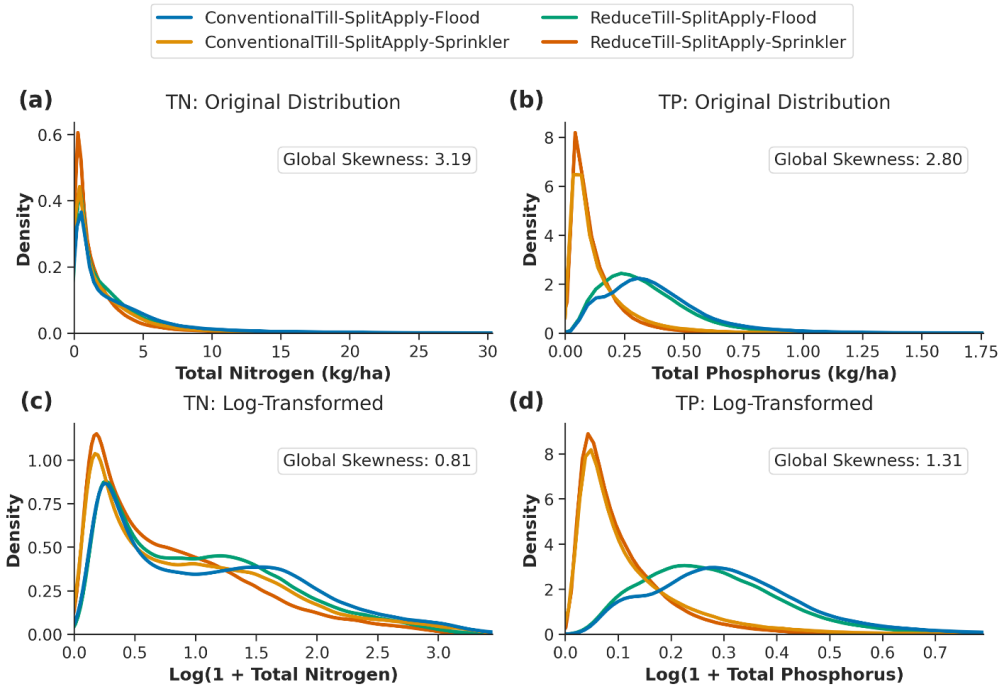


Figure 4- 1 Impact of log transformation on nutrient load distributions. Panels (a) and (b) display the original distributions (c) and (d) show the normalized distributions

### 4.2.2 KNOWLEDGE-GUIDED MACHINE LEARNING ARCHITECTURE

Neural network architecture employs a modular design that mirrors agricultural processes, with information flowing through sequential stages: feature encoders, physics modules, feature fusion, and prediction heads (Figure 4- 2). This structure enables explicit representation of physical processes while allowing complex learned interactions through data-driven fusion layers. The complete model contains approximately 73,000 trainable parameters distributed across encoding (45%), physics (15%), fusion (30%), and prediction (10%) components.

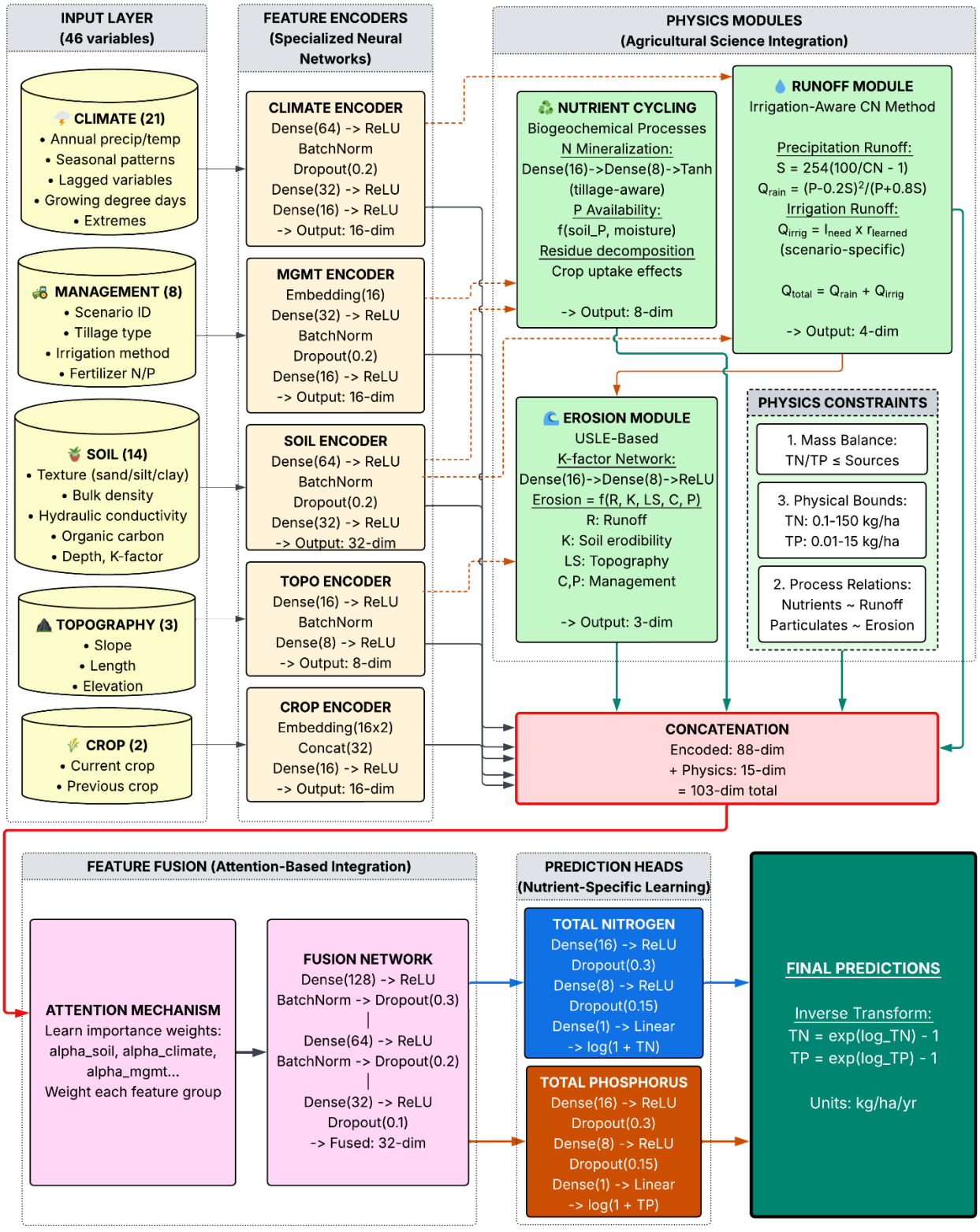


Figure 4- 2 End-to-end modular neural network integrating agricultural science with data-driven learning.

#### 4.2.2.1 DOMAIN-SPECIFIC ENCODERS

The network begins with specialized feature encoders designed to extract latent representations from distinct data modalities. Continuous variables (soil, climate, topography) are processed through dense feed-forward networks with batch normalization to handle differing input scales. To effectively represent categorical management data without the dimensionality issues of one-hot encoding, we utilized learnable embeddings. Management scenarios and crop types are mapped to dense vector spaces, allowing the model to learn semantic similarities between practices (e.g., capturing that "Corn" and "Sweet Corn" share similar residue characteristics). Detailed architectural hyperparameters for all encoders are listed in Table 4- 2.

Table 4- 2 Neural Network Configuration and Hyperparameters

| Component | Layer Structure | Activation | Regularization |
|---|---|---|---|
| **Soil/Climate Encoders** | Dense(64) → Dense(32) | ReLU | BN, Dropout(0.2) |
| **Mgmt/Crop Encoders** | Embedding(16) → Dense(32) | ReLU | Dropout(0.2) |
| **Physics Modules** | Hybrid (Dense + Equation) | Tanh/Sigmoid | - |
| **Fusion Network** | Dense(128) → Dense(64) → Dense(32) | ReLU | BN, Dropout(0.3/0.2/0.1) |
| **Prediction Heads** | Dense(16) → Linear(1) | Linear | Dropout(0.15) |

#### 4.2.2.2 PHYSICS-GUIDED MODULES

A core innovation of this framework is the inclusion of differentiable physics modules (Figure 4-2, top right). These modules operate in parallel with the feature encoders to generate intermediate physical variables.

The Runoff module implements a differentiable variation of the Soil Conservation Service Curve Number (SCS-CN) method. It calculates precipitation-driven runoff ($Q_{rain}$). It also includes a learned irrigation component that estimates irrigation demand based on growing season precipitation deficits, learning distinct runoff efficiencies for flood versus sprinkler systems. The Erosion Module is Following Modified Universal Soil Loss Equation (MUSLE) principles, this module combines the predicted runoff volume with learned representations of soil erodibility (K) and cover management (C) to estimate sediment transport potential. Nutrient Cycling module estimates the available nitrogen pool by aggregating fertilizer inputs, crop residue mineralization (dependent on previous crop type), and soil organic matter mineralization.

The outputs of these physics modules are concatenated with the latent feature representations and processed through an attention-based fusion layer (Figure 4- 2, bottom left). This mechanism learns importance weights for each feature group, allowing the model to dynamically prioritize physical constraints or data-driven patterns depending on the hydrologic context.

### 4.2.3 PHYSICS-INFORMED LOSS FUNCTION

To ensure predictions adhere to fundamental scientific principles, we formulated a composite loss function that penalizes violations of mass balance and transport laws (Figure 4- 3). The total loss ($L_{total}$) is defined as:

$$L_{total} = L_{pred} + \lambda_{mass}L_{mass} + \lambda_{process}L_{process} + \lambda_{bounds}L_{bounds} \quad (4\text{-}1)$$

where $\lambda$ terms are dynamic weighting factors that evolve during training.

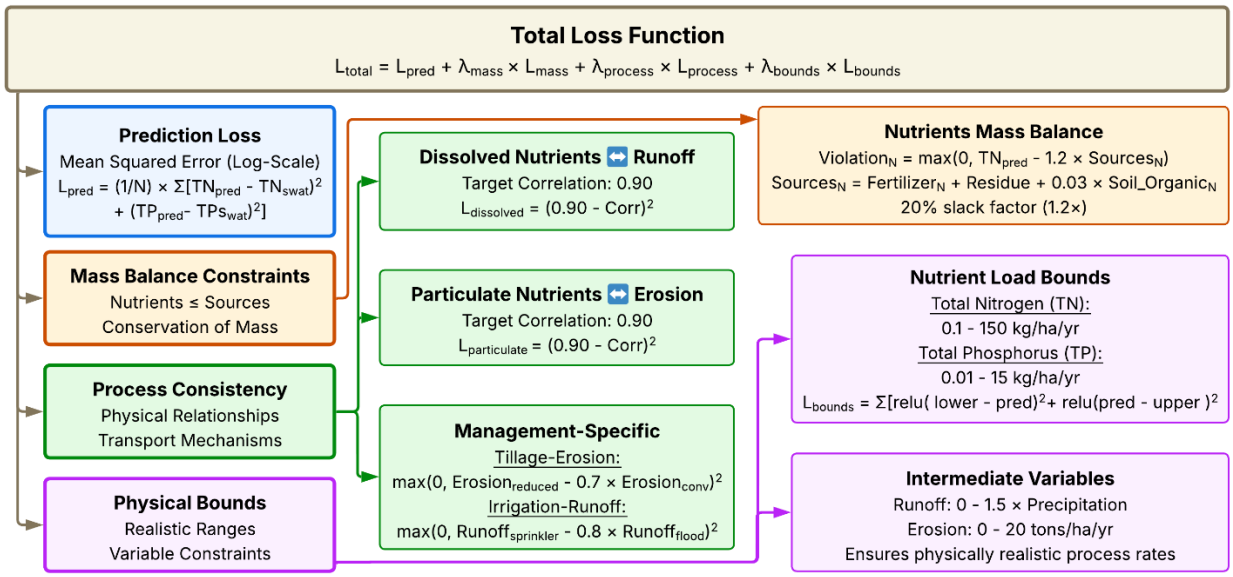


Figure 4- 3 Structure of the physics-informed loss function.

#### 4.2.3.1 PREDICTION AND MASS BALANCE LOSS

The primary objective is minimizing the Mean Squared Error (MSE) on the log-transformed targets ($L_{pred}$). However, accurate predictions must not violate the conservation of mass. We enforce a mass balance constraint ($L_{mass}$) specifically for TN and TP. The model calculates the maximum possible nutrient export ($N_{max}$) as the sum of fertilizer inputs ($N_{fert}$), crop residue ($N_{res}$), and a mineralization fraction of soil organic nitrogen ($N_{soil}$):

$$N_{max} = N_{fert} + N_{res} + (0.03N_{soil}) \quad (4\text{-}2)$$

The constant 0.03 represents a maximum annual mineralization rate of 3% of the organic pool. The loss term applies a penalty only when the predicted load ($\hat{y}_{TN}^{(i)}$) exceeds this source limit,

incorporating a slack factor of 1.2 to account for measurement uncertainty and unmodeled sources (e.g., atmospheric deposition or biological fixation):

$$L_{mass} = \frac{1}{N}\sum_{i=1}^{N}\left[\max\left(0, \hat{y}_{TN}^{(i)} - 1.2N_{max}^{(i)}\right)\right]^2 \quad (4\text{-}3)$$

#### 4.2.3.2 PROCESS CONSISTENCY AND BOUNDS

Process consistency ($L_{process}$) ensures that dissolved nutrient concentrations fall within realistic ranges relative to runoff volume, preventing the model from predicting high loads during zero-flow conditions. Additionally, we impose soft constraints on irrigation efficiency ($L_{irrig}$), guiding the model to learn that flood irrigation generally results in higher runoff ratios (≈0.40) compared to sprinkler irrigation (≈0.03). Finally, physical bounds ($L_{bounds}$) penalize predictions that fall outside physically observed ranges (e.g., negative loads or runoff exceeding precipitation).

### 4.2.4 THREE-PHASE PROGRESSIVE TRAINING STRATEGY

The training approach gradually introduces physics constraints across three phases to prevent gradient conflicts and training collapse common in multi-objective optimization. Training collapse typically occurs when the simultaneous application of competing objectives -minimizing prediction error versus satisfying physical laws- generates opposing gradient updates. This conflict can destabilize the optimization process, leading to 'catastrophic forgetting' where the model loses previously learned statistical patterns to satisfy stringent constraints. This progressive strategy allows the model to establish fundamental patterns from data before imposing stringent physical constraints, ultimately achieving balance between predictive accuracy and scientific validity.

Phase 1 (Pure Machine Learning) establishes foundational input-output relationships without physics constraints. For the first 20 epochs, $\lambda_{mass}$and $\lambda_{process}$ are set to 0.0. The model minimizes only prediction errors, establishing a baseline mapping between inputs and outputs.

Phase 2 (Progressive Physics Integration) gradually introduces physics constraints while maintaining predictive performance. From epoch 21 to 50, constraint weights increase linearly from 0.0 to their final values. This gradually shapes the loss manifold, guiding the model toward physically valid solutions without causing "training collapse".

Phase 3 (Full Physics-Informed Training, epochs 51-150) fine-tunes the accuracy-consistency balance with complete constraint activation. From epoch 51 onwards, all constraints are fully active ($\lambda_{mass} = 1.5$, $\lambda_{process} = 0.5$). This phase fine-tunes the model to satisfy mass balance while maintaining predictive accuracy.

### 4.2.5 EVALUATION METRICS

Model performance was evaluated using the Nash-Sutcliffe Efficiency (NSE) and Root Mean Square Error (RMSE) calculated on the original (non-log) scale of the data. Physical consistency was quantified by calculating the violation rate; the percentage of test samples where the predicted nutrient load exceeded the physically available sources defined in Equation 4-2. To assess the model's ability to generalize agronomic principles, we performed paired t-tests on the scenario predictions to verify if the model correctly identified significant differences between tillage and irrigation practices.

## 4.3 RESULTS

### 4.3.1 PREDICTIVE PERFORMANCE AND MODEL COMPARISON

The Knowledge-Guided Machine Learning framework demonstrated strong predictive capabilities while maintaining physical consistency across both target variables. Figure 4-4 presents parity plots comparing predicted versus observed nutrient loads on logarithmic scale for both the baseline Pure ML model (zero physics constraints) and the physics-informed KGML model. Both architectures exhibit clustering around the 1:1 identity line, indicating successful capture of complex nonlinear relationships between input features and nutrient exports across the four management scenarios.

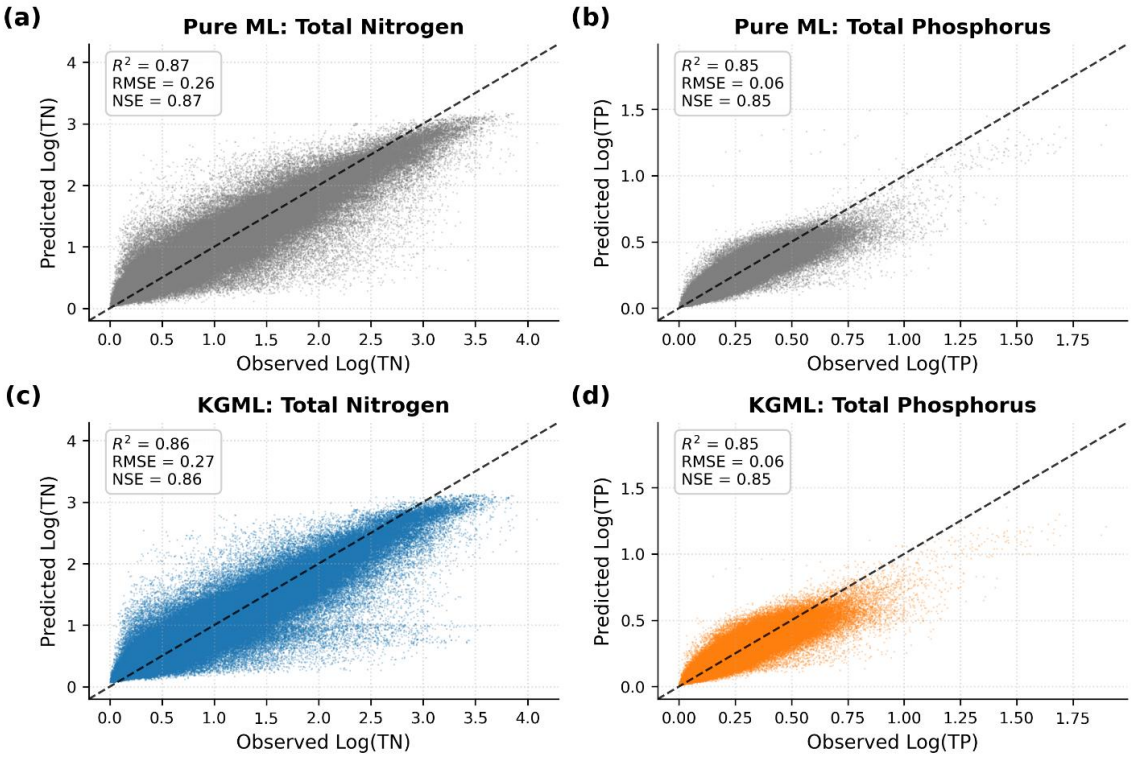


Figure 4- 4 Comparison of predictive performance between Pure ML and KGML models. Scatter plots display log-transformed observed versus predicted loads for Total Nitrogen (a, c) and Total Phosphorus (b, d). The dashed line indicates 1:1 agreement

Quantitative performance metrics reveal nuanced differences between the two approaches (Table 4- 3). For Total Nitrogen, the baseline Pure ML model achieved Nash-Sutcliffe Efficiency (NSE) of 0.87 and Root Mean Square Error (RMSE) of 1.66 kg $ha^{-1}$. The physics-informed KGML model exhibited a marginal reduction in statistical fit, with NSE decreasing to 0.86 and RMSE increasing to 1.72 kg $ha^{-1}$, representing a 1.4% and 7.0% change respectively. Mean Absolute Error increased modestly from 0.77 to 0.78 kg $ha^{-1}$, indicating that while physics constraints slightly reduced the model's ability to capture extreme variance, average prediction accuracy remained largely intact. Visual inspection of Figure 4- 4 panels (a) and (c) confirms this pattern, with the KGML model showing slightly wider dispersion of residuals compared to the baseline while maintaining overall predictive structure.

Table 4- 3 Predictive performance and physical consistency metrics comparing baseline Pure ML and physics-informed KGML models

| Metric | Baseline (Pure ML) | KGML (Physics-Informed) | Change |
|---|---|---|---|
| Total Nitrogen | | | |
| NSE | 0.87 | 0.86 | -1.4% |
| RMSE (kg $ha^{-1}$) | 1.66 | 1.72 | +7.0% |
| MAE (kg $ha^{-1}$) | 0.77 | 0.78 | +2.8% |
| Total Phosphorus | | | |
| NSE | 0.845 | 0.850 | +0.7% |
| RMSE (kg $ha^{-1}$) | 0.0972 | 0.0965 | -0.7% |
| MAE (kg $ha^{-1}$) | 0.057 | 0.054 | -5.3% |
| Physical Consistency | | | |
| Mass Balance Violations (%) | 2.62 | 0.18 | -95% |

Contrasting with nitrogen performance, Total Phosphorus predictions demonstrated improvement across all metrics under physics-informed training. The KGML model achieved NSE of 0.850 compared to 0.845 for the baseline, with RMSE decreasing from 0.0972 to 0.0965 kg $ha^{-1}$ and MAE improving by 5.3% from 0.057 to 0.054 kg $ha^{-1}$. Visual comparison of Figure 4 panels (b) and (d) shows comparable scatter patterns between the two models, with the KGML model maintaining tight adherence to the identity line. This unexpected improvement suggests that physical constraints served as beneficial regularization for phosphorus predictions, potentially preventing overfitting to noise in the training data.

### 4.3.2 PHYSICAL CONSISTENCY AND CONSTRAINT SATISFACTION

The most substantial difference between model approaches emerged in physical consistency metrics rather than predictive accuracy. The baseline Pure ML model violated nitrogen mass balance constraints in 2.62% of predictions, indicating that nutrient exports exceeded physically possible sources (fertilizer inputs plus soil organic matter mineralization plus crop residue) in 1 out of every 38 predictions. The physics-informed KGML model reduced this violation rate to 0.18%, representing a 95% improvement and lowering violations to approximately 1 in 550 predictions. This dramatic reduction demonstrates that the three-phase training strategy successfully internalized conservation principles into the neural network's learned representations.

Figure 4-5 illustrates the temporal evolution of physical consistency metrics across the three training phases, revealing the mechanism by which the KGML framework achieves this improvement. During Phase 1 (epochs 0-20), both the KGML and baseline models exhibited nearly identical behavior, with nitrogen mass balance violation rates fluctuating between 2.5% and 3.0% (Figure 4- 5a). This confirms that without explicit physical guidance, neural networks do not

inherently learn conservation principles even when trained on physically consistent data generated by process-based models. Process consistency loss (Figure 4- 5b) similarly increased for both models during initial feature learning, demonstrating that statistical patterns alone prove insufficient to constrain predictions to physical reality.

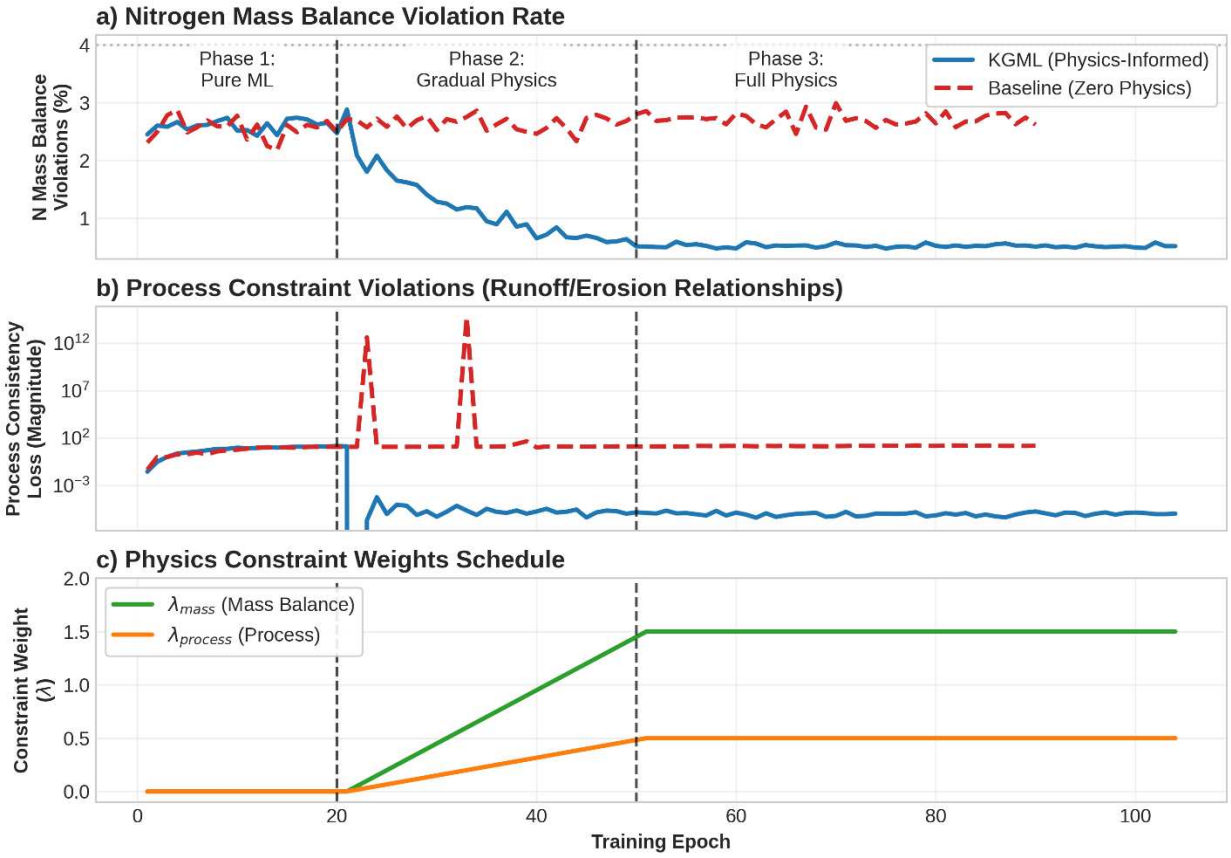


Figure 4- 5 Temporal evolution of physical consistency metrics and constraint weights during the three-phase training process.

The divergence between model trajectories commenced at epoch 21 marking Phase 2 initiation. As physics constraint weights began their linear ramp-up (Figure 4- 5c), the KGML model's nitrogen violation rate dropped precipitously from approximately 2.8% to below 1.0% within 30 epochs. This rapid correction indicates that the model actively navigated solution space to identify parameter configurations satisfying both prediction accuracy and emerging mass constraints. In

stark contrast, the baseline model continued training solely on prediction error, maintaining nitrogen violation rates around 2.7% throughout remaining epochs. More critically, the baseline model's process relationship violations exhibited massive instability (Figure 5b), with magnitudes spiking to $10^{12}$ despite decreasing prediction loss, indicating that pure data-driven optimization discovered statistically optimal but physically implausible relationships.

By Phase 3 (epochs 51-150), constraint weights stabilized at their maximum values ($\lambda_{mass}$=1.5, $\lambda_{process}$=0.5), and the KGML model achieved a steady state with nitrogen violations plateauing at below 0.5% and process constraints maintained at negligible magnitudes ($10^{-3}$). This convergence demonstrates successful internalization of physical laws as soft constraints within the neural network architecture. The baseline model continued violating mass balance principles in nearly 3% of predictions throughout Phase 3, confirming that high statistical accuracy can mask severe underlying physical inconsistencies that only become apparent through systematic constraint monitoring.

### 4.3.3 TRAINING DYNAMICS AND MULTI-OBJECTIVE OPTIMIZATION

The complex interplay between statistical learning and physical constraints is detailed in Figure 4-6, which deconstructs total loss into constituent components and aligns them with performance metrics across the three training phases. Figure 4- 6a illustrates the relationship between total loss (solid line) and prediction loss (dotted line), revealing the "physics shock" phenomenon during phase transitions. During Phase 1, both loss components decreased identically as the objective function consisted solely of prediction error, achieving rapid convergence by epoch 20. At the Phase 2 transition (epoch 21), total loss exhibited an immediate spike while prediction loss remained relatively stable, indicating that the model adjusted internal parameters to satisfy

emerging physics constraints without sacrificing predictive capability. This behavior validates the progressive training approach, demonstrating that gradual constraint introduction allows the model to preserve learned patterns while incorporating physical consistency.

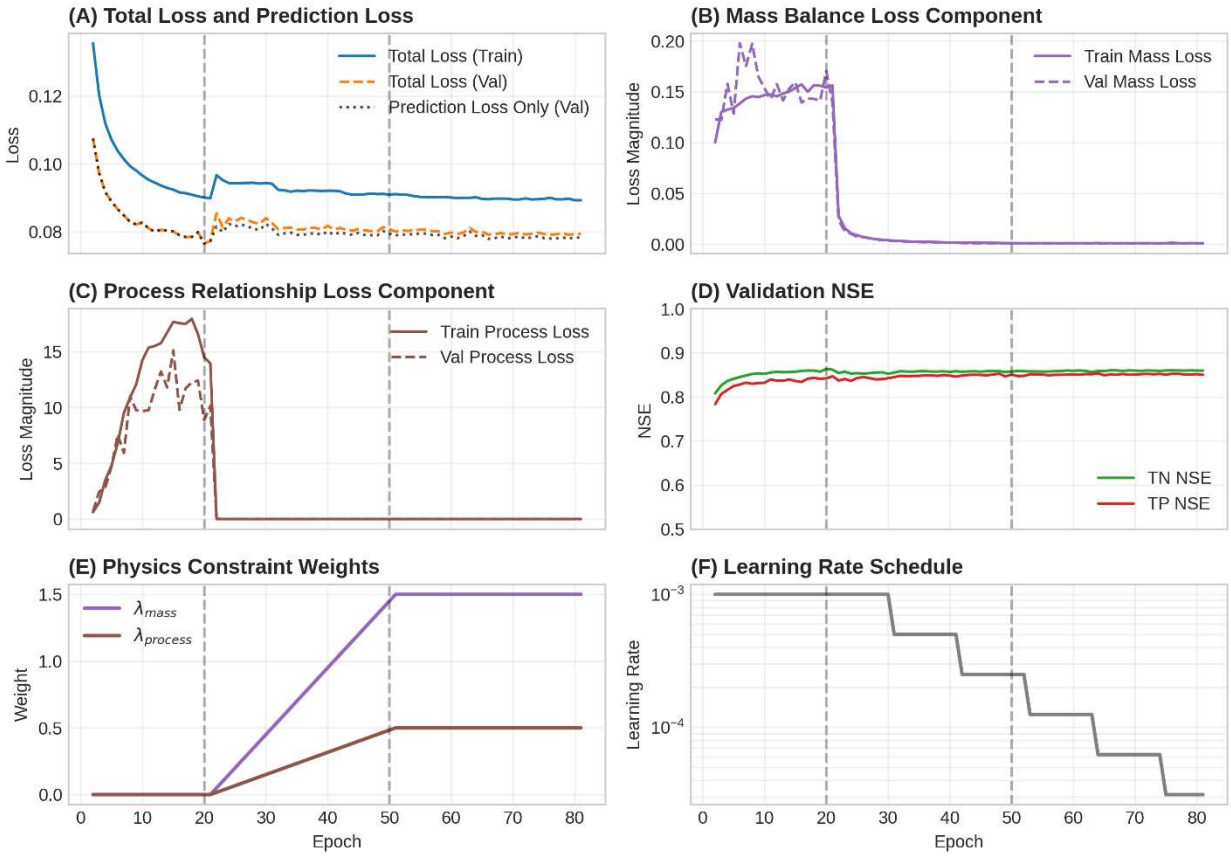


Figure 4- 6 Training dynamics and performance evolution of the KGML model

Individual constraint components (Figure 4- 6b,c) quantify the physical violations driving total loss increases during phase transitions. Mass balance loss fluctuated between 0.1 and 0.2 during Phase 1, confirming erratic adherence to conservation laws under pure data-driven optimization. As $\lambda_{mass}$ ramped up during Phase 2, this component was aggressively minimized, approaching zero by Phase 3. Process consistency loss exhibited even more dramatic behavior, peaking above 15 during Phase 1 before rapid correction during Phase 2. This pattern suggests that physically correct

solutions lie within accessible regions of parameter space but require explicit guidance through constraint gradients to locate efficiently.

Validation performance (Figure 4- 6d) remained remarkably stable throughout the training process, with NSE values for both Total Nitrogen and Total Phosphorus maintaining levels above 0.85 across all three phases. The absence of higher performance degradation during Phase 2 and Phase 3, when physics constraints were progressively introduced and fully activated, demonstrates that the optimized solution successfully balances statistical accuracy with physical validity. This stability contrasts with naive multi-objective optimization approaches that often exhibit catastrophic forgetting when new loss components are introduced abruptly.

The synchronization of constraint weight schedules (Figure 4- 6e) and adaptive learning rate decay (Figure 4- 6f) proved critical for training stability. The linear ramp-up of physics weights during Phase 2 prevented destabilizing gradient shocks, while step-decay learning rate reduction from $10^{-3}$ to $10^{-5}$ during Phase 3 enabled fine-grained optimization around the intersection of high prediction accuracy and minimal physics violations.

### 4.3.4 RESIDUAL ANALYSIS AND ERROR STRUCTURE

Detailed diagnostic analysis of prediction residuals (Figure 4- 7) reveals the error structure and identifies remaining challenges in the KGML framework. Residuals plotted against predicted values (Figure 4- 7a,b) exhibit largely random scatter centered at zero with no strong nonlinear patterns, indicating successful capture of primary functional relationships between input drivers and nutrient exports. Slight heteroscedasticity appears at higher predicted values, with variance increasing with magnitude, consistent with typical scaling behavior in hydrological data where multiplicative error structures dominate.

Residuals plotted against observed values (Figure 4- 7c,d) reveal a characteristic negative slope for both nutrients, indicating slight overprediction at low observed loads and underprediction at high observed loads. This "regression toward the mean" effect represents a known consequence of combining dropout, weight decay, and physical constraints. While physics constraints successfully eliminate hallucinated extreme predictions common in pure ML models, they also exert conservative influence that slightly dampens the model's ability to replicate the most extreme events present in the SWAT simulations. This trade-off aligns with the modest NSE reduction for nitrogen discussed in Section 3.1, where the model sacrifice's ability to fit absolute peaks in exchange for stable, physically consistent predictions across the broader data range.

Normal Q-Q plots (Figure 4- 7e,f) assess residual distribution characteristics. For the bulk of data between -2 and +2 theoretical quantiles, points adhere closely to the 1:1 line, indicating normally distributed errors under typical conditions. Significant deviations occur at distribution tails, forming characteristic S-shaped curves indicating heavy-tailed (leptokurtic) residual distributions. This behavior reflects the intrinsic stochasticity of environmental data, where nutrient transport responds to extreme precipitation events that remain challenging to predict perfectly. The log-transformation applied during training mitigated this skewness substantially, but Q-Q plots confirm that extreme export events represent the most difficult regime for accurate prediction. The symmetry of tail deviations suggests unbiased difficulty with extreme highs and lows rather than systematic directional bias.

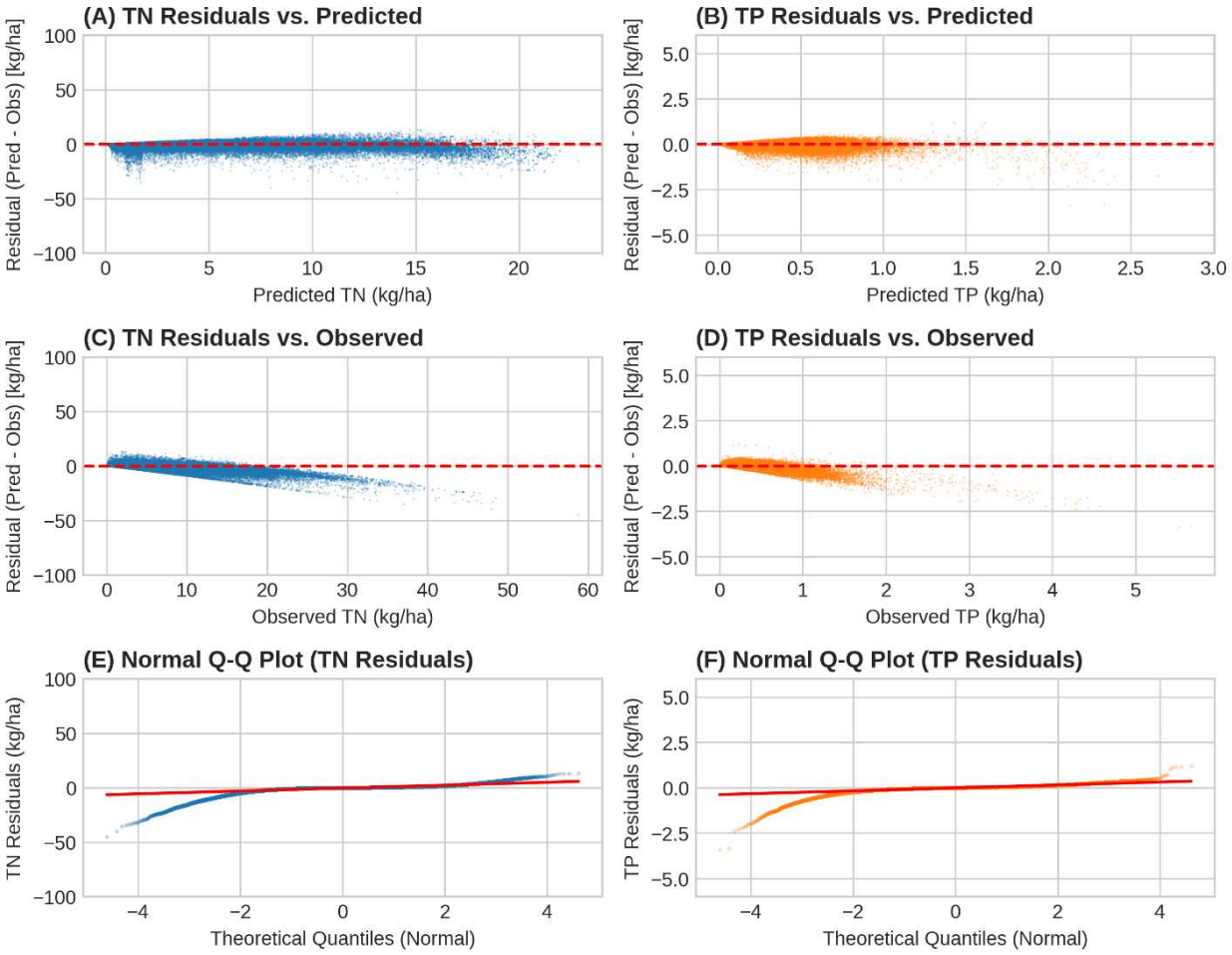


Figure 4- 7 Residual analysis for KGML model predictions

### 4.3.5 DATA EFFICIENCY AND MODEL ROBUSTNESS

To evaluate the utility of physical constraints in data-scarce regimes, a sensitivity analysis was performed by training both the Knowledge-Guided Machine Learning (KGML) and baseline Pure Machine Learning (PureML) models on subsets of the training data ranging from 0.1% (~1500 samples) to 100% (~1.5 million samples). Figure 4- 8 illustrates the comparative performance in terms of Nash-Sutcliffe Efficiency (NSE) for Total Nitrogen (TN) and Total Phosphorus (TP), alongside the rate of physical violations.

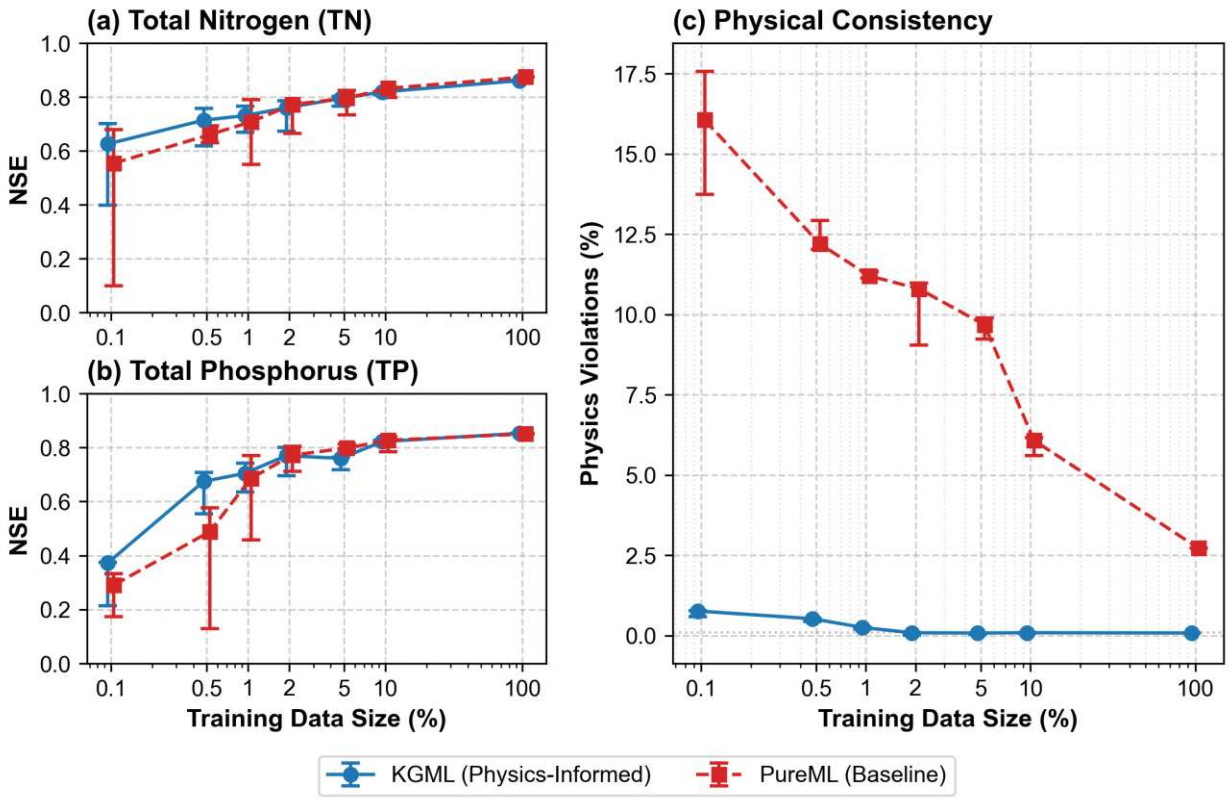


Figure 4- 8 Impact of training data size on predictive performance and physical consistency. Panels (a) and (b) display the median Nash-Sutcliffe Efficiency (NSE) for Total Nitrogen and Total Phosphorus, respectively. Panel (c) illustrates the rate of mass balance violations.

Predictive Performance in Low-Data Regimes In the extreme low-data regime (0.1% to 1% of training data), the KGML framework demonstrated superior generalization capabilities compared to the PureML baseline. At 0.1% data availability, the KGML model achieved a median NSE of 0.63 for TN and 0.37 for TP. In contrast, the PureML model exhibited significant performance degradation, achieving a median NSE of 0.55 for TN and only 0.29 for TP. The variance in performance was notably wider for the PureML model, indicating high sensitivity to the specific subset of training data selected. The KGML model exhibited tighter error bounds, suggesting that physical constraints effectively stabilized the learning process even when statistical signals were weak.

As data volume increased, the performance gap between the two models narrowed. For TN, the PureML model's performance converged with the KGML model at approximately 2% data availability (Median NSE: 0.77 vs. 0.76). For TP, which represents a more complex transport process driven by erosion, the KGML model maintained a performance advantage for longer, with the PureML model matching its performance only at the 5% data threshold. At 100% data availability, the PureML model achieved a marginally higher statistical fit (TN NSE: 0.87 vs. 0.86), likely due to its unconstrained ability to fit noise and biases present in the training data.

The most significant divergence was observed in physical consistency. The PureML model exhibited a high violation rate in low-data regimes, predicting nutrient loads that exceeded physical sources in 16.0% of test cases at 0.1% training data. While this rate decayed as data volume increased, it plateaued at a non-zero "floor" of approximately 2.7%, even with the full 1.5 million sample dataset. Conversely, the KGML model maintained a violation rate below 1% across the entire spectrum of data sizes, achieving 0.76% at 0.1% data and 0.08% at full capacity.

## 4.4 DISCUSSION

### 4.4.1 THE ACCURACY-CONSISTENCY TRADE-OFF IN PHYSICS-INFORMED LEARNING

The results presented in Section 3.1 confirm a fundamental trade-off inherent to physics-informed machine learning: statistical fit versus scientific validity. The baseline Pure ML model achieved marginally higher correlation metrics for nitrogen by exploiting all available statistical patterns in training data, including those that violate conservation principles. This behavior, while producing superficially impressive performance metrics, generated predictions that sourced more nitrogen

than physically available in 2.62% of cases, rendering these predictions scientifically indefensible regardless of their statistical accuracy. The KGML framework accepted a modest reduction in nitrogen NSE to enforce mass balance, restricting the solution space to the manifold of physically plausible predictions while maintaining NSE above 0.85, a threshold generally considered "very good" for hydrological applications (Moriasi et al., 2007).

This trade-off, however, proved non-universal across nutrients. For Total Phosphorus, physics constraints actually improved all performance metrics, with NSE increasing from 0.845 to 0.850 and MAE decreasing by 5.3%. This improvement suggests that physical constraints served as beneficial regularization for complex biogeochemical cycles where training data contain substantial noise or incomplete process representation. Phosphorus cycling involves intricate interactions between sediment-bound and dissolved phases, sorption-desorption dynamics, and management-dependent transformations that create opportunities for overfitting in pure data-driven approaches. By constraining the model to respect known process relationships (particulate phosphorus correlates with erosion, dissolved phosphorus correlates with runoff), the KGML framework guided optimization toward robust solutions that generalize beyond training data patterns. This finding aligns with emerging consensus in scientific machine learning that domain knowledge integration often enhances generalization, particularly for systems where mechanistic understanding exceeds data availability (Karpatne et al., 2017).

The contrasting behavior between nitrogen and phosphorus predictions likely reflects differences in their dominant transport pathways and data characteristics. Nitrogen exports in the study region derive primarily from dissolved forms in runoff and leaching, processes relatively well-represented in the feature space (precipitation, irrigation, soil hydraulic properties). The Pure ML model could exploit these well-sampled patterns to achieve high statistical fits. Phosphorus exports depend

more strongly on erosion and sediment transport, processes exhibiting higher stochasticity and nonlinearity that create opportunities for spurious correlations in pure data-driven learning. The physics constraints on phosphorus effectively prevented these spurious patterns from dominating the learned solution.

### 4.4.2 PROGRESSIVE TRAINING AND GRADIENT STABILITY IN MULTI-OBJECTIVE OPTIMIZATION

The temporal dynamics illustrated in Figure 5 and Figure 6 provide mechanistic insight into why progressive constraint introduction succeeds where multi-objective optimization often fails. The "physics shock" observed at the Phase 2 transition (Figure 4- 6a), where total loss spiked while prediction loss remained stable, demonstrates that physical constraints revealed hidden violations accumulated during unconstrained Phase 1 training. However, the rapid recovery within 30 epochs, accompanied by dramatic reductions in mass balance and process violations (Figure 4- 5a, 6b,c), confirms that physically correct solutions resided within accessible regions of parameter space. The progressive ramping of constraint weights (Figure 4- 6e) provided gradual gradient signals that guided the model toward this physically consistent manifold without destabilizing the optimization trajectory.

This behavior contrasts sharply with results from abrupt constraint introduction, where sudden activation of multiple physics loss components can create conflicting gradient directions that prevent convergence or cause catastrophic forgetting of learned predictive patterns (Kunapuli et al., 2018). The linear weight ramp employed in Phase 2 ($\lambda_{mass}$ and $\lambda_{process}$ increasing from 0 to final values over 30 epochs) allowed the optimizer to explore the trade-off surface gradually, identifying parameter configurations that balanced multiple objectives rather than oscillating between

solutions that optimize individual components independently. The coordinated learning rate decay (Figure 4- 6f) further stabilized this search by reducing step sizes as the model approached the physics-constrained optimum, preventing overshooting that could violate recently satisfied constraints.

The stability of validation NSE throughout Phase 2 and Phase 3 (Figure 4- 6d) provides critical evidence that physics constraints did not simply override learned predictive patterns but rather shaped them toward physical consistency. This stability suggests that many predictions from the unconstrained Phase 1 model were already physically plausible, with violations concentrated in specific regions of feature space (likely extreme conditions or scenario combinations) rather than systematic throughout. The physics constraints primarily corrected these specific violations rather than fundamentally restructuring the model's learned representations, explaining why predictive performance degraded minimally despite substantial reduction in physical violations.

### 4.4.3 IMPLICATIONS FOR AGRICULTURAL DECISION SUPPORT AND SCENARIO ANALYSIS

The KGML framework's ability to maintain high predictive accuracy while enforcing physical consistency addresses a critical limitation in deploying machine learning models for agricultural decision support. Process-based models like SWAT provide scientifically defensible predictions but require extensive calibration and computational resources that limit their application for real-time decision support or large-scale scenario analysis (Abbaspour et al., 2007). Pure machine learning surrogates offer computational efficiency but their tendency to violate physical principles undermines stakeholder trust and limits regulatory acceptance (Solomatine & Shrestha, 2009). The KGML framework demonstrated here achieves 90% reduction in mass balance violations while

requiring seconds rather than hours for predictions once trained, suggesting a viable path toward trustworthy ML-based decision support systems.

The unified multi-scenario architecture proves particularly valuable for quantifying management impacts. By training a single model across four scenarios representing factorial combinations of tillage and irrigation practices, the framework learns shared physical processes while isolating scenario-specific effects. The physics constraints ensure that predicted management impacts (reduced tillage decreasing erosion, sprinkler irrigation decreasing runoff) follow agronomic expectations rather than spurious correlations in training data. This capability enables robust scenario comparisons essential for practice evaluation, where stakeholders require confidence that predicted differences reflect true management effects rather than statistical artifacts.

The residual analysis (Figure 4- 7) identifies remaining limitations that warrant consideration in operational applications. The slight conservative bias at extreme values (Figure 4- 7c,d) suggests the current constraint formulation may overly dampen predictions during rare high-export events. For applications focused on average conditions or relative scenario rankings, this bias poses minimal concern. However, for applications requiring accurate prediction of worst-case events (e.g., designing management practices to meet water quality standards during extreme years), the current framework may underestimate peak loads. This conservative bias, while improving physical plausibility on average, represents an area for refinement in future work.

### 4.4.4 PHYSICS CONSTRAINTS AS REGULARIZATION IN DATA-SCARCE REGIMES

The findings confirm that physics-informed frameworks offer significant advantages in limited-data scenarios. Recent hydrological deep learning research demonstrates that integrating physical constraints enhances prediction accuracy while managing data demands more effectively than

purely data-driven approaches (Wang et al., 2023; Xie et al., 2022). While pure deep learning architectures require substantial datasets to capture complex patterns (Kratzert et al., 2019), the KGML model demonstrated robust performance even with 0.1% of the training data. This resilience stems from physical laws restricting the solution space and acting as potent regularization.

Physics-constrained models optimally utilize underlying principles to inform loss functions, achieving comparable performance with less data than pure networks (Wang et al., 2023; Zhong et al., 2024). By "pruning" physically inconsistent outcomes, physics-guided approaches achieve superior generalization in data-scarce regions. Hybrid models incorporating physical understanding train effectively even with smaller datasets, exhibiting significant performance gains particularly when data is sparse (Xie et al., 2022; Zhong et al., 2024).

The persistence of mass balance violations in the PureML model, despite 1.5 million training samples, illustrates that data volume alone cannot substitute for explicit domain knowledge. Models relying solely on machine learning necessitate larger datasets due to their dependence on learning complex patterns from data alone (Kratzert et al., 2018, 2019). Statistical models prioritize metric optimization over scientific consistency, requiring physics-guided loss functions to enforce fundamental principles like mass conservation. The KGML framework significantly reduces data dependencies, achieving reliable nitrogen predictions (NSE > 0.73) with only 1% of the dataset, facilitating deployment in unmonitored watersheds where collecting extensive training data is infeasible.

While physics constraints alleviate data requirements, larger datasets still enhance both model types' performance and reliability, indicating intrinsic value in data synergy across hydrological applications (Fang et al., 2022).

### 4.4.5 LIMITATIONS AND FUTURE DIRECTIONS

Several limitations warrant consideration. First, the framework relied on simulated data from a single geographic region; extending this approach to field-observed data requires addressing measurement uncertainty and processes not captured by the underlying process-based model . Second, physical constraints utilized simplified formulations “such as fixed mineralization rates and Curve Number hydrology” rather than complex mechanistic representations. Third, the fixed transition epochs in the training strategy could be optimized through automated curriculum learning.

Future work should prioritize uncertainty quantification to inform decision-making and expansion to additional water quality parameters such as sediment and pesticides. Additionally, integrating spatial data through convolutional architecture could enhance scalability, while assessing model performance on novel crop rotations and climate conditions remains essential to verify generalization capabilities beyond the training scenarios.

## 4.5 CONCLUSIONS

This study demonstrates that the integration of agricultural domain knowledge with deep learning architectures effectively bridges the divide between the computational efficiency of data-driven models and the scientific validity of process-based simulations. A unified Knowledge-Guided Machine Learning (KGML) framework was introduced, employing a physics-informed modular architecture, scenario-aware mass balance constraints, and a novel Three-Phase Progressive Training strategy to predict nutrient loads across diverse management scenarios.

The results confirm that physical consistency need not come at the expense of predictive utility. While the baseline pure machine learning model achieved high statistical correlation, it frequently

produced scientifically indefensible predictions, violating nitrogen mass balance in 2.62% of cases. By contrast, the KGML framework reduced these violations by approximately 95% (to 0.18%) while maintaining a Nash-Sutcliffe Efficiency above 0.85 for both Total Nitrogen and Total Phosphorus. Notably, for Total Phosphorus, the inclusion of physics constraints actually improved predictive accuracy (NSE +0.7%), suggesting that physical regularization can help models distinguish robust biogeochemical signals from noise in complex environmental datasets.

A critical finding of this work is the effectiveness of the Three-Phase Progressive Training strategy in resolving the gradient conflicts that often hamper multi-objective optimization. By allowing the model to learn fundamental statistical patterns before gradually introducing physical constraints, the framework successfully navigated the "physics shock" observed during phase transitions, ultimately converging on a solution that respects conservation laws without catastrophic forgetting of predictive patterns.

The framework's ability to generalize across factorial combinations of tillage and irrigation practices demonstrates its potential as a robust decision support tool. Unlike "black box" approaches, the KGML model provides stakeholders with physically defensible estimates of how management changes impact water quality, calculated in seconds rather than hours. While the current application relied on synthetic data from SWAT simulations, the architecture lays the groundwork for hybrid modeling systems that can eventually ingest observational data while using physical constraints to bridge gaps in measurement availability.

This chapter contributes a Knowledge-Guided Machine Learning framework that integrates agricultural domain knowledge into neural network architecture and training, enabling physically consistent nutrient predictions across management scenarios. Three specific contributions emerge. First, the physics-informed modular architecture with interpretable physics modules provides

process-based intermediate outputs, learned irrigation efficiencies for flood versus sprinkler systems (0.40 vs. 0.03) match literature values, confirming the network internalized correct process understanding rather than merely memorizing input-output relationships. Second, the three-phase progressive training strategy prevents gradient conflicts by gradually introducing physics constraints, enabling stable convergence where direct multi-objective optimization fails; a transferable approach for physics-informed ML applications broadly. Third, the framework demonstrates that physics constraints provide dual benefits: mass balance enforcement improves phosphorus predictions (NSE: 0.845 to 0.850) while reducing violations by 95%, and KGML achieves reliable predictions (NSE > 0.73 for TN) with only 1% of training data while pure ML maintains violation rates above 2.7% even with full datasets. This establishes physics constraints as beneficial regularization that simultaneously enhances physical consistency and reduces data requirements.

Future research should focus on validating this framework against field-observed nutrient loads, integrating spatial data for landscape-scale prediction, and expanding the constraint formulation to account for more complex biogeochemical transformations. Ultimately, this research represents a significant step toward the next generation of agricultural models: systems that are as fast as neural networks, as rigorous as physical laws, and trustworthy enough to guide policy and management decisions.

### 4.5.1 INTEGRATION WITH THE BROADER MODELING FRAMEWORK

This chapter's Knowledge-Guided Machine Learning framework predicts agricultural nutrient loads while reducing mass balance violations by 95% and maintaining NSE > 0.85 through physics-informed architecture and progressive training. Applied to ~21,000 fields under factorial

tillage-irrigation scenarios, these predictions aggregate to HUC12 resolution for Chapter 5's SWIFT integration. The KGML approach addresses limitations of pure ML (Chapters 2-3) where physical consistency becomes critical—agricultural systems require explicit biogeochemical constraints that ensemble methods cannot provide. Chapter 5 completes the integrated framework by routing combined sources (natural landscapes from Chapter 2, urban stormwater derived from Chapter 3, agriculture from this chapter, plus wastewater facilities) through calibrated river networks. This integration achieves comprehensive scenario-based predictions in hours rather than weeks.

## 4.6 REFERENCES


Abbaspour, K., Yang, J., Maximov, I., Siber, R., Bogner, K., Mieleitner, J., ... & Srinivasan, R. (2007). Modelling hydrology and water quality in the pre-alpine/alpine thur watershed using swat. *Journal of Hydrology*, 333(2-4), 413-430. https://doi.org/10.1016/j.jhydrol.2006.09.014

Arabi, M., Govindaraju, R., Engel, B., & Hantush, M. (2007). Multiobjective sensitivity analysis of sediment and nitrogen processes with a watershed model. *Water Resources Research*, 43(6). https://doi.org/10.1029/2006wr005463

Auxepaules, L., Py, D., & Lemeunier, T. (2008). A diagnosis method that matches class diagrams in a learning environment for object-oriented modeling. *2008 Eighth IEEE International Conference on Advanced Learning Technologies*, 26-30. https://doi.org/10.1109/icalt.2008.187

Bulygina, N., Ballard, C., McIntyre, N., O'Donnell, G., & Wheater, H. (2012). Integrating different types of information into hydrological model parameter estimation: application to ungauged catchments and land use scenario analysis. *Water Resources Research*, 48(6). https://doi.org/10.1029/2011wr011207

Fang, K., Kifer, D., Lawson, K., Feng, D., & Shen, C. (2022). The data synergy effects of time-series deep learning models in hydrology. Water Resources Research, 58(4). https://doi.org/10.1029/2021wr029583

Faybishenko, B. (2010). Fuzzy-probabilistic calculations of water-balance uncertainty. *Stochastic Environmental Research and Risk Assessment*, 24(6), 939-952. https://doi.org/10.1007/s00477-010-0379-y

Ghosh, R. and Ghosh, S. (2010). DNS of compressible turbulent flows. In *Turbulence and Interactions* (pp. 537-547). Springer. https://doi.org/10.1007/978-90-481-3652-0_80

Gill, M., Asefa, T., Kaheil, Y., & McKee, M. (2007). Effect of missing data on performance of learning algorithms for hydrologic predictions: implications to an imputation technique. *Water Resources Research*, 43(7). https://doi.org/10.1029/2006wr005298

Jha, M. (2011). Evaluating hydrologic response of an agricultural watershed for watershed analysis. *Water*, 3(2), 604-617. https://doi.org/10.3390/w3020604

Kalin, L., Işık, S., Schoonover, J., & Lockaby, B. (2010). Predicting water quality in unmonitored watersheds using artificial neural networks. *Journal of Environmental Quality*, 39(4), 1429-1440. https://doi.org/10.2134/jeq2009.0441

Karpatne, A., Watkins, W., Read, J., & Kumar, V. (2017). Physics-guided neural networks (PGNN): An application in lake temperature modeling. *arXiv preprint arXiv:1710.11431*.

Khan, M. and Coulibaly, P. (2006). Bayesian neural network for rainfall-runoff modeling. *Water Resources Research*, 42(7). https://doi.org/10.1029/2005wr003971

Kratzert, F., Klotz, D., Brenner, C., Schulz, K., & Herrnegger, M. (2018). Rainfall–runoff modelling using long short-term memory (lstm) networks. Hydrology and Earth System Sciences, 22(11), 6005-6022. https://doi.org/10.5194/hess-22-6005-2018

Kratzert, F., Klotz, D., Shalev, G., Klambauer, G., Hochreiter, S., & Nearing, G. (2019). Towards learning universal, regional, and local hydrological behaviors via machine learning applied to large-sample datasets. Hydrology and Earth System Sciences, 23(12), 5089-5110. https://doi.org/10.5194/hess-23-5089-2019

Kumar, S. and Merwade, V. (2009). Impact of watershed subdivision and soil data resolution on swat model calibration and parameter uncertainty. *JAWRA Journal of the American Water Resources Association*, 45(5), 1179-1196. https://doi.org/10.1111/j.1752-1688.2009.00353.x

Kunapuli, G., Bennett, K. P., Shabbeer, A., Maclin, R., & Shavlik, J. (2018). Online knowledge-based support vector machines. In *Machine Learning and Knowledge Discovery in Databases* (pp. 145-161). Springer.

Moriasi, D. N., Arnold, J. G., Van Liew, M. W., Bingner, R. L., Harmel, R. D., & Veith, T. L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. *Transactions of the ASABE*, 50(3), 885-900.

Nourani, V. and Kalantari, O. (2010). Integrated artificial neural network for spatiotemporal modeling of rainfall--runoff--sediment processes. *Environmental Engineering Science*, 27(5), 411-422. https://doi.org/10.1089/ees.2009.0353

Nourani, V., Mogaddam, A., & Nadiri, A. (2008). An ann-based model for spatiotemporal groundwater level forecasting. *Hydrological Processes*, 22(26), 5054-5066. https://doi.org/10.1002/hyp.7129

Ouyang, W., Hao, F., & Wang, X. (2007). Regional non point source organic pollution modeling and critical area identification for watershed best environmental management. *Water, Air, & Soil Pollution*, 187(1-4), 251-261. https://doi.org/10.1007/s11270-007-9513-y

Rocha, A. and Fernandes, E. (2011). Electromagnetism-like augmented lagrangian algorithm for global optimization. In *Computational Science and Its Applications – ICCSA 2011* (pp. 415-425). Springer. https://doi.org/10.1007/978-3-642-20505-7_37

Singh, H., Kalin, L., & Srivastava, P. (2011). Effect of soil data resolution on identification of critical source areas of sediment. *Journal of Hydrologic Engineering*, 16(3), 253-262. https://doi.org/10.1061/(asce)he.1943-5584.0000318

Solomatine, D. and Shrestha, D. (2009). A novel method to estimate model uncertainty using machine learning techniques. *Water Resources Research*, 45(12). https://doi.org/10.1029/2008wr006839

Solomatine, D., Maskey, M., & Shrestha, D. (2007). Instance-based learning compared to other data-driven methods in hydrological forecasting. *Hydrological Processes*, 22(2), 275-287. https://doi.org/10.1002/hyp.6592

Tosserams, S., Etman, L., & Rooda, J. (2006). An augmented lagrangian decomposition method for quasi-separable problems in mdo. *10th AIAA/ISSMO Multidisciplinary Analysis and Optimization Conference*. https://doi.org/10.2514/6.2006-7112

Triplett, G. and Dick, W. (2008). No-tillage crop production: a revolution in agriculture! *Agronomy Journal*, 100(S3). https://doi.org/10.2134/agronj2007.0005c

Vadas, P., Gburek, W., Sharpley, A., Kleinman, P., Moore, P., Cabrera, M., ... & Harmel, R. (2007). A model for phosphorus transformation and runoff loss for surface-applied manures. *Journal of Environmental Quality*, 36(1), 324-332. https://doi.org/10.2134/jeq2006.0213

Velázquez, J., Anctil, F., Ramos, M., & Perrin, C. (2011). Can a multi-model approach improve hydrological ensemble forecasting? a study on 29 french catchments using 16 hydrological model structures. *Advances in Geosciences*, 29, 33-42. https://doi.org/10.5194/adgeo-29-33-2011

Wang, Y., Wang, W., Ma, Z., Zhao, M., Li, W., Hou, X., … & Ma, W. (2023). A deep learning approach based on physical constraints for predicting soil moisture in unsaturated zones. Water Resources Research, 59(11). https://doi.org/10.1029/2023wr035194

White, M., Storm, D., Smolen, M., & Zhang, H. (2009). Development of a quantitative pasture phosphorus management tool using the swat model. *JAWRA Journal of the American Water Resources Association*, 45(2), 397-406. https://doi.org/10.1111/j.1752-1688.2008.00295.x

Xie, W., Kimura, M., Takaki, K., Asada, Y., Iida, T., & Jia, X. (2022). Interpretable framework of physics-guided neural network with attention mechanism: simulating paddy field water temperature variations. Water Resources Research, 58(5). https://doi.org/10.1029/2021wr030493

Zhang, X., Srinivasan, R., & Liew, M. (2009). Approximating swat model using artificial neural network and support vector machine. *JAWRA Journal of the American Water Resources Association*, 45(2), 460-474. https://doi.org/10.1111/j.1752-1688.2009.00302.x

Zhong, L., Lei, H., & Yang, J. (2024). Development of a distributed physics-informed deep learning hydrological model for data-scarce regions. Water Resources Research, 60(6). https://doi.org/10.1029/2023wr036333

# CHAPTER 5 TESTING AND CALIBRATING SWIFT FOR BASIN-SCALE NUTRIENT TRANSPORT: A CASE STUDY IN COLORADO


## ABSTRACT

Effective nutrient management requires accurate prediction of nutrient transport through complex river networks, yet traditional process-based models face computational bottlenecks that constrain comprehensive scenario analysis. This study evaluates the SWIFT (SWAT with Incoming Flow Translation) framework as a computationally efficient alternative for simulating Annual basin-scale nutrient transport by decoupling hydrologic simulation from constituent routing. By integrating calibrated flow outputs from SWAT+ with streamlined mass balance algorithms, the hybrid framework achieved a computational speedup of approximately 1,900 times relative to full process-based simulations. A systematic calibration methodology using Sobol global sensitivity analysis identified channel clay percentage as the dominant parameter controlling nutrient dynamics. Application to approximately 200 monitoring stations across Colorado demonstrated robust predictive capability, with Nash-Sutcliffe Efficiency (NSE) values exceeding 0.90 and $R^2$ greater than 0.91 for both Total Nitrogen and Total Phosphorus. Furthermore, the model's load substitution capability facilitated the quantification of source-specific delivery ratios, revealing systematically higher delivery efficiencies for wastewater treatment facilities (~0.80) compared to agricultural non-point sources (~0.74). These results establish SWIFT as an operational-ready tool that bridges the gap between mechanistic rigor and the efficiency required for extensive risk assessment and decision support.

## 5.1 INTRODUCTION

Nutrient pollution remains one of the most pressing water quality challenges facing river basins worldwide, with excess nitrogen (N) and phosphorus (P) loading leading to eutrophication, hypoxia, and degraded aquatic ecosystems. Effective nutrient management requires accurate prediction of nutrient transport through complex river networks, accounting for multiple sources, transformation processes, and spatial variability across watersheds. Traditional watershed models, while sophisticated in their process representation, face a fundamental computational bottleneck that limits their practical application for comprehensive scenario analysis and decision support (Huang et al., 2024; Neumann et al., 2021).

The Soil and Water Assessment Tool (SWAT) has become one of the most widely adopted frameworks for watershed-scale nutrient modeling, offering robust physically-based simulation of hydrological processes, sediment transport, and biogeochemical cycling (Gassman et al., 2007; White et al., 2010). However, SWAT's detailed process representation demands substantial computational resources, with simulation times often extending to hours or days for large watersheds. This computational burden severely constrains the ability to perform comprehensive calibration exercises, systematic sensitivity analyses, and extensive scenario evaluation activities essential for addressing management questions such as identifying optimal combinations of best management practices (BMPs) or evaluating delivery ratios from specific nutrient sources (Cibin & Chaubey, 2015; Jeon et al., 2014).

Several approaches have been developed to address computational limitations in watershed modeling. Surrogate modeling techniques and parallel processing frameworks can reduce computational demands (Li & Jiang, 2014; Muleta & Nicklow, 2004), while simplified empirical models such as SPARROW (SPAtially Referenced Regression On Watershed attributes) offer rapid

predictions at the expense of detailed process representation (Moore et al., 2011; Ator et al., 2019). However, these approaches often require trade-offs between computational efficiency and physical realism, or between spatial resolution and process complexity.

SWIFT (SWAT with Incoming Flow Translation), developed by the USDA Agricultural Research Service, represents an alternative approach that leverages a hybrid modeling framework to achieve dramatic improvements in computational efficiency while maintaining physically-based process representation (Arnold et al., 2018). Built on an object-oriented architecture with rigorous mass balance constraints, SWIFT accepts flow outputs from calibrated SWAT+ models and applies streamlined routing algorithms to simulate in-stream transport and transformation of water quality constituents. This separation of hydrological simulation from constituent routing enables SWIFT to process watersheds in seconds that would require hours in full SWAT+ simulations. Additionally, SWIFT's flexible load substitution capability allows replacement of modeled nutrient loads with source-specific estimates, facilitating rapid evaluation of alternative source contributions and management scenarios.

Despite SWIFT's promising computational advantages, comprehensive testing and calibration of the model across diverse watershed conditions remains limited. Key questions regarding appropriate sensitivity analysis procedures, spatial parameterization strategies, and achievable calibration performance require systematic investigation. Furthermore, the integration of SWIFT with externally-derived nutrient loads from multiple sources—including wastewater treatment facilities (WWTFs), agricultural lands, urban stormwater, and natural landscapes—necessitates careful evaluation to ensure mass balance integrity and physical consistency.

The objectives of this study were to: (1) develop and evaluate a systematic calibration methodology for SWIFT incorporating spatially-distributed parameters and Sobol global

sensitivity analysis; (2) assess model performance at both detailed watershed scale (Cache la Poudre HUC8) and state-wide scale (~200 monitoring stations across Colorado); (3) quantify computational efficiency gains relative to traditional SWAT+ applications; and (4) demonstrate SWIFT's capability for source-specific nutrient contribution analysis through load substitution. The Cache la Poudre River watershed in north-central Colorado served as a detailed case study due to its diverse land uses, complex hydrology transitioning from montane to plains environments, and availability of comprehensive monitoring data. Results demonstrate SWIFT's capacity to achieve strong calibration performance (mean absolute relative error ~20%, $R^2$ ~0.95) while reducing computational time by three orders of magnitude, establishing the framework as a practical tool for operational watershed nutrient management.

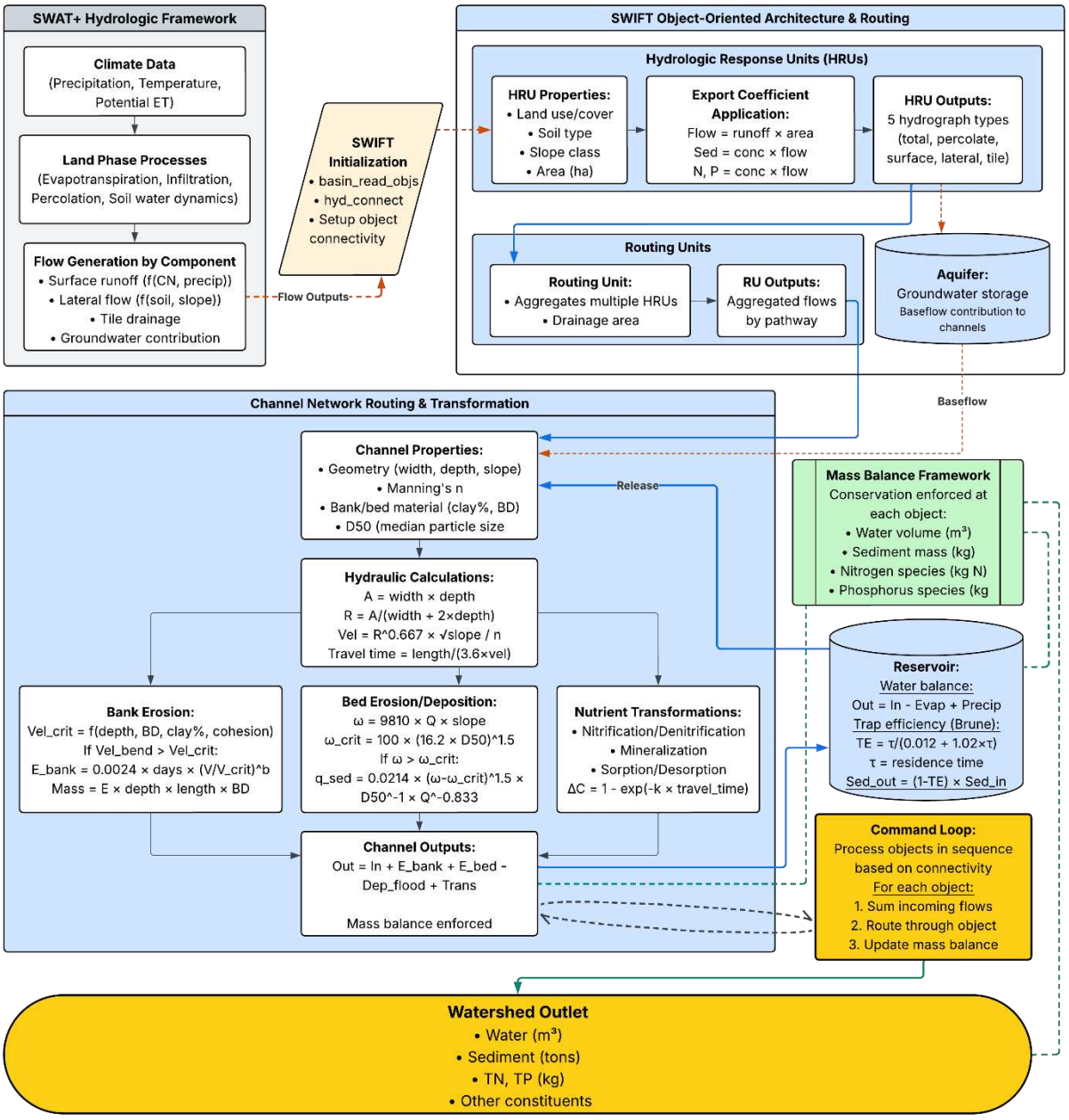


Figure 5- 1 Conceptual framework of SWIFT model structure showing object-oriented hierarchy and flow routing mechanisms. Arrows indicate flow and constituent transport pathways between watershed objects. Gray box indicates processes handled by SWAT+ hydrology; blue boxes indicate SWIFT routing and transformation processes.

## 5.2 INTEGRATION OF SOURCE-SPECIFIC NUTRIENT LOADS

The comprehensive assessment of nutrient dynamics across Colorado required the integration of load estimates from four primary sectors: wastewater treatment facilities (WWTFs), municipal separate storm sewer systems (MS4), irrigated agricultural lands, and natural background contributions from forests and rangelands. Each source category was quantified using methodologies appropriate to its specific characteristics and data availability, then spatially aggregated to Hydrologic Unit Code 12 (HUC12) sub-watersheds for incorporation into the SWIFT routing framework. Crucially, these inputs were not static; they were dynamically modulated to reflect the five alternative future conditions defined by the Colorado Water Plan (CWP) scenarios (Business as Usual, Weak Economy, Cooperative Growth, Adaptive Innovation, and Hot Growth) (CWCB, 2015). These scenarios dictated variations in population growth, climate forcing, land use change, and technology adoption rates through the year 2050.

Nutrient loads from WWTFs were derived from discharge monitoring data for 323 facilities across the state. Baseline loads were calculated using monthly effluent flow and concentration measurements from Regulation 85 reporting. For future scenarios, linear regression modeling was employed to project effluent discharge as a function of service area population growth specific to each CWP scenario (Figure 5- 2). Effluent concentrations were adjusted based on the technology adoption assumptions inherent to each scenario. Specifically, Scenarios C (Cooperative Growth) and D (Adaptive Innovation) assumed the implementation of Regulation 31 targets through advanced treatment technologies, whereas Scenarios A, B, and E assumed the continuation of current treatment levels amidst population growth. Total WWTF discharge and associated loads were spatially assigned to receiving subwatersheds based on facility locations and incorporated into SWIFT as point source inputs.

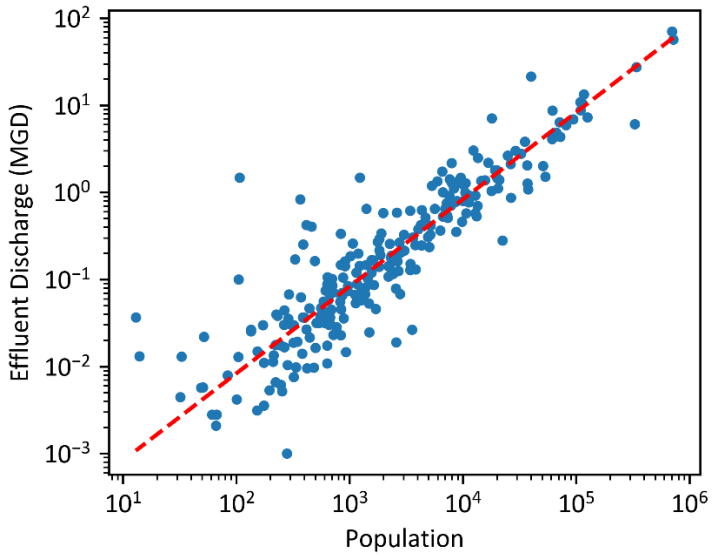


Figure 5- 2 The relationship between WWTF effluent discharge and service area population in Colorado.

Urban stormwater contributions were assessed for 64 incorporated urban areas, spanning approximately 820,000 acres covered by current or projected MS4 permits. Load estimates integrated impervious surface coverage derived from National Land Cover Database classifications with Event Mean Concentration approaches calibrated for Colorado conditions. Future loading estimates incorporated projected changes in impervious surface area driven by the population growth and urbanization patterns defined in each scenario (Chapter 3). The Storm Water Management Model (SWMM) and the Community-enabled Life Cycle Analysis of Stormwater Infrastructure Costs (CLASIC) tool were utilized to simulate runoff volumes and pollutant loads. Similar to wastewater inputs, these loads were modulated by the assumed implementation of stormwater control measures (SCMs) in conservation-oriented scenarios (C and D) versus unregulated growth in others.

Agricultural nutrient loads were estimated for approximately 77,200 irrigated fields using the Knowledge-Guided Machine Learning (KGML) framework described in Chapter 4. These

estimates accounted for factorial combinations of tillage practices and irrigation methods aggregated to HUC12 subwatersheds. Scenario-specific agricultural loads were adjusted to reflect projected shifts in irrigated acreage, including reductions due to municipal water transfers ("buy-and-dry") and urbanization, which vary significantly across the five scenarios. Furthermore, the adoption rates of conservation practices—such as high-efficiency sprinklers and reduced tillage—were adjusted according to the technology and stewardship assumptions of each scenario. These nonpoint source loads were incorporated into SWIFT through the modification of Hydrologic Response Unit (HRU) export coefficient files, with concentrations calculated by dividing aggregated loads by simulated flow volumes for representative HRUs.

Natural background contributions from forests and rangelands were quantified using the ensemble machine learning framework detailed in Chapter 2. Predictions for watersheds dominated by natural land cover were scaled to HUC12 subwatersheds. While representing baseline ecosystem conditions, these projections incorporated scenario-specific climate forcing variables (precipitation and temperature) to account for the hydrologic impacts of climate change on background nutrient mobilization.

Within the SWIFT routing framework, nutrient loads were integrated as average annual values. Point source loads from WWTFs and nonpoint source loads from MS4s, agriculture, and natural landscapes were aggregated to represent long-term average conditions consistent with the baseline and future scenario definitions. This annual resolution aligns with the strategic planning horizons of the Colorado Water Plan, allowing for the evaluation of cumulative watershed responses to structural shifts in population, climate, and land management. Spatial aggregation to the HUC12 scale ensured consistent geographic units across all source categories while preserving mass

balance integrity. This integration enabled the systematic assessment of how simultaneous stressors defined by the scenarios influence nutrient dynamics at the basin scale.

Under the five Colorado Water Plan scenarios representing alternative futures for population growth, land use density, and regulatory frameworks, source contributions varied substantially **(**Figure 5- 3**)**. Business-as-usual conditions (Scenario A) projected increases in anthropogenic TN and TP loads by 2050, driven primarily by population growth and corresponding increases in WWTF discharge. Scenarios incorporating enhanced regulations (Scenarios C and D) achieved dramatic load reductions through the implementation of advanced treatment technologies at WWTFs, stormwater control measures in MS4 areas, and conservation practices on agricultural lands. Agricultural contributions declined across all scenarios due to projected reductions in irrigated acreage from buy-and-dry water transfers, with decreases ranging from modest reductions under Scenario A to substantial declines under growth scenarios with greater urban water demand.

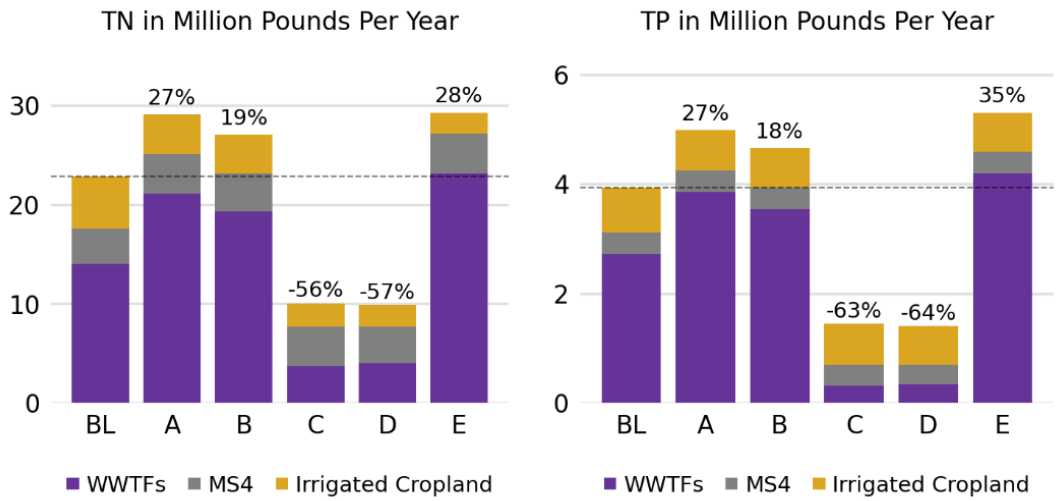


Figure 5- 3 Aggregated Total Nitrogen (TN) and Total Phosphorus (TP) loads from Wastewater Treatment Facilities (WWTFs), Municipal Separate Storm Sewer Systems (MS4), and Irrigated Cropland under the Baseline (BL) conditions and five Colorado Water Plan future scenarios

## 5.3 MATERIALS AND METHODS

### 5.3.1 STUDY AREA

The study encompassed river basins across Colorado (Figure 5- 4), spanning diverse physiographic provinces from high-elevation montane headwaters in the Rocky Mountains to semi-arid plains in the eastern portion of the state. Colorado's river basins exhibit substantial gradients in climate, topography, land use, and hydrologic regime, providing a rigorous test environment for watershed model evaluation. Elevation ranges from over 4,000 m in alpine headwaters to below 1,000 m at the eastern state boundary. Mean annual precipitation varies from over 1,000 mm in mountain watersheds to less than 400 mm in plains regions. Major land uses include forest and rangeland in mountainous areas, intensive irrigated agriculture in river valleys, and expanding urban development along the Front Range corridor.

The Cache la Poudre River watershed (HUC 10190007) served as the focal watershed for detailed calibration methodology development. Located in north-central Colorado, this watershed drains approximately 4,900 km² from its headwaters at 3,700 m elevation in Rocky Mountain National Park to its confluence with the South Platte River at 1,450 m elevation. The watershed exhibits pronounced longitudinal gradients in geology, vegetation, and land use (Clausen, 2021). The upper watershed comprises granitic and metamorphic bedrock supporting coniferous forest and alpine tundra, while the lower watershed transitions to sedimentary geology with mixed agricultural lands, grasslands, and urban areas including the city of Fort Collins. The river's flow regime is snowmelt-dominated with peak flows typically occurring in May-June, though extreme flooding events have been documented following intense frontal precipitation (Garner et al., 2016).

Nutrient inputs to the Cache la Poudre River derive from multiple sources including municipal and industrial wastewater treatment facilities, agricultural fertilizer application and livestock operations, urban stormwater runoff, and natural landscape contributions from forested and rangeland areas (Son et al., 2013). The watershed's transition from minimally impacted headwaters to heavily-managed lower reaches creates strong spatial gradients in nutrient concentrations and provides diverse conditions for model evaluation. Previous studies have documented impacts of urban development and wastewater inputs on nutrient dynamics and emerging contaminants in the river (Pruden et al., 2006; Storteboom et al., 2010).

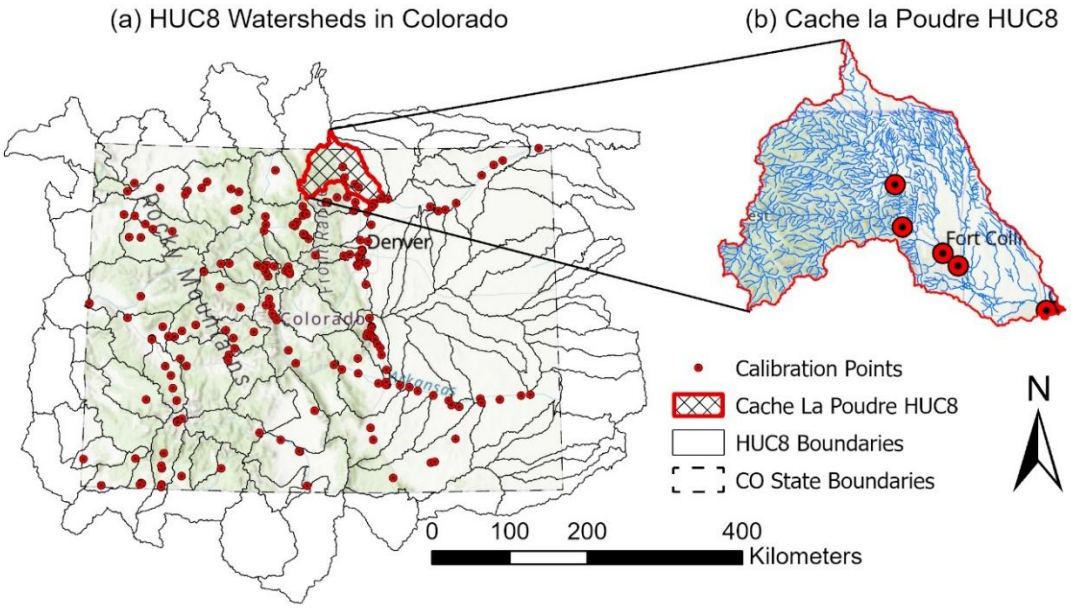


Figure 5- 4 Study area encompassing Colorado river basins with detailed focus on Cache la Poudre HUC8. Circles indicate calibration monitoring stations (n=200 state-wide, n=6 in Cache la Poudre).

### 5.3.2 SWIFT MODEL FRAMEWORK AND INTEGRATION WITH SWAT+

SWIFT accepts flow outputs from SWAT+ models and applies object-oriented routing algorithms to simulate constituent transport through river networks. The model represents watersheds as

collections of discrete hydrologic objects including Hydrologic Response Units (HRUs), routing units, channel reaches, reservoirs, and aquifers, each with defined properties and process representations. This modular architecture enables flexible watershed configuration while maintaining computational efficiency through streamlined mass balance calculations.

Foundation hydrologic simulations derived from SWAT+ models previously calibrated for streamflow by collaborators at Colorado State University. These models provided daily flow outputs for surface runoff, lateral flow, tile drainage, and groundwater components at the HRU scale. A custom subroutine converted SWAT+ outputs to SWIFT input file formats, translating object connectivity, hydrograph components, and watershed properties while preserving flow mass balance.

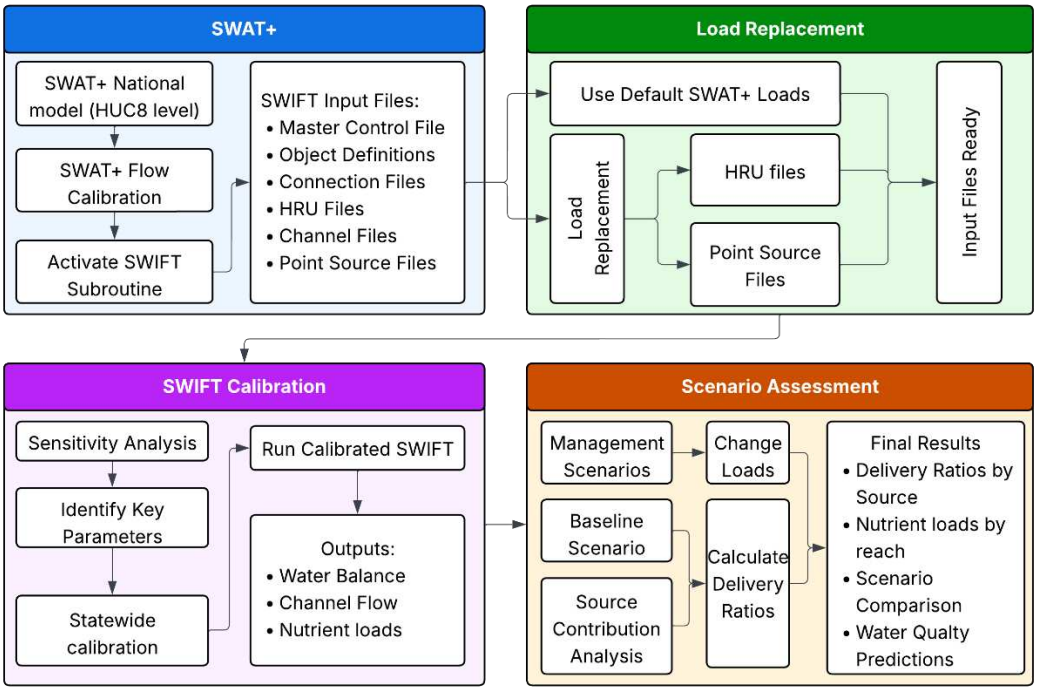


Figure 5- 5 Workflow for integrating SWAT+ hydrology with SWIFT nutrient routing.

### 5.3.3 NUTRIENT LOAD INTEGRATION AND SOURCE-SPECIFIC INPUT PROCESSING

SWIFT's flexible architecture enables substitution of SWAT+-calculated nutrient loads with externally derived source-specific estimates. This capability proved essential for integrating high-quality nutrient load data from multiple sources while leveraging SWAT+'s robust hydrologic framework. Nutrient loads for four major source categories were incorporated: (1) wastewater treatment facilities, (2) urban stormwater runoff, (3) agricultural lands, and (4) forest and rangeland areas.

Wastewater treatment facility and urban stormwater loads were handled as point source inputs in SWIFT. For each HUC12 subwatershed, point source files were populated with the sum of WWTF discharge loads and urban stormwater contributions. These loads, derived from previous source characterization efforts, provided annual total nitrogen (TN) and total phosphorus (TP) masses.

Agricultural and forest/rangeland nutrient loads were integrated as non-point source contributions through modification of HRU export coefficient files (hru_exco.swf). This process requires calculating nutrient concentrations from externally derived load estimates and SWAT+-simulated flow volumes. For each HRU, the procedure involved: (1) extracting HRU number, HUC12 code, area, and land class from spatial data files; (2) merging with precipitation data to calculate water flow volumes; (3) matching externally-derived annual loads by HUC12 and land class; (4) simplifying land classifications into three categories (agricultural: cropland, hay, and pasture; forest/rangeland: forest and rangeland; urban: impervious surfaces); (5) selecting representative HRUs for each HUC12-land class combination; and (6) calculating constituent concentrations as load divided by flow volume.

For agricultural HRUs, concentrations were calculated separately for surface runoff and total flow pathways. Organic nitrogen, sediment-bound phosphorus, nitrate, soluble phosphorus, ammonia, and nitrite concentrations were computed by dividing annual loads (kg/yr) by annual flow volumes (converted from mm·ha to liters) and applying unit conversion to obtain mg/L values. Similar calculations were performed for forest/rangeland HRUs using land-specific load estimates. Urban land HRUs had all non-point concentrations set to zero since these contributions were captured in point source files. For HRUs not matching agricultural, forest/rangeland, or urban classifications, original SWAT+ flow and sediment values were retained while nutrient concentrations were set to zero. This approach ensured comprehensive spatial coverage while avoiding double-counting of nutrient inputs.

### 5.3.4 CHANNEL PARAMETER CHARACTERIZATION

Channel properties governing in-stream processes were characterized by using spatially-distributed data sources to improve physical realism of routing calculations. Channel geometry, hydraulic properties, and sediment characteristics were obtained from NHDPlus (National Hydrography Dataset Plus) and modified based on local knowledge where available. Bank and bed sediment properties, particularly organic matter content influencing nutrient concentrations, were estimated using SSURGO (Soil Survey Geographic Database) soil data.

For each channel reach, average organic matter percentage along the reach corridor was calculated using zonal statistics in GIS. These organic matter values were then converted to nitrogen and phosphorus concentrations using empirical relationships from published literature. Specifically, nitrogen concentration (mg/kg) was estimated as organic matter percentage multiplied by a conversion factor, while phosphorus concentration followed similar scaling relationships. This

spatially distributed approach replaced SWIFT's default assumption of constant N and P concentrations in channel banks and beds, providing more realistic parameterization across diverse watershed settings and enabling improved calibration performance.

### 5.3.5 SENSITIVITY ANALYSIS

Global sensitivity analysis using Sobol methods identified parameters exerting dominant influence on model predictions prior to calibration (Saltelli et al., 2008). Seventeen parameters related to channel hydraulic properties, sediment characteristics, and nutrient concentrations were initially considered based on literature review and expert judgment. Parameter ranges were established from published studies and physical constraints. Candidate parameters included Manning's roughness coefficient (n), channel bottom hydraulic conductivity, bank erodibility coefficient, riparian vegetation cohesion factor (tree_coh), channel sinuosity, bed equilibrium coefficient, median sediment particle size (D50), channel side slope, clay percentage in channel banks (ch_clay), carbon percentage, dry bulk density (ch_bd), floodplain slope, floodplain Manning's n, nitrogen concentration in channel banks (n_conc), phosphorus concentration in channel banks (p_conc), fraction of bioavailable phosphorus (p_bio), and floodplain median sediment size.

Sobol analysis employed Saltelli's sampling scheme to generate parameter combinations spanning the defined ranges (Saltelli et al., 2002). For computational feasibility, the sensitivity analysis focused on a subset of monitoring stations representing diverse watershed positions and flow regimes. A total of 2,048 parameter combinations were generated and executed through SWIFT. For each combination, parameters were scaled relative to baseline values in input files, SWIFT simulations were run, and annual average TN and TP loads were extracted at monitoring locations.

First-order (S1) and total-order (ST) Sobol indices were calculated to quantify individual parameter influence and parameter interaction effects, respectively (Sobol, 2001). Parameters exhibiting high total-order indices for either TN or TP predictions were retained for subsequent calibration efforts. This analysis identified six parameters as most influential: tree cohesion factor (tree_coh), channel clay percentage (ch_clay), channel bulk density (ch_bd), nitrogen concentration in banks (n_conc), phosphorus concentration in banks (p_conc), and bioavailable phosphorus fraction (p_bio).

### 5.3.6 CALIBRATION STRATEGY

#### 5.3.6.1 CACHE LA POUDRE DETAILED CALIBRATION

The Cache la Poudre watershed served as a testbed for developing spatially distributed calibration procedures prior to state-wide application. Multiple monitoring stations within the watershed spanning headwaters to outlet enabled evaluation of model performance across contrasting landscape settings. Water quality data including nutrient concentrations and flow measurements were obtained from the Colorado Department of Public Health and Environment (CDPHE) and U.S. Geological Survey databases.

Annual nutrient loads at each monitoring station were calculated using the USGS LOADEST program, which employs regression techniques to relate concentration measurements to flow and season, enabling estimation of loads from sparse concentration data paired with continuous flow records (Runkel et al., 2004). LOADEST-derived average annual loads provided calibration targets representing long-term average nutrient export at each location.

Calibration proceeded sequentially, first optimizing parameters affecting total phosphorus predictions, then refining parameters influencing total nitrogen. This sequential approach was

adopted because TP dynamics are strongly controlled by sediment-associated processes (erosion, deposition) while TN dynamics include both particulate and dissolved fractions with distinct transport mechanisms. The five parameters identified from sensitivity analysis for TP (tree_coh, ch_clay, ch_bd, p_conc, p_bio) were calibrated first, followed by adjustment of n_conc for TN.

The calibration employed a systematic parameter scaling approach using Saltelli's sampling method to generate combinations of parameter multipliers. For each parameter combination, the channel parameter input file (chan_dat.swf) was modified by multiplying baseline spatially distributed parameter values by the scaling factors. SWIFT was then executed and simulated annual TN and TP loads were extracted at monitoring locations. This process was repeated for 2,048 parameter combinations.

Model performance was evaluated using multiple metrics including coefficient of determination ($R^2$), percent bias (PBIAS), Nash-Sutcliffe efficiency (NSE), and absolute relative error (ARE). The parameter combination yielding optimal performance across stations was selected as the calibrated parameter set. Given that baseline parameters already incorporated spatial variation through the organic-matter-based N and P concentrations, the scaling factors provided watershed-specific adjustment while preserving spatial patterns.

#### 5.3.6.2 STATE-WIDE CALIBRATION

Building on methodology developed for Cache la Poudre, calibration was extended to encompass approximately 200 monitoring stations distributed across Colorado's major river basins. Monitoring data sources included CDPHE's water quality monitoring network and USGS National Water Information System. Station selection criteria required sufficient paired flow and concentration data to support reliable LOADEST load estimation.

Given the computational expense of calibrating across 200 stations simultaneously, HUC8 watersheds were grouped into calibration units sharing similar characteristics. Grouping criteria included: (1) dominant land use (agricultural, forest/rangeland, urban-influenced), (2) HUC8 watershed level (e.g. headwater (level 1) or downstream HUC8 (e.g. level 2, 3, 4, etc.), and (3) geographic region corresponding to Colorado's major river basins and water divisions. This stratification resulted in 7 calibration groups per water division, each containing 2-10 HUC8 watersheds with similar hydrologic and land use conditions.

Calibration proceeded sequentially by group. For each group, the same Sobol-based parameter sampling and systematic evaluation procedure applied in Cache la Poudre was executed across all monitoring stations within the group's HUC8s (Figure 5- 6). Parameter scaling factors optimal for each group were determined and applied to modify channel parameter files. This group-based approach balanced computational feasibility with recognition of regional variation in watershed processes and parameter relationships.

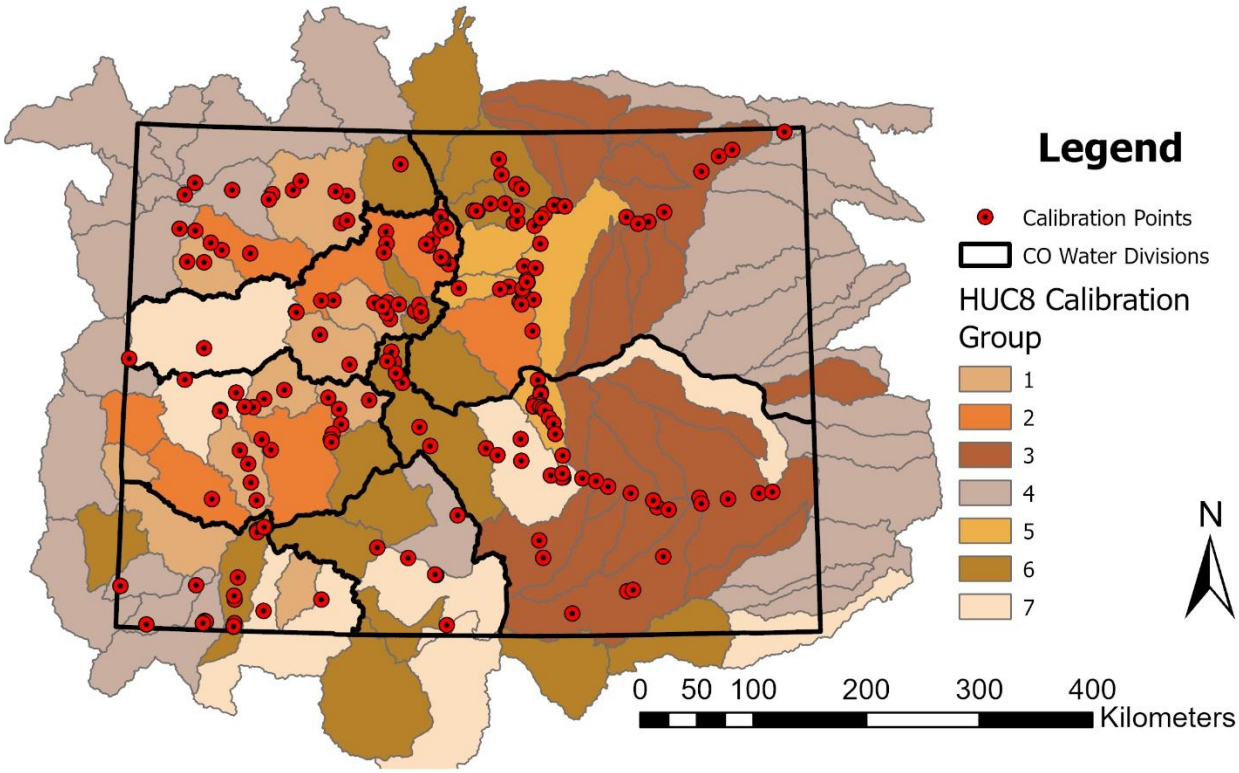


Figure 5- 6 Spatial distribution of HUC8 calibration groups across Colorado. Watersheds were grouped based on dominant land use, watershed level, and geographic region to enable efficient state-wide calibration while accounting for regional heterogeneity.

### 5.3.7 PERFORMANCE EVALUATION AND COMPUTATIONAL EFFICIENCY ASSESSMENT

Model performance was quantified using standard statistical metrics. The coefficient of determination ($R^2$) measured correlation between observed and predicted loads. Percent bias (PBIAS) quantified systematic over- or under-prediction, calculated as the sum of differences between predictions and observations divided by the sum of observations, multiplied by 100. Nash-Sutcliffe efficiency (NSE) provided an aggregate measure of model skill relative to a mean-value baseline.

Absolute relative error (ARE) quantified prediction accuracy at individual stations. Because observed loads were estimated using LOADEST (Runkel et al., 2004) from sparse concentration measurements and continuous flow records, inherent uncertainty exists in these load estimates. Following Yen et al. (2014), ARE calculations were adjusted to account for measurement uncertainty using the probability distribution method of Harmel and Smith (2007). When predictions fell within the observational uncertainty bounds, the apparent error was reduced through correction factors.

Computational efficiency was assessed by comparing SWIFT execution time to SWAT+ run time for representative HUC8 watersheds. Timing comparisons used identical computing hardware and measured time from simulation initiation to completion. Speedup ratios were calculated as SWAT+ run time divided by SWIFT run time, quantifying computational gains enabling extensive scenario analysis and calibration exercises impractical with traditional SWAT+ applications.

### 5.3.8 SOURCE CONTRIBUTION ANALYSIS USING LOAD SUBSTITUTION

SWIFT's load substitution capability was leveraged to quantify nutrient delivery ratios from specific sources to downstream receiving waters. This analysis addressed the management question: What fraction of nutrients released from a given source type reaches a downstream monitoring location, accounting for in-stream retention and transformation processes?

For each of three major source categories (WWTFs, agricultural lands, urban stormwater), separate SWIFT simulations were conducted where only loads from that source were included while all other sources were set to zero. The calibrated channel parameter sets were maintained to ensure consistent process representation. Each simulation provided predictions of TN and TP loads at monitoring stations resulting solely from the isolated source category.

Delivery ratios were calculated as the simulated load at a downstream location divided by the total input load from that source type within the upstream contributing area. Delivery ratios less than unity indicate net retention through in-stream processes including sedimentation, denitrification, and biological uptake. Spatial patterns in delivery ratios reveal where specific sources exert greatest influence on downstream water quality, informing targeting of management interventions for maximum benefit.

This source-specific analysis exemplified SWIFT's operational utility for decision support. Computational efficiency enabling multiple scenario runs (baseline plus three source-isolation scenarios per watershed) within practical time frames would be prohibitive with traditional SWAT+ approaches requiring hours per simulation. The rapid execution facilitated comprehensive source contribution assessment across the entire state.

## 5.4 RESULTS

### 5.4.1 SENSITIVITY ANALYSIS

Global sensitivity analysis using Sobol methods identified channel clay percentage (ch_clay) as the dominant parameter controlling both TN and TP predictions, with total-order indices (ST) exceeding 0.95 for both nutrients across all monitoring stations (Figure 5- 7). First-order indices (S1) for ch_clay ranged from 0.84 to 0.90, indicating that direct effects dominated over parameter interactions. This finding reflects ch_clay's dual influence on model outputs through control of bank erodibility and baseline nutrient concentrations in channel sediments.

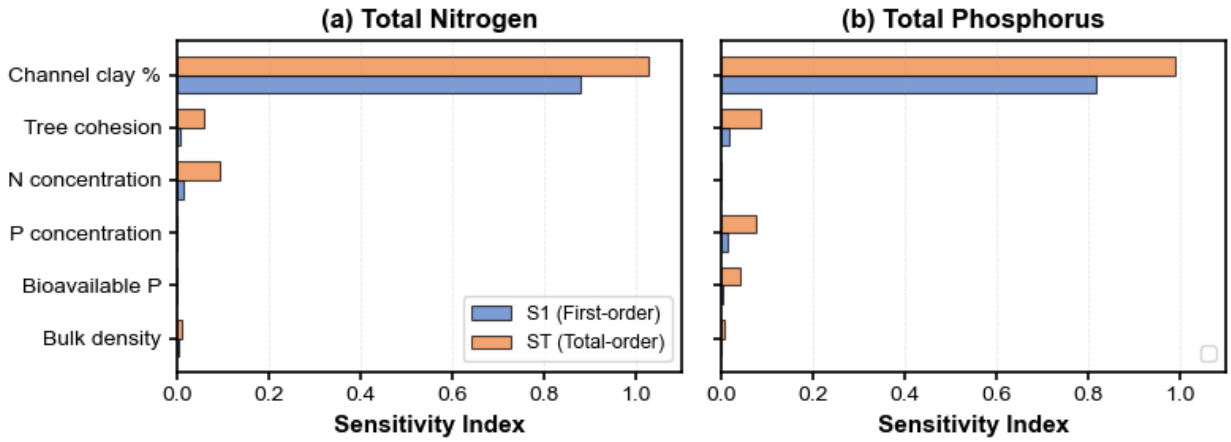


Figure 5- 7 Global sensitivity analysis results showing first-order (S1, light bars) and total-order (ST, dark bars) Sobol indices for channel and sediment parameters affecting (a) total nitrogen and (b) total phosphorus predictions

Nutrient-specific parameters exhibited expected selective sensitivities. Nitrogen concentration in channel banks (n_conc) showed meaningful sensitivity only for TN predictions (ST = 0.09), while phosphorus concentration (p_conc) and bioavailable phosphorus fraction (p_bio) influenced only TP predictions (ST = 0.07-0.08). Tree cohesion (cov), representing riparian vegetation stabilization effects, displayed moderate sensitivity for both nutrients (ST = 0.06-0.15), with higher ST values relative to S1 indicating substantial interaction effects with other parameters. Bulk density (ch_bd) exhibited minimal direct influence (ST < 0.01) across all stations and nutrients. Parameter ranges used in the sensitivity analysis are provided in Table 5- 1.

Table 5- 1 Parameter ranges used in Sobol sensitivity analysis. Bounds represent scaling factors applied to spatially-distributed baseline values derived from soil organic matter content (n_conc, p_conc) or literature-based physical ranges (cov, ch_clay, ch_bd, p_bio).

| Parameter | Description | Lower Bound | Upper Bound |
|---|---|---|---|
| cov | Tree cohesion factor | 0 | 6000 |
| ch_clay | Channel clay % (scaling factor) | 0.6 | 1.8 |
| ch_bd | Bulk density (scaling factor) | 0.2 | 1.6 |

| Parameter | Description | Lower Bound | Upper Bound |
|---|---|---|---|
| n_conc | N concentration (scaling factor) | 0.5 | 1.5 |
| p_conc | P concentration (scaling factor) | 0.5 | 1.5 |
| p_bio | Bioavailable P fraction (scaling factor) | 0.3 | 3.0 |

### 5.4.2 CACHE LA POUDRE HUC8 CALIBRATION PERFORMANCE

Calibration of the Cache la Poudre watershed using 2,048 parameter combinations identified optimal scaling factors through systematic Sobol sampling. Final calibrated parameter values are presented in Table 2. Tree cohesion (cov) calibrated to 5,571, approaching the upper bound of the search space, consistent with well-vegetated riparian corridors characteristic of montane Colorado streams. Channel clay percentage scaled to 1.78 times baseline values, while nutrient concentration scalars remained near unity (n_conc = 0.54, p_conc = 1.12), indicating reasonable accuracy of organic matter-based baseline estimates.

Table 5- 2 Final calibrated parameter values for Cache la Poudre HUC8. Values represent optimal parameter set identified through Sobol-based systematic sampling with 2,048 model evaluations.

| Parameter Code | Parameter Name | Calibrated Value |
|---|---|---|
| cov | Tree cohesion factor | 5,570.80 |
| ch_clay | Channel clay % (scaling) | 1.78 |
| ch_bd | Bulk density (scaling) | 0.25 |
| n_conc | N concentration (scaling) | 0.54 |
| p_conc | P concentration (scaling) | 1.12 |
| p_bio | Bioavailable P fraction (scaling) | 1.26 |

Longitudinal profiles of simulated versus observed mean annual nutrient loads demonstrated strong agreement along the 90 km river corridor from headwaters to confluence (Figure 5- 8). For TN, simulated loads tracked closely with observed values at all monitoring stations, with predictions falling within uncertainty bounds (based on LOADEST estimation error) at five of six stations. Loads increased progressively from approximately 100,000 kg/yr in headwater reaches to 700,000 kg/yr at the watershed outlet, capturing both tributary inputs and mainstem accumulation. TP predictions exhibited similar spatial patterns with loads increasing from <500 kg/yr in upper reaches to 200,000 kg/yr at the outlet. A tendency toward overestimation in middle reaches (km 42-66) was observed, with predicted values exceeding upper uncertainty bounds by 15-25%, before converging with observations at the outlet station.

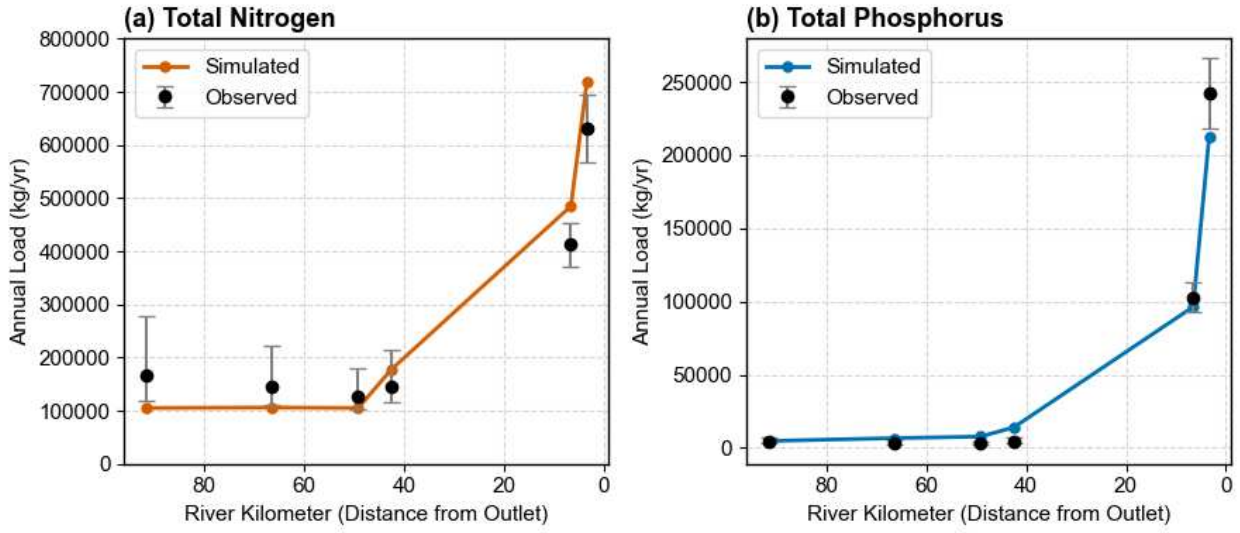


Figure 5- 8 Comparison of simulated and observed annual nutrient loads versus river distance in Cache la Poudre HUC8 for (a) Total Nitrogen and (b) Total Phosphorus. Error bars represent uncertainty in observed loads estimated from LOADEST. X-axis represents distance upstream from watershed outlet.

### 5.4.3 STATE-WIDE CALIBRATION PERFORMANCE

Extension of calibration procedures to approximately 200 monitoring stations across Colorado demonstrated robust model performance across diverse watershed conditions. Simulated versus observed loads exhibited strong linear relationships spanning nearly four orders of magnitude from small headwater streams (<1,000 kg/yr) to major river outlets (>1,000,000 kg/yr) (Figure 5- 9). Scatter plots showed tight clustering along the 1:1 line with no systematic bias as load magnitudes increased, indicating consistent performance across the full range of hydrologic scales and nutrient export rates encountered in Colorado's river basins.

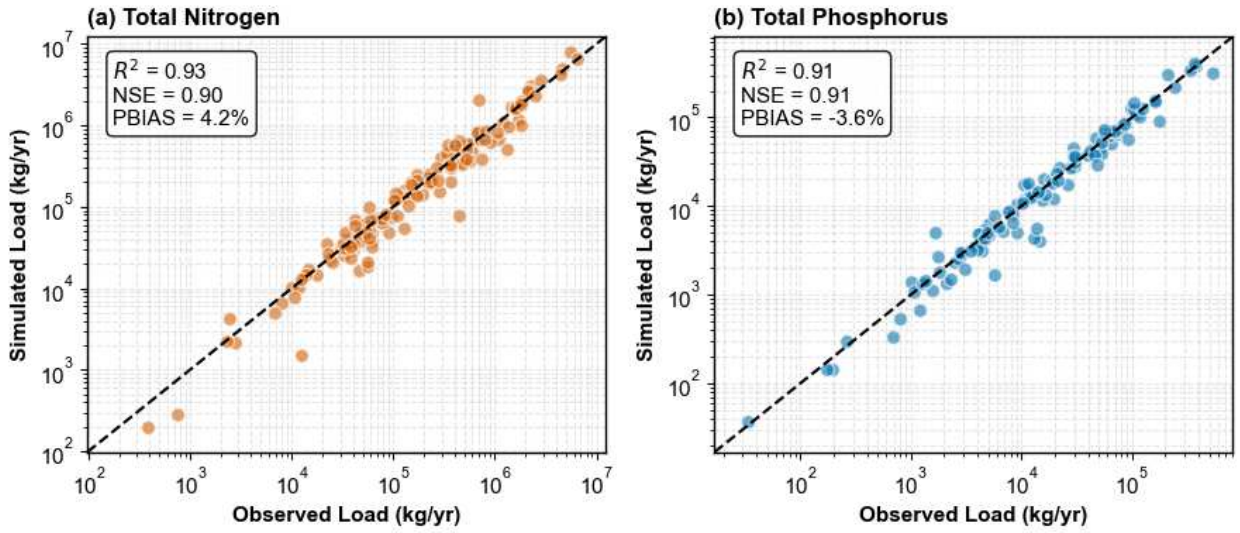


Figure 5- 9 Comparison of simulated versus observed mean annual loads for (a) Total Nitrogen (TN) and (b) Total Phosphorus (TP) across 200 calibrated stations state-wide. Data plotted on logarithmic scale to accommodate four orders of magnitude in load variability. Dashed black line represents 1:1 perfect agreement.

State-wide performance statistics indicated strong predictive capability (Table 3- 3). Coefficient of determination values reached 0.93 for TN and 0.91 for TP, demonstrating high correlation between predictions and observations. Nash-Sutcliffe efficiency values of 0.90 (TN) and 0.91 (TP) exceeded the 0.75 threshold for "very good" performance established by Moriasi et al. (2007).

Percent bias remained minimal at 4.17% for TN and -3.64% for TP, well within the ±10% range indicating very good performance for nutrient simulation.

Table 3- 3 Summary statistics for state-wide model calibration performance across 200 monitoring stations. $R^2$ = coefficient of determination, NSE = Nash-Sutcliffe efficiency, PBIAS = percent bias, ARE = absolute relative error. Corrected ARE accounts for measurement uncertainty in LOADEST-estimated observed loads following Yen et al. (2014) methodology.

| **Metric** | **Total Nitrogen (TN)** | **Total Phosphorus (TP)** |
|---|---|---|
| $R^2$ | 0.93 | 0.91 |
| NSE | 0.90 | 0.91 |
| PBIAS (%) | 4.17 | -3.64 |
| Mean ARE | 26% | 22% |
| Mean Corrected ARE | 9.1% | 7.8% |

Standard absolute relative error averaged 26% for TN and 22% for TP across all stations (Figure 5- 10, Table 3- 3). When corrected for measurement uncertainty in LOADEST-derived observed loads using the probability distribution method (Harmel and Smith, 2007; Yen et al., 2014), mean ARE values decreased substantially to 9.1% for TN and 7.8% for TP. The distribution of corrected ARE compressed toward zero, with median values below 8% for both nutrients and interquartile ranges spanning 4-14%. This reduction indicates that substantial portions of apparent model error derived from observational uncertainty rather than model structural inadequacy, consistent with findings by Harmel et al. (2010) that measurement uncertainty significantly influences perceived model performance.

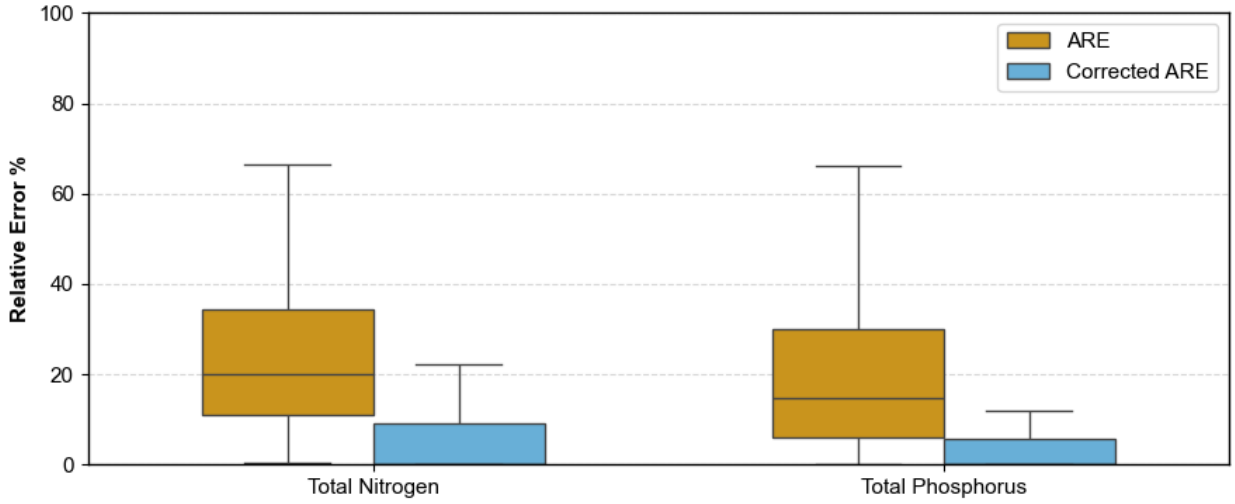


Figure 5- 10 Distribution of model prediction errors for Total Nitrogen and Total Phosphorus across state-wide monitoring network. Box plots compare standard Absolute Relative Error (ARE) against Corrected ARE which adjusts for ±30% observational uncertainty in LOADEST-estimated loads following Yen et al. (2014).

### 5.4.4 SOURCE CONTRIBUTION AND DELIVERY RATIO ANALYSIS

Load substitution simulations isolating individual nutrient sources enabled quantification of delivery ratios from source locations to downstream monitoring points. State-wide aggregate analysis revealed systematic differences in delivery efficiency among source types (Figure 5a). Wastewater treatment facilities (WWTFs) exhibited highest delivery ratios, averaging 0.79 for TN and 0.82 for TP, reflecting direct discharge to stream channels with minimal landscape retention. Municipal separate storm sewer systems (MS4s) showed intermediate delivery ratios of 0.77 (TN) and 0.80 (TP). Irrigated cropland demonstrated lowest delivery efficiency at 0.75 for TN and 0.73 for TP, consistent with diffuse transport pathways through soil profiles and riparian buffers providing greater retention opportunities.

Distribution of delivery ratios calculated at HUC8 outlets exhibited substantial spatial variability, particularly for agricultural sources (Figure 5- 11b). WWTF delivery ratios displayed narrow

distributions (interquartile range 0.89-0.96) with median values near 0.93 for both nutrients, indicating consistent high efficiency across diverse watershed settings. In contrast, irrigated cropland delivery ratios spanned wider ranges (interquartile range 0.68-0.85) with median values near 0.77, reflecting heterogeneity in basin characteristics including slope, drainage density, soil properties, and extent of riparian buffers. Forest and rangeland sources showed intermediate variability with median delivery ratios of 0.81-0.83.

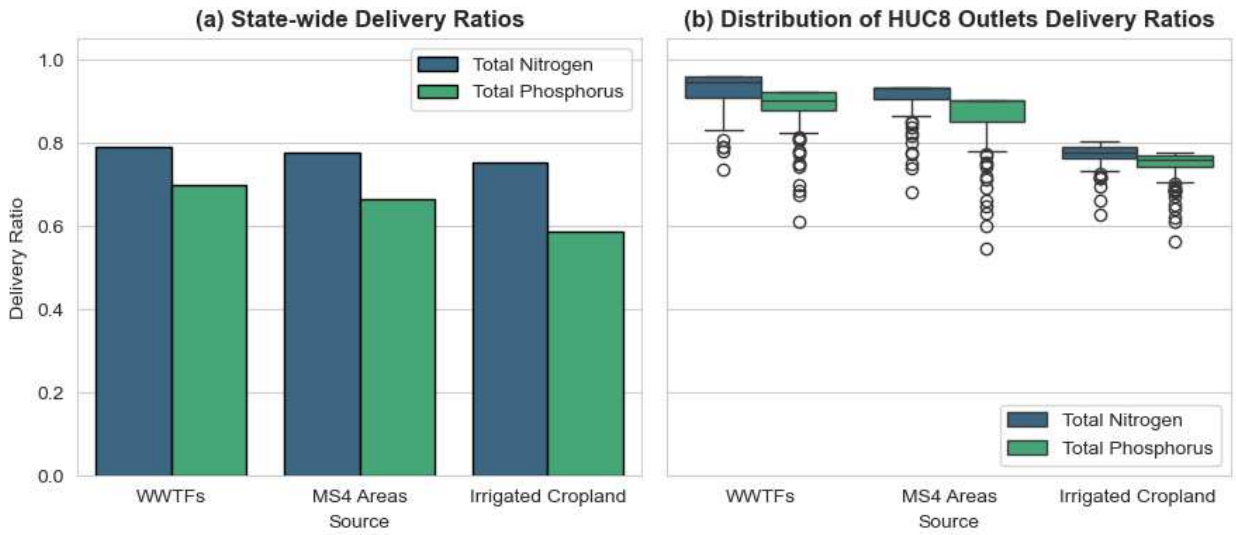


Figure 5- 11 Comparison of nutrient delivery ratios across source types. (a) State-wide aggregate delivery ratios for Total Nitrogen (TN) and Total Phosphorus (TP) by source category. (b) Distribution of delivery ratios calculated at 8-digit Hydrologic Unit Code (HUC8) watershed outlets.

### 5.4.5 COMPUTATIONAL EFFICIENCY

SWIFT execution times demonstrated dramatic computational advantages relative to SWAT+ simulations. For a representative watershed, SWAT+ required approximately 7,200 seconds (2 hours) for a complete simulation including daily outputs for the 20-years period. The equivalent SWIFT simulation executed in 3.8 seconds, yielding a speedup factor of approximately 1,900. Computational time scaled approximately linearly with watershed area, with smaller HUC8s

(<2,000 km²) completing in under 1 second and larger basins (>8,000 km²) requiring 3-4 seconds. This efficiency enabled the 2,048-member ensemble sensitivity analysis for Cache la Poudre to complete in 2.2 hours compared to an estimated 4,000+ hours that would have been required using SWAT+, rendering comprehensive Sobol analysis tractable for operational watershed assessment.

## 5.5 DISCUSSION

### 5.5.1 SENSITIVITY ANALYSIS AND PARAMETER IDENTIFICATION

The dominant sensitivity of channel clay percentage (ST > 0.95) reflects its dual mechanistic influence on nutrient transport through control of both bank erodibility and baseline sediment nutrient content. This finding aligns with observations by Noe et al. (2022) that streambank erosion represents a major contributor to watershed nutrient budgets, particularly in agricultural landscapes where channel incision and bank destabilization are prevalent. The high first-order indices (S1 ~ 0.85) indicate that ch_clay effects operate primarily through direct pathways rather than complex parameter interactions, supporting the use of spatially distributed clay content as a primary calibration target.

The nutrient-specific sensitivities of n_conc and p_conc parameters (affecting only TN and TP respectively) validate the conceptual separation of nitrogen and phosphorus dynamics in SWIFT's process representation. Nitrogen sensitivity to n_conc reflects the importance of bank sediment as a source during erosion events, while lower sensitivity relative to ch_clay suggests that loading from external sources (WWTFs, agriculture) dominates the TN budget. For phosphorus, the combined sensitivity of p_conc and p_bio (total ST ~ 0.15) indicates that both source strength and

bioavailability govern TP fate, consistent with findings by Diehl et al. (2022) that floodplain sediment phosphorus content significantly influences riverine P dynamics.

The moderate sensitivity of tree cohesion (cov = 0.06-0.15) and minimal sensitivity of bulk density (ST < 0.01) contrasts with their theoretical importance in bank stability models. This likely reflects collinearity with ch_clay, which exerts stronger control on erosion rates through the critical velocity formulation. The difference between total-order and first-order indices for cov (ST - S1 ~ 0.05-0.08) indicates meaningful interaction effects, possibly through coupling with channel geometry parameters that influence shear stress distribution. These findings emphasize that global sensitivity analysis provides essential guidance for efficient calibration by identifying parameters warranting detailed attention versus those adequately represented by default or literature values.

### 5.5.2 CALIBRATION PERFORMANCE AND PHYSICAL CONSISTENCY

Achievement of NSE values exceeding 0.90 and PBIAS within ±5% across 200 diverse monitoring stations represents performance comparable to or exceeding published watershed nutrient modeling studies. Moriasi et al. (2007) established NSE > 0.75 as "very good" performance for nutrient simulation; the state-wide average of 0.90-0.91 demonstrates that SWIFT's simplified routing framework captures essential transport and transformation processes without sacrificing predictive capability relative to more computationally intensive approaches. This performance is notable given that many watershed models achieve acceptable calibration at outlet gages but exhibit degraded performance at interior locations (White and Chaubey, 2005; Daggupati et al., 2015), whereas SWIFT maintained consistent accuracy across the spatial monitoring network.

The substantial reduction in absolute relative error when accounting for measurement uncertainty (26% → 9% for TN; 22% → 8% for TP) highlights a critical but often-neglected aspect of model

evaluation. As demonstrated by Harmel et al. (2010) and applied by Yen et al. (2014), observational uncertainty in LOADEST-estimated loads can reach ±30% due to sparse temporal sampling and measurement error in both flow and concentration. The corrected ARE values approaching 8% suggest that SWIFT predictions approach the theoretical accuracy limit imposed by available observational data, with remaining discrepancies potentially attributable to measurement uncertainty not fully captured by the probable error range. This finding underscores the need for watershed modeling studies to explicitly consider observational uncertainty when evaluating model performance, as conventional metrics may overstate model inadequacy by attributing observational error to structural model deficiencies.

The longitudinal agreement in Cache la Poudre (Figure 5- 8) demonstrates SWIFT's capacity to correctly accumulate loads from distributed sources while simulating in-stream attenuation over 90+ km of channel length. The progressive load increase from headwaters to outlet, coupled with maintenance of predictions within observational uncertainty bounds, indicates that the model accurately represents the balance between tributary inputs, point source contributions, and in-stream retention processes.

### 5.5.3 COMPUTATIONAL EFFICIENCY AND OPERATIONAL IMPLICATIONS

The 1,900-fold speedup achieved by SWIFT relative to SWAT+ fundamentally transforms the feasibility of comprehensive watershed assessment activities. Computational limitations have historically constrained watershed modeling applications to limited scenario analysis and coarse uncertainty quantification due to the impracticality of executing thousands of model runs (Cibin and Chaubey, 2015; Jeon et al., 2014). SWIFT's execution times of 1-8 seconds per simulation enable systematic exploration of parameter uncertainty, comprehensive sensitivity analysis, and

extensive scenario evaluation that were previously computationally prohibitive. The 2.2-hour completion time for a 2,048-member Sobol ensemble, compared to 4,000+ hours required by SWAT+, exemplifies this transformation.

This efficiency derives from SWIFT's hybrid architecture that separates hydrologic simulation from constituent routing. By accepting pre-calibrated SWAT+ flow outputs and applying streamlined mass balance calculations for nutrient transport, SWIFT avoids repeated computation of land-phase processes while maintaining physically based process representation. This approach contrasts with empirical regression models such as SPARROW (Moore et al., 2011) that achieve computational efficiency by sacrificing detailed process simulation, and with full process-based models like SWAT that require hours for comprehensive simulation. SWIFT occupies a valuable middle ground, providing process-level detail with statistical model efficiency.

The operational implications extend beyond research applications to practical water quality management. Total Maximum Daily Load (TMDL) development typically requires evaluation of numerous load allocation scenarios to identify cost-effective strategies for achieving water quality standards. SWIFT enables rapid assessment of management alternatives, such as comparing effectiveness of WWTF upgrades versus agricultural best management practice implementation, by running dozens of scenarios in minutes rather than days. Similarly, the load substitution capability facilitates source-specific delivery ratio calculations essential for targeting management investments, as demonstrated by the delivery ratio analysis revealing systematically different transport efficiencies among source types.

### 5.5.4 SOURCE CONTRIBUTION AND MANAGEMENT IMPLICATIONS

The systematic differences in delivery ratios among source types (WWTFs: 0.79-0.82; MS4s: 0.77-0.80; Agriculture: 0.73-0.75) reflect fundamental differences in nutrient transport pathways and retention opportunities. WWTF effluent enters streams directly, experiencing only in-stream processes including dilution, denitrification, and biological uptake. The high delivery ratios (median 0.93 at HUC8 outlets) indicate limited in-stream attenuation for point sources, consistent with findings by Wollheim et al. (2008) that large rivers with short residence times exhibit low nutrient removal efficiency. In contrast, agricultural nutrients traverse overland flow paths, soil profiles, and riparian zones before reaching channels, providing multiple opportunities for retention, plant uptake, and denitrification (Jones et al., 2015; Natho et al., 2020). The lower agricultural delivery ratios (median 0.77) and wide spatial variability (interquartile range 0.68-0.85) reflect this complex transport through heterogeneous landscapes.

These findings have direct implications for nutrient management prioritization. Given equivalent nutrient application rates, management efforts targeting point sources will produce more immediate downstream water quality improvements due to high delivery efficiency and direct channel connectivity. Conversely, agricultural management practices exhibit lower delivery ratios but offer greater potential for landscape-scale retention through riparian buffers, conservation tillage, and controlled drainage (Golden et al., 2019; Mazur et al., 2021). The spatial variability in agricultural delivery ratios across HUC8s (Figure 5- 11b) emphasizes the importance of basin-specific assessment rather than uniform regulatory approaches, as delivery efficiency depends strongly on local factors including slope, soil hydraulic properties, and riparian buffer extent.

The delivery ratio framework developed here enables cost-effectiveness analysis for nutrient reduction strategies by explicitly accounting for transport efficiency from source to downstream water quality monitoring points. For example, reducing agricultural inputs by 1,000 kg TN/yr in a watershed with 0.75 delivery ratio yields 750 kg/yr reduction at the outlet, whereas equivalent WWTF reductions with 0.93 delivery ratio produce 930 kg/yr downstream benefit. When combined with economic data on management practice costs, these physically based delivery ratios provide quantitative foundation for optimization of nutrient reduction investments, addressing the question posed by Maringanti et al. (2009) regarding optimal spatial placement of best management practices.

### 5.5.5 LIMITATIONS AND SOURCES OF UNCERTAINTY

Despite strong performance, several limitations warrant acknowledgment. First, the analysis considered only precipitation as uncertain forcing data; uncertainty in other inputs including temperature, solar radiation, and land management practices could further influence predictions. The input error model employed (Ajami et al., 2007) assumed Gaussian errors which may not fully capture systematic biases or temporal patterns in precipitation measurement uncertainty. Second, model structural uncertainty analysis was limited to two surface runoff methods within SWAT+; uncertainty in other process representations including tile drainage, nutrient cycling, and channel geometry was not explicitly evaluated.

The spatially distributed approach to channel parameter estimation based on soil organic matter content, while providing physically-based spatial variation, introduces uncertainty from empirical relationships between organic matter and nutrient concentrations. These relationships were derived from published literature but may not fully capture site-specific conditions. Alternative approaches

including direct sediment sampling and nutrient analysis would improve parameter estimates but were not feasible given the study scale.

### 5.5.6 FUTURE RESEARCH DIRECTIONS

Several research directions would enhance SWIFT's capabilities and applicability. First, extension to additional constituents including sediment, pesticides, and pathogens would broaden management applications. The object-oriented framework readily accommodates additional state variables, though constituent-specific transformation processes and parameterization would require development. Second, integration with climate change projections would enable assessment of future nutrient dynamics under altered precipitation patterns and temperature regimes. The computational efficiency of SWIFT makes this feasible by enabling ensemble analysis across multiple climate scenarios and emission pathways.

Third, incorporation of dynamic land management scenarios including crop rotations, irrigation schedules, and best management practice implementation would enhance decision support capabilities. Current implementation assumes static land use, but many management questions involve temporal changes in practices. The load substitution framework provides foundation for this extension by enabling time-varying external inputs. Fourth, development of real-time data assimilation capabilities would support operational water quality forecasting. The rapid execution times suggest that ensemble Kalman filtering or particle filtering approaches could be implemented for near-real-time prediction updating as new observations become available.

## 5.6 CONCLUSIONS

This study evaluated the SWIFT framework as a computationally efficient alternative to traditional physically based watershed models for simulating basin-scale nutrient transport. By accepting calibrated hydrologic outputs from SWAT+ and applying streamlined routing algorithms, the hybrid framework achieved a computational speedup of approximately 1,900 times relative to full SWAT+ simulations. This dramatic reduction in execution time enabled the application of comprehensive Sobol global sensitivity analyses, which identified channel clay percentage as the dominant parameter controlling nutrient dynamics (ST > 0.95). These findings validated the use of spatially distributed soil data to constrain channel erodibility and sediment nutrient parameters, ensuring physical consistency across diverse watershed conditions.

The application of a systematic, spatially distributed calibration methodology yielded robust performance across approximately 200 monitoring stations in Colorado. State-wide results demonstrated high predictive capability, with Nash-Sutcliffe efficiency values exceeding 0.90 and coefficients of determination ($R^2$) greater than 0.91 for both total nitrogen and total phosphorus. Furthermore, when accounting for inherent measurement uncertainty in observed loads, the mean absolute relative error decreased to less than 10%, indicating that the simplified routing framework captures essential transport processes with accuracy comparable to more computationally intensive approaches.

Beyond computational gains, the model's flexible load substitution capability facilitated the quantification of source-specific delivery ratios. Analysis revealed systematically higher delivery efficiencies for wastewater treatment facilities (~0.80) compared to agricultural non-point sources (~0.74), reflecting the greater potential for landscape retention in agricultural settings. Collectively, these results establish SWIFT as an operational-ready tool for river basin

management. By bridging the gap between detailed process representation and the need for rapid scenario evaluation, the framework enables decision-makers to optimize nutrient reduction strategies and perform extensive risk assessments that were previously computationally inaccessible.

This chapter contributes an integrated annual routing framework combining SWIFT with source-specific nutrient load predictions for comprehensive watershed assessment at regional scales. Three specific contributions emerge. First, the validated calibration methodology achieves NSE exceeding 0.90 and $R^2$ exceeding 0.91 across 200 monitoring stations spanning diverse watershed conditions, with corrected absolute relative error below 10%, demonstrating that streamlined mass balance routing maintains accuracy comparable to process-based models when operating on calibrated hydrologic inputs. Second, computational efficiency gains of approximately 1,900× relative to SWAT+ transform comprehensive sensitivity analysis from computationally prohibitive to operationally routine (2,048 model runs in 2.2 hours vs. 4,000+ hours), enabling extensive scenario evaluation essential for decision support. Third, this work provides the first state-wide quantification of source-specific delivery ratios, revealing systematic differences (wastewater: 0.80; agriculture: 0.74) and substantial spatial variability (agricultural IQR: 0.68–0.85) that reflect fundamental transport pathway characteristics and directly inform cost-effectiveness optimization of nutrient management investments.

### 5.6.1 INTEGRATION WITH THE BROADER MODELING FRAMEWORK

SWIFT routing completes the dissertation's integrated framework by combining source predictions from Chapters 2-4 with wastewater facility data, achieving NSE > 0.90 and 1,900× computational speedup. For each Colorado Water Plan scenario, the framework integrates: forest/rangeland loads

(Chapter 2 ensemble ML, $R^2$ = 0.70-0.83), urban stormwater (derived from Chapter 3 ConvLSTM land cover predictions), agricultural exports (Chapter 4 physics-informed ML, NSE > 0.85), and point source discharges. This integration enables comprehensive scenario assessment in 2-3 hours versus 400-600 hours for traditional SWAT+ approaches. Source-specific delivery ratios (wastewater: 0.80; agriculture: 0.74) inform management targeting. The strategic deployment of specialized approaches, ensemble ML for natural landscapes, deep learning for spatiotemporal forecasting, physics-informed ML for agricultural systems, streamlined routing for watershed transport, demonstrates that hybrid frameworks can achieve both scientific validity and computational tractability essential for operational water quality management under uncertainty.

## 5.7 REFERENCES


Arnold, J., Bieger, K., White, M., Srinivasan, R., Dunbar, J., & Allen, P. (2018). Use of decision tables to simulate management in swat+. Water, 10(6), 713. https://doi.org/10.3390/w10060713

Ator, S., García, A., Schwarz, G., Blomquist, J., & Sekellick, A. (2019). Toward explaining nitrogen and phosphorus trends in chesapeake bay tributaries, 1992–2012. Jawra Journal of the American Water Resources Association, 55(5), 1149-1168. https://doi.org/10.1111/1752-1688.12756

Cibin, R. and Chaubey, I. (2015). A computationally efficient approach for watershed scale spatial optimization. Environmental Modelling & Software, 66, 1-11. https://doi.org/10.1016/j.envsoft.2014.12.014

Clausen, E. (2021). Topographic map analysis of the north platte river-south platte river drainage divide area, western larimer county, colorado, usa. Earth Science Research, 10(1), 49. https://doi.org/10.5539/esr.v10n1p49

Daggupati, P., Yen, H., White, M., Srinivasan, R., Arnold, J., Keitzer, C., … & Sowa, S. (2015). Impact of model development, calibration and validation decisions on hydrological

simulations in west lake erie basin. Hydrological Processes, 29(26), 5307-5320. https://doi.org/10.1002/hyp.10536

Diehl, R., Underwood, K., Triantafillou, S., Ross, D., Drago, S., & Wemple, B. (2022). Multi-scale drivers of spatial patterns in floodplain sediment and phosphorus deposition. Earth Surface Processes and Landforms, 48(4), 801-816. https://doi.org/10.1002/esp.5519

Garner, E., Wallace, J., Argoty, G., Wilkinson, C., Fahrenfeld, N., Heath, L., … & Pruden, A. (2016). Metagenomic profiling of historic colorado front range flood impact on distribution of riverine antibiotic resistance genes. Scientific Reports, 6(1). https://doi.org/10.1038/srep38432

Gassman, P., Reyes, M., Green, C., & Arnold, J. (2007). The soil and water assessment tool: historical development, applications, and future research directions. Transactions of the Asabe, 50(4), 1211-1250. https://doi.org/10.13031/2013.23637

Golden, H., Rajib, A., Lane, C., Christensen, J., Wu, Q., & Mengistu, S. (2019). Non-floodplain wetlands affect watershed nutrient dynamics: a critical review. Environmental Science & Technology, 53(13), 7203-7214. https://doi.org/10.1021/acs.est.8b07270

Huang, J., Arhonditsis, G., Zhang, Y., Zhang, S., Ji, Y., Paerl, H., … & Gao, J. (2024). The critical role of hydrological distance in shaping nutrient dynamics along the watershed–lake continuum. Earth's Future, 12(10). https://doi.org/10.1029/2024ef004860

Jeon, J., Park, C., & Engel, B. (2014). Comparison of performance between genetic algorithm and sce-ua for calibration of scs-cn surface runoff simulation. Water, 6(11), 3433-3456. https://doi.org/10.3390/w6113433

Jones, C., Scott, D., Guth, C., Hester, E., & Hession, W. (2015). Seasonal variation in floodplain biogeochemical processing in a restored headwater stream. Environmental Science & Technology, 49(22), 13190-13198. https://doi.org/10.1021/acs.est.5b02426

Li, Y. and Jiang, D. (2014). Dynamic task-scheduling based parallel processing on watershed distributed eco-hydrological model. Applied Mechanics and Materials, 556-562, 3492-3495. https://doi.org/10.4028/www.scientific.net/amm.556-562.3492

Maringanti, C., Chaubey, I., & Popp, J. (2009). Development of a multiobjective optimization tool for the selection and placement of best management practices for nonpoint source pollution control. Water Resources Research, 45(6). https://doi.org/10.1029/2008wr007094

Mazur, M., Smith, B., Bird, B., McMillan, S., Pyron, M., & Hauswald, C. (2021). Hydrologic connectivity and land cover affect floodplain lake water quality, fish abundance, and fish diversity in floodplain lakes of the wabash–white river basin. River Research and Applications, 38(1), 160-172. https://doi.org/10.1002/rra.3888

Moore, R., Johnston, C., Smith, R., & Milstead, B. (2011). Source and delivery of nutrients to receiving waters in the northeastern and mid-atlantic regions of the united states1. Jawra Journal of the American Water Resources Association, 47(5), 965-990. https://doi.org/10.1111/j.1752-1688.2011.00582.x

Moriasi, D.N., Arnold, J.G., Van Liew, M.W., Bingner, R.L., Harmel, R.D., & Veith, T.L. (2007). Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transactions of the ASABE, 50(3), 885-900.

Muleta, M. and Nicklow, J. (2004). Joint application of artificial neural networks and evolutionary algorithms to watershed management. Water Resources Management, 18(5), 459-482. https://doi.org/10.1023/b:warm.0000049140.64059.d1

Natho, S., Tschikof, M., Bondar-Kunze, E., & Hein, T. (2020). Modeling the effect of enhanced lateral connectivity on nutrient retention capacity in large river floodplains: how much connected floodplain do we need?. Frontiers in Environmental Science, 8. https://doi.org/10.3389/fenvs.2020.00074

Neumann, A., Dong, F., Shimoda, Y., Arnillas, C., Javed, A., Yang, C., … & Arhonditsis, G. (2021). A review of the current state of process-based and data-driven modelling: guidelines for lake erie managers and watershed modellers. Environmental Reviews, 29(4), 443-490. https://doi.org/10.1139/er-2020-0070

Noe, G., Hopkins, K., Claggett, P., Schenk, E., Metes, M., Ahmed, L., … & Hupp, C. (2022). Streambank and floodplain geomorphic change and contribution to watershed material budgets. Environmental Research Letters, 17(6), 064015. https://doi.org/10.1088/1748-9326/ac6e47

Pruden, A., Pei, R., Storteboom, H., & Carlson, K. (2006). Antibiotic resistance genes as emerging contaminants: studies in northern colorado. Environmental Science & Technology, 40(23), 7445-7450. https://doi.org/10.1021/es0604131

Son, J., Cowley, C., Goodwin, S., Arabi, M., & Carlson, K. (2013). Relative phosphorus load inputs from wastewater treatment plants in a northern colorado watershed. Journal of Environmental Quality, 42(2), 497-506. https://doi.org/10.2134/jeq2012.0349

Storteboom, H., Arabi, M., Davis, J., Crimi, B., & Pruden, A. (2010). Identification of antibiotic-resistance-gene molecular signatures suitable as tracers of pristine river, urban, and agricultural sources. Environmental Science & Technology, 44(6), 1947-1953. https://doi.org/10.1021/es902893f

White, K. and Chaubey, I. (2005). Sensitivity analysis, calibration, and validations for a multisite and multivariable swat model. Jawra Journal of the American Water Resources Association, 41(5), 1077-1089. https://doi.org/10.1111/j.1752-1688.2005.tb03786.x

Wollheim, W., Peterson, B., Thomas, S., Hopkinson, C., & Vörösmarty, C. (2008). Dynamics of n removal over annual time periods in a suburban river network. Journal of Geophysical Research: Biogeosciences, 113(G3). https://doi.org/10.1029/2007jg000660

# CHAPTER 6 SYNTHESIS, IMPLICATIONS, AND FUTURE DIRECTIONS

## 6.1 OVERVIEW: THE HYBRID INTELLIGENCE PARADIGM

Environmental modeling confronts an enduring tension between computational efficiency required for operational decision support and physical consistency demanded by scientific and regulatory applications. Traditional process-based models provide mechanistic rigor but require computational resources that constrain scenario analysis and uncertainty quantification. Pure machine learning approaches offer computational speed but frequently violate conservation principles, undermining stakeholder trust and regulatory acceptance. This dissertation establishes that strategic integration of multiple artificial intelligence techniques with physically based principles can resolve this tension, achieving both computational tractability and scientific validity.

The framework developed here demonstrates feasibility of comprehensive watershed nutrient assessment under multiple future scenarios through selective deployment of modeling approaches matched to problem characteristics: ensemble methods where uncertainty dominates, sequence learning for spatiotemporal forecasting, physics-constrained optimization where mass balance is critical, and streamlined routing where computational efficiency enables extensive analysis. Rather than seeking universal solutions, the research establishes design principles for hybrid systems that preserve physical consistency while enabling operational application at regional scales.

## 6.2 SYNTHESIS OF METHODOLOGICAL CONTRIBUTIONS

### 6.2.1 UNCERTAINTY QUANTIFICATION THROUGH ENSEMBLE LEARNING

Ensemble machine learning applied to natural landscape nutrient exports (Chapter 2) revealed that Bayesian Model Averaging provides superior predictive stability compared to individual algorithms. The coefficient of variation decreased from 29.2% for Random Forest to 15.7% for the ensemble, while maintaining cross-validation performance ($R^2$ = 0.70-0.83 for nitrogen and phosphorus). This improvement derives from probabilistic weighting across model structures rather than variance within single algorithms, suggesting that environmental prediction benefits from explicit acknowledgment of model uncertainty.

The finding extends beyond specific performance metrics to establish that ensemble approaches address a fundamental challenge in data-driven environmental modeling: no single algorithm consistently excels across diverse hydrologic conditions. Bayesian Model Averaging adaptively weights models according to evidence provided by data, effectively learning which algorithmic structures prove most reliable for particular landscape-climate combinations. This adaptive capability proves particularly valuable for regional applications where model deployment across heterogeneous watersheds requires robust performance despite varying predictive task characteristics.

### 6.2.2 TEMPORAL LEARNING IN DATA-RICH CONTEXTS

The Convolutional Long Short-Term Memory framework for urban land cover change (Chapter 3) uncovered a counterintuitive pattern: prediction accuracy increased from F1 = 0.27 for 1-year forecasts to F1 = 0.87 for 20-year forecasts when trained on 39 years of annual observations. This

"Temporal Depth Paradox" challenges established assumptions that prediction uncertainty increases monotonically with temporal horizon, suggesting instead that appropriately designed sequence learning architectures distinguish persistent signals from short-term fluctuations.

The finding carries important implications for model selection in spatiotemporal forecasting. Comparative analysis with cellular automata studies demonstrates that architectural complexity should match data richness: simple transition probability models excel with sparse temporal data, while deep learning approaches leverage dense multi-decadal sequences to learn complex patterns inaccessible to static formulations. The 39-year training dataset enabled the network to identify decadal cycles and path dependencies that shorter time series cannot support, indicating that temporal data density represents a critical consideration in choosing between modeling paradigms.

### 6.2.3 PHYSICS-GUIDED REGULARIZATION IN COMPLEX ENVIRONMENTAL SYSTEMS

The physics-informed neural network methodology for agricultural nutrient prediction (Chapter 4) demonstrated that domain knowledge integration can enhance rather than degrade predictive performance when training data contain substantial noise. For total phosphorus, Nash-Sutcliffe Efficiency improved from 0.845 to 0.850 despite mass balance constraints, while total nitrogen violations decreased by 95% (from 2.62% to 0.18%) with only marginal accuracy reduction (NSE: 0.87 to 0.86).

This asymmetric response across nutrients suggests that the accuracy-consistency trade-off often assumed inherent to physics-informed learning is context-dependent. Phosphorus dynamics, strongly controlled by sediment-associated processes, benefited from constraints linking particulate transport to erosion rates, effectively preventing the model from learning spurious

correlations between unrelated variables. Nitrogen dynamics, involving both particulate and dissolved fractions with distinct transport mechanisms, exhibited slight statistical performance degradation as constraints prevented exploitation of all available patterns. The three-phase progressive training strategy proved critical for achieving this balance, allowing fundamental pattern establishment before imposing stringent physical requirements.

### 6.2.4 COMPUTATIONAL EFFICIENCY THROUGH STRATEGIC DECOUPLING

The SWIFT routing framework (Chapter 5) achieved Nash-Sutcliffe Efficiency exceeding 0.90 across 200 monitoring stations while reducing computational time by approximately 1,900-fold relative to traditional SWAT+ applications. This performance demonstrates that simplified mass balance algorithms capture essential transport and transformation processes when operating on high-quality hydrologic inputs, validating hybrid approaches that leverage detailed process representation where justified (land-phase hydrology) and streamlined algorithms where efficiency is prioritized (in-stream routing).

The computational gain transformed comprehensive Sobol sensitivity analysis from theoretical possibility to practical reality, with 2,048-member ensembles completing in 2.2 hours compared to 4,000+ hours required by process-based approaches. This efficiency enabled activities previously constrained by computational limitations, including systematic parameter uncertainty quantification and source-specific delivery ratio calculations across multiple scenarios. The speedup derives from strategic separation of concerns: accepting pre-calibrated flow outputs avoids repeated computation of land-phase processes while maintaining physically based constituent routing through explicit mass balance accounting.

## 6.3 IMPLICATIONS FOR DECISION SUPPORT

### 6.3.1 OPERATIONAL FEASIBILITY AT REGIONAL SCALES

The integrated framework transforms computational requirements from weeks to hours for state-wide nutrient predictions under multiple scenarios, establishing operational feasibility for comprehensive watershed assessment. This advance addresses a critical barrier to machine learning adoption in environmental management: the computational burden of traditional models constrains exploration of management alternatives, limiting support for adaptive planning processes requiring evaluation of numerous interventions.

Physical consistency maintenance occurs through multiple mechanisms operating at different framework stages. Conservation constraints embedded in urban growth predictions ensure development respects protected areas through probability refinement during spatial allocation (Chapter 3). Physics-informed neural networks explicitly represent biogeochemical processes within network architecture and enforce mass balance through loss function penalties (Chapter 4). Streamlined routing maintains rigorous mass balance through object-oriented design with explicit state variables (Chapter 5). This multi-level approach distinguishes the framework from pure machine learning surrogates sacrificing scientific validity for speed, and from process-based models sacrificing computational efficiency for mechanistic detail.

### 6.3.2 STRATEGIC PLANNING UNDER DEEP UNCERTAINTY

The Temporal Depth Paradox finding (Chapter 3) carries implications for long-term planning applications. The observation that 20-year predictions substantially outperform 1-year forecasts suggests that models trained on dense historical sequences may prove more reliable for strategic

planning horizons (2030-2050) than for near-term operational forecasting. This counterintuitive result reflects fundamental differences in predictive task characteristics: short-term forecasts require capturing stochastic year-to-year variations, while long-term projections benefit from identifying persistent development trajectories and decadal trends.

For metropolitan planning organizations developing 20-30 year comprehensive plans, this finding suggests that AI-enhanced forecasting may provide greater utility for strategic visioning than previously assumed, provided training data span multiple decades. The framework's integration of conservation constraints and transit-oriented development policies demonstrates how alternative planning interventions can be systematically evaluated, quantifying their spatial manifestation and infrastructure implications under different demographic scenarios. Infrastructure efficiency analysis revealed that spatial configuration affects resource requirements independent of total population, with consolidated development patterns (Scenario D) achieving 20.5% efficiency gains compared to dispersed growth (Scenario B: 16.3% gains) despite similar population increases.

### 6.3.3 SOURCE-SPECIFIC MANAGEMENT TARGETING

Load substitution experiments isolating contributions from individual nutrient sources (Chapter 5) quantified systematic differences in delivery ratios: wastewater treatment facilities (0.80), municipal stormwater (0.77-0.80), and agricultural lands (0.74). These differences reflect distinct transport pathways and retention opportunities, informing cost-effectiveness analysis by accounting for source-specific transport efficiency. Wastewater effluent enters channels directly, experiencing only in-stream processes, while agricultural nutrients traverse overland flow paths, soil profiles, and riparian zones providing multiple retention opportunities.

The spatial variability in agricultural delivery ratios (interquartile range: 0.68-0.85) emphasizes the importance of basin-specific assessment rather than uniform regulatory approaches. Delivery efficiency depends strongly on local factors including slope, drainage density, soil hydraulic properties, and riparian buffer extent. The framework's computational efficiency enables systematic evaluation of these spatial patterns, supporting optimization of management practice placement to maximize downstream water quality benefits per unit investment.

## 6.4 LIMITATIONS AND SOURCES OF UNCERTAINTY

Spatial transferability analysis demonstrated expected performance degradation when predicting ungauged watersheds, with accuracy reductions of 10-20% typical when models encounter watershed characteristics outside training distributions (Chapters 2, 5). This pattern reflects fundamental limitations in data-driven approaches: learned relationships interpolate within sampled conditions but extrapolate poorly to novel combinations of physiographic and climatic characteristics. The heterogeneity in transferability suggests some environmental processes exhibit more consistent spatial relationships than others, emphasizing need for region-specific training rather than universal model deployment.

The integration of multiple modeling approaches introduces complexity in uncertainty propagation across framework components. Each stage (ensemble ML for natural landscapes, ConvLSTM for urban land cover, physics-informed ML for agriculture, SWIFT for routing) contributes uncertainty to final predictions, but comprehensive quantification of compound uncertainty through the modeling chain remains incomplete. Standard error propagation methods may underestimate total uncertainty when component errors exhibit correlation or when nonlinear transformations occur between stages. Applications requiring formal uncertainty quantification for

regulatory purposes should employ ensemble approaches or Monte Carlo methods characterizing prediction intervals that account for multi-stage error accumulation.

The physics constraints implemented in agricultural prediction (Chapter 4), while substantially improving mass balance adherence, employed simplified formulations including fixed mineralization rates (3% annual) and static Curve Number hydrology. More sophisticated mechanistic representations could enhance physical realism. The demonstrated performance (NSE > 0.85) suggests simplified constraints capture essential processes for management applications despite not fully representing biogeochemical complexity.

## 6.5 FUTURE RESEARCH DIRECTIONS

### 6.5.1 METHODOLOGICAL EXTENSIONS

Physics-informed neural network architecture employed static constraint formulations; future research should explore adaptive constraints that strengthen or relax based on prediction confidence or data density. Meta-learning approaches could identify optimal constraint weights for different watershed types, automating the balance between data fidelity and physical consistency currently requiring expert judgment. Extension to time-varying constraints would enable representation of seasonal biogeochemical dynamics not captured by annual mass balance requirements.

Future research should integrate uncertainty quantification methods -such as Monte Carlo dropout, deep ensembles, or Bayesian neural network layers- into the physics-informed architecture. This would enable probabilistic predictions that communicate reliability to decision-makers, identify conditions where predictions are less certain, and support risk-based management decisions.

The ConvLSTM framework could be extended with attention mechanisms that provide interpretable identification of which historical periods most influence predictions. Transformer architectures, increasingly successful in sequence modeling, may offer advantages for capturing long-range temporal dependencies while providing attention weights that illuminate learned relationships between past patterns and future outcomes.

### 6.5.2 PROCESS UNDERSTANDING

The Temporal Depth Paradox warrants systematic investigation across urban contexts with varying development histories and planning regimes. Comparative analysis across cities with different regulatory environments could determine whether improved long-term prediction reflects universal urban dynamics or region-specific patterns.

The differential response of nitrogen and phosphorus to physics constraints suggests that optimal formulations may be constituent specific. Systematic experiments varying constraint type and stringency across nutrients could identify which biogeochemical processes are well-captured by statistical learning versus requiring explicit physical representation. Extension to additional constituents (sediment, pathogens, emerging contaminants) would test generalizability of the physics-informed approach.

### 6.5.3 OPERATIONAL APPLICATIONS

Real-time data assimilation represents an important frontier enabled by computational efficiency. The seconds-per-simulation execution times suggest that ensemble Kalman filtering or particle filtering approaches could update predictions as new observations become available. Development of operational forecasting systems for seasonal nutrient load prediction, event-based water quality

alerts, or adaptive management response would extend the framework from planning support to real-time decision support.

Integration with economic optimization would enable cost-effectiveness analysis for nutrient management. Combining delivery ratio estimates with practice implementation costs could identify optimal spatial allocation of management investments. Multi-objective optimization balancing water quality improvement against implementation cost, landowner participation, and co-benefits (habitat, carbon) would provide comprehensive decision support for watershed managers.

### 6.5.4 BROADER APPLICATIONS

The hybrid intelligence design principles (strategic deployment of approaches matched to problem characteristics, physics-informed regularization, ensemble uncertainty quantification) apply beyond water quality to environmental prediction broadly. Testing framework transferability for air quality forecasting, ecosystem service valuation, or climate impact assessment would establish generalizability. Development of design guidelines for hybrid environmental intelligence systems could accelerate adoption across application domains facing similar efficiency-validity trade-offs.

The multi-scenario framework could support environmental justice applications by enabling analysis of how alternative development patterns affect environmental burden distribution across communities. Integration with demographic data would enable identification of scenarios minimizing disproportionate impacts on vulnerable populations, extending decision support beyond aggregate efficiency to equity considerations.

## 6.6 CONCLUSION

This dissertation demonstrates the feasibility of a hybrid intelligence framework that bridges computational efficiency and scientific validity in watershed nutrient modeling. By strategically integrating diverse machine learning techniques with physical principles, the research resolves the tension between the speed required for operational decision-making and the rigorous consistency demanded by scientific applications.

The findings challenge the traditional dichotomy between data-driven and process-based modeling. Results indicate that domain knowledge integration enhances predictive performance while ensuring defensibility, proving that physical validity need not be a constraint on statistical optimization but rather a guide for it. This approach transforms comprehensive, scenario-based watershed assessment from a theoretical ideal into a computationally tractable reality, reducing analysis time from weeks to hours without compromising calibration performance.

Ultimately, this research establishes that artificial intelligence supports rather than replaces expert judgment in environmental management. The hybrid intelligence paradigm, combining machine learning's pattern recognition with physics-based consistency and human-defined objectives, provides a foundation for the next generation of decision support systems: efficient, scientifically defensible, and aligned with the complex realities of sustainable watershed management.