ARTIFICIAL INTELLIGENCE FOR FORECASTING GROUNDWATER DEPLETION IN BANGLADESH FOR SUSTAINABLE GROUNDWATER MANAGEMENT: A DATA-DRIVEN APPROACH

by

Ibtida Ahmed

August 2026

A thesis submitted to the
Faculty of the Graduate School of
the University at Buffalo, State University of New York
in partial fulfillment of the requirements for the
degree of

MS in Computer Science and Engineering

School of Engineering and Applied Sciences

Acknowledgments

I would like to give my thanks to my committee for their guidance throughout my work, especially to Professor Christopher Lowry of the Earth Sciences Department for providing me hydrogeological insights that I otherwise would have overlooked, and to also sincerely give my thanks to my primary advisor, Professor Varun Chandola, for helping me stay on track for a timely graduation over what I consider to be an otherwise compressed timeline due to circumstances beyond my own control. Even so, I would also like to thank PHD Student Yuting Hu's efforts to help me stay on track with my work during the spring semester of 2026.

I would also like to sincerely give my thanks to the Bangladesh Water Development Board for providing me the rest of their groundwater well data for up to 2020. The rest of this study would not have been possible without their contributions, which I greatly appreciate.

This has been an especially stressful period working on what I consider to be my most complex project to date, and I couldn't have been happier to finally get it out.

Abstract

Groundwater is a vital resource in upholding agriculture and food security, as well as drinking water supply in Bangladesh. However, decades of intensive dry-season irrigation have driven substantial depletion, particularly in the Barind and Madhupur Tracts, and in the Dhaka Metropolitan Area. Efforts to forecast and mitigate this depletion remain limited in both spatial resolution and predictive skill, constraining the country's capacity to anticipate—rather than to just forecast—groundwater depletion. This thesis develops and benchmarks three spatiotemporal deep learning architectures—a convolutional-recurrent hybrid (CNN-LSTM), a spatiotemporal deep graph neural network (ST-GNN), and a spatiotemporal graph transformer network (ST-GTN)—to monitor fluctuations in seasonal water table depth across Bangladesh at a high spatial resolution. These models were trained on a multi-source covariate dataset spanning from 2000 to 2020, integrating Bangladesh Water Development Board (BWDB) well observations with climate, MODIS NDVI, land-use, and population layers, and evaluated across validation (2017-2018), test (2019), and forecast (2020) periods, with quantile-mapping applied to correct residual distribution bias. The ST-GTN model achieved the strongest and most consistent performance, with a Nash-Sutcliffe efficiency reaching 0.956 and a validation RMSE as low as 0.288 m, outperforming both the ST-GNN and CNN-LSTM baselines after converging in fewer training epochs. Cross-model feature importance analysis—integrating and combining both permutation importance and gradient-input attribution—shows that land use and temperature-related covariates, rather than precipitation and vegetation indices as initially expected, are the two variables that most consistently explain forecast accuracy, a pattern attributed to vadose-zone lag and the dominance of anthropogenic extraction

over climatic recharge in irrigation-intensive regions. All models, however, have systematically underestimated peak drawdown in the Barind Tract and furthermore exposed deteriorated predicting skill during the 2020 monsoon period, which we attribute to missing abstraction data and limited spatial resolution as the primary constraints on model forecast quality. These findings position graph-transformer architectures as a promising prospect for groundwater forecasting in otherwise data-scarce, monsoon-driven deltaic systems, with a concrete pathway outlined that would incorporate irrigation-extraction covariates, finer spatial resolution, and calibrated uncertainty. This pathway would contribute toward an operational AI decision-support tool for the Bangladesh Water Development Board in support of the Bangladesh Delta Plan 2100.

Table of Contents

List of Tables

List of Figures

Introduction

Background and Motivation

Groundwater is a crucial resource for human survival and economic growth, making up about 30% of the world's freshwater supply [1]. A resource generally more stable than surface water, it provides a reliable source of clean water that persists especially during droughts [86]. However, numerous factors such as population growth, urbanization, and poor management contribute to significant groundwater depletion, thus affecting water security and food production [61, 66]. Around 90% of the world's freshwater is used for irrigation, with groundwater supporting up to 40% of agricultural output [66]. Consequently, aquifer levels are rapidly declining in critical farming areas—including, though not limited to California's Central Valley and much of South Asia. Areas where groundwater is essential for agriculture are most vulnerable to depletion [59, 67].

In Bangladesh, groundwater is crucial for both agriculture and domestic water needs. 80% of irrigation water is supplied by Groundwater, especially during the dry boro-rice season, during the months of December and January where surface water is scarce [56]. Between 1991 and 2013, rice production in Bangladesh nearly tripled, primarily from increased groundwater irrigation [41]. However, excessive pumping has caused significant drops in groundwater levels, particularly in the Barind and Madhupur regions [68]. Information from NASA's GRACE satellite show that Bangladesh has lost about 37.5 billion cubic meters of water from 2002 to 2020, with irrigation being a primary factor [20]. In Dhaka city, groundwater levels have fallen by more than 60 meters in the last 50 years. If current patterns continue, predictions indicate a potential drop of up to 161 meters by 2050. [47].

Problem Statement

Groundwater depletion in Bangladesh presents several interconnected issues such as lower recharge rates, a decline in water quality (such as arsenic contamination), increasing pumping costs, and limited access to water for vulnerable communities [47, 56]. The Bangladesh Delta Plan 2100 identifies groundwater depletion as a significant threat to the country's food and water security, but it lacks effective tools for real-time monitoring and decision-making [20]. Traditional models like ARIMA and linear regression often fail to accurately predict groundwater levels due to their inability to address the complex nature of groundwater systems [10, 76]. Thus, there is a pressing need for AI-driven approaches that can utilize various data sources, especially geological information, to better forecast groundwater levels and support sustainable management.

Artificial intelligence (AI) has great potential to improve groundwater forecasting and management through techniques like machine learning (ML) and deep learning (DL). AI models—including Long Short-Term Memory (LSTM) networks, Convolutional Neural Networks (CNNs), and Graph Neural Networks (GNNs)—have shown to effectively analyze complex relationships between climate factors (like rainfall and evaporation), human activities (such as irrigation and urban water use), and groundwater levels [4, 45]. Hybrid models-like CNN-LSTM and Spatio-Temporal Graph Neural Networks (ST-GNNs)-have demonstrated better accuracy in predicting groundwater levels in diverse aquifers [79]. Additionally, using remote sensing data from sources like NASA's GRACE and InSAR improves the accuracy of AI models by offering insights into large-scale water storage and ground movement [77].

Incorporating AI to manage groundwater can greatly improve agriculture and food security in Bangladesh, as it can help us see how weather and landuse practices affect water supply. It can help guide farmers to use water more efficiently and still grow quality crop yields. In addition, remote sensing technology provides important data for monitoring and managing groundwater resources.

These tools can change how we manage groundwater by offering real-time forecasts to contribute to data-driven decision-making, and especially offer early warnings about areas running low on water. However, challenges still exist when utilizing these tools. The groundwater monitoring network in the country is limited and inconsistent, access to data is often restricted, and the ability of institutions to implement AI tools is still developing [70]. To overcome these problems, we need efficient, easy to comprehend AI systems that can work with otherwise limited data. These systems should help create practical policies for sustainable farming and food security in Bangladesh.

This research, titled ”Artificial Intelligence for Forecasting Groundwater Depletion in Bangladesh for Sustainable Groundwater Management: A Data-Driven Approach,” aims to address these gaps by developing and comparing advanced AI-based spatiotemporal models to predict groundwater dynamics across Bangladesh's diverse hydrogeological settings.

Research Questions

1. How do advanced spatiotemporal AI models (CNN-LSTM, ST-GNN, ST-GTN) compare when forecasting groundwater depth across Bangladesh's aquifer systems?

2. How much do location and environmental covariates (temperature, rainfall, land use, vegetation) contribute to forecast accuracy?

3. How robust are the models to the realities of Bangladesh's monitoring data—gaps, intermittent records, and a single-year forecast horizon?

4. Can combining AI with satellite remote sensing improve the basis for early warning of depletion-prone areas?

5. What institutional and technical constraints condition the operational use of such tools?

Research Objectives

- We will study groundwater depletion in Bangladesh using various types of data, including geological, climate, and satellite information.
- Our goal is to create AI models that can predict groundwater levels over time and in different seasons. The models we will develop are:
  - Convolutional Neural Network and Long-Short-Term Memory Hybrid (CNN-LSTM)
  - Spatio-Temporal Graph Neural Network (ST-GNN)
  - Spatio-Temporal Graph Transformer Network (ST-GTN)
- We will compare how well these models perform against a standard CNN-LSTM model under various data conditions.
- We aim to identify the main factors that cause groundwater depletion and provide insights for policymakers and water resource managers.

Outputs and Applications

1. We have developed high-resolution models to forecast groundwater levels and predict depletion trends in Bangladesh's major aquifer areas.
2. An AI and satellite-based tool helps water authorities locate depletion hotspots and plan irrigation schedules effectively.
3. We provide insights that support the Bangladesh Water Development Board (BWDB) and the Ministry of Water Resources in carrying out the Bangladesh Delta Plan 2100.
4. Our work contributes to the field of hydroinformatics by applying AI, especially in areas lacking data and facing climate challenges.

Thesis Organization

The structure of this Thesis is organized as follows. Chapter 2 goes over past literature on global and local Groundwater depletion within Bangladesh, documenting groundwater forecasting development by first highlighting traditional statistical models and how development transitions to machine learning and spatiotemporal deep learning. Furthermore chapter 2 acknowledges the research gaps identified in this study. Chapter 3 emphasizes the study area and the multi-source dataset that we've used, consisting of Bangladesh Water Development Board (BWDB) water table observations, climate data, MODIS NDVI data, land-use, and population covariates, all spanning the time frame of 2000 to 2020. This dataset is then run through a preprocessing pipeline for standardization into a common spatiotemporal grid. Chapter 4 identifies and formalizes the forecasting problem, and introduces the three model architectures developed and used (CNN-LSTM baseline, ST-GNN, and ST-GTN), as well as the experimental design, the ground-truth metrics used for model evaluation, and the quantile-mapping bias-correction procedure. Chapter 5 reports Bangladesh's groundwater dynamics from model performance, analyzing model training, validation, test performance, forecast-year evaluation, and the bias-corrected predictions made. Chapter 6 interprets these results, examining the superior performance of the ST-GTN model, the roles of climatic and anthropogenic drivers, the persistent errors found in predicting groundwater dynamics within the Barind Tract and during the 2020 monsoon season, as well as highlighting the study's limitations. In turn, consequences for groundwater management under the Bangladesh Delta Plan 2100 are also discussed. Chapter 7 concludes by acknowledging each research question while summarizing the thesis's contributions, before outlining directions for future research under this discipline.

Literature Review

Groundwater depletion is a significant issue that continues to threaten water security and agriculture globally. As the largest accessible source of freshwater, Groundwater is crucial for billions of people and global food production, especially in drought-prone areas [61, 86]. However, rapid population growth, urbanization, and heavy irrigation have also contributed to severe declines in groundwater levels, particularly in regions throughout South Asia [59, 66]. In Bangladesh, where almost 80% of irrigation relies on groundwater, excessive extraction has resulted in serious depletion, especially in the Barind Tract and Dhaka city, threatening food security and access to clean water resources [56, 70]. To tackle this issue, advanced data-driven solutions, including artificial intelligence, have been proposed and utilized to help forecast groundwater changes and promote sustainable management.

Groundwater Depletion: Global Context

Freshwater is an essential reservoir for clean drinking water and for agriculture, which is key for global food production [1]. Agriculture uses about 70–80% of freshwater resources and supports roughly 40% of global food production [6]. Groundwater in particular makes up about 30% of the world's freshwater supply and is the largest reservoir, far more resilient short-term climate changes and seasonal droughts compared to surface water, which makes it a more reliable resource especially across arid regions [86]. Despite technological advances in agriculture, freshwater has remained crucial to uphold global food security [61]. Ultimately, groundwater is vital for addressing surface water shortages and its role in food production cannot be overstated [61].

The global population has increased drastically, from under 1.7 billion in the early 1900s to about 7.2 billion by 2014. It is projected to reach 8.2 billion by 2025 and 9.5 billion by 2050 [76, 86]. This population growth has led to a surge in water demand, with global water withdrawals rising from 500 cubic kilometers in 1900 to around 3,000 cubic kilometers by 2000 [86]. By 2010, annual groundwater withdrawals were about 950 cubic kilometers and eventually rose to over 4,000 km^3 by 2020 [14, 86], largely for agricultural irrigation to meet food needs [66]. Irrigation practices have consumed roughly 90% of freshwater, supported 40% of food production, and has occupied 20% of cropland [66]. In turn, around 30% of total freshwater use comes from groundwater withdrawals, exceeding 1,000 km^3 annually, mainly for agricultural irrigation, which sustains the global food supply [59, 66, 74]. This rising demand places considerable pressure on water-scarce areas, thus risking groundwater depletion. Looking forward, population growth, urbanization, and climate change are expected to increase water stress, especially in arid and semi-arid regions. The overeliance on groundwater in these areas could lead to aquifer depletion and long-term water and food security issues [14, 19, 21].

Intensive groundwater irrigation has profound consequences, particularly in arid and semi-arid regions. In the North China Plain, extensive irrigation for crops such as wheat and maize has significantly lowered groundwater levels. Similarly, California's Central Valley, a key agricultural hub, faces severe challenges from excessive groundwater extraction [59, 66]. In the United States as a whole, approximately 60% of irrigation depends on groundwater, with California's Central Valley accounting for nearly half of the nation's groundwater depletion since 1900 [66]. In South Asia, countries like India and Pakistan depend heavily on groundwater for irrigation due to limited rainfall, leading to significant depletion of these resources [67]. For Bangladesh in particular, groundwater is crucial for growing rice, especially boro rice in the dry season, with about 80% of the country's groundwater used for irrigation. As the population grows—from 145 million in 2008 to an expected 182 million by 2030—rice demand is projected to rise to 39 million tons, putting more pressure on groundwater resources and threatening sustainability [55].

Over-extraction of groundwater further threatens food security by disrupting natural aquifer recharge. As water table boundaries plummet, access to freshwater becomes more precarious [16]. This can lead to issues like land subsidence and seawater intrusion, especially in dry areas dependent on limited groundwater [67]. Vulnerable communities are particularly affected, facing increased water scarcity and costs due to both physical and economic barriers [61]. Over time, declining water levels can leave urban poor populations without reliable water sources, worsening public health, gender equity, social cohesion, and economic stability [47].

While Groundwater supports billions in food production, unsustainable extraction and contamination of the resource pose serious long-term risks. To avoid irreversible damage to aquifers, immediate and coordinated action is essential, including science-based management, fair water policies, and sustainable agricultural practices. Protecting groundwater is critical not only for food security but also for environmental integrity and social equity, and it should be prioritized alongside climate and biodiversity efforts.

Groundwater Depletion in Bangladesh

Bangladesh is home to a vast network of over 230 rivers, including 57 that cross into India and Myanmar [12]. It lies in the lower region of the Ganges-Brahmaputra-Meghna (GBM) river system, which covers 1.72 million square kilometers, Bangladesh occupying about 8% of this area [73]. These rivers are essential for agriculture, transportation, and local industries [8]. Within Bangladesh, the river system extends over 22,000 kilometers and provides an estimated 1,211 km^3 of renewable water each year, including 21 km^3 from groundwater that is partly replenished by surface water [56]. However, surface water irrigation is still not widely used due to historical neglect and inadequate infrastructure [55].

Groundwater is a fundamental resource in Bangladesh's agricultural economy, particularly for rice farming, which occupies 75% of the arable land and produces 90% of the country's food grains [56]. Rice is cultivated in three seasons: aus (pre-monsoon), aman (monsoon), and boro (dry season), boro rice yielding the most [41]. From 1991 to 2013, rice production increased from

6.8 to 18.8 million tons, mainly due to the growth of boro rice [56]. Boro rice, primarily grown in the dry season, constitutes over 70% of total crop output and relies on groundwater for 80% of its irrigation [56]. By 2022, the average rice yield in Bangladesh was 4.8 tons per hectare, surpassing that of India and Thailand [20].

During the dry season, around 5 million hectares are irrigated, 73% of the water sourced from groundwater [56]. Since 2002, there has been a loss of 37.5 billion cubic meters of terrestrial water due to irrigation, as shown by NASA's GRACE satellite data [20]. As a result of this depletion, farmers are now extracting water from depths that are 20 meters deeper than they were two decades ago [20]. Rainfall has decreased by 10% over the last 20 years, and winter river flows in the Ganges-Brahmaputra-Meghna (GBM) system have been cut in half from 1993 to 2021 due to less rainfall and increased groundwater pumping in India [20]. The Barind and Madhupur Tracts are particularly impacted, with the Barind Tract facing the highest groundwater withdrawals [68]. Additionally, rising sea levels, at 5 mm per year, have resulted in the loss of 490 km^2 of coastal land since 2001 due to subsidence and other causes [20].

By 2030, the population is expected to hit 182 million, further fueling groundwater demand. To satisfy this population demand, the supply of rice needs to reach 39 million tons, and wheat and maize will also see substantial increases—124% more than in 2008 [56]. This rising demand for water is putting further pressure on aquifer recharge, making it harder for more impoverished communities to access water especially in Dhaka, where many tubewells are failing [47]. Dhaka city—with a population of 36.6 million as of 2025—heavily depends on groundwater, with the water authority extracting 5.9 million m^3/day, including 2.75 million m^3 from 920 deep tubewells. Over the last 50 years, groundwater levels have dropped by 60 meters, and projections indicate they could decline by up to 161 meters by 2050 [47].

The Bangladesh Delta Plan 2100 is a long-term strategy aimed at tackling water issues like groundwater depletion and contamination. The plan suggests ways to reduce arsenic contamination—such as by promoting deeper wells, surface water piping, and using alternative water sources [20]—but it still lacks clear implementation plans [20]. By August 2022, real-time water quality monitoring systems were set up in 78 out of Dhaka's 145 districts, helping cut water losses from 40% to 15% through demand-based rationing [20]. The total cost of the plan is estimated at $38 billion, with $1.8 billion already pledged by the World Bank for infrastructure [20].

Access to hydrological data—like streamflow, groundwater levels, rainfall, and water quality—remains limited in Bangladesh. However, satellite technologies are helping to bridge these gaps. For instance, satellite-based irrigation systems have been effective in improving water use and decreasing groundwater extraction in northern India and eastern Pakistan [20]. In Bangladesh, large-scale modeling shows that using shallow groundwater for dry-season rice can help recharge aquifers, especially when rainfall is stable or slightly reduced [70]. Data from 465 sites in the Bengal Basin show increased groundwater recharge from 1965 to 2017, especially between 1976–1980 and 2011–2015 [70]. However, monsoon recharge often doesn't fully compensate for groundwater use in dry seasons, particularly in low-permeability areas like the Barind Tract and Ganges floodplain, where groundwater depth can exceed 8 meters [70]. These areas might need to rely more on the Brahmaputra floodplains, which have better recharge potential.

Through real-time monitoring data and predictive models, AI-driven groundwater forecasting can greatly improve water management, helping to predict water demand needs and identifying the underlying depletion risks to develop and support sustainable extraction needs. This approach can effectively balance groundwater use with natural recharge, ensuring long-term resilience amid climate change and population growth. By integrating real-time monitoring systems with AI tools, we can reduce water losses and improve quality control, especially for arsenic contamination. By investing in strong hydrological models and expanded monitoring networks via AI technologies, Bangladesh will be able to balance groundwater use with recharge and ensure resilience against climate change and population growth to meet the needs of 182 million people by 2030.

Groundwater Level Forecasting Methods

Estimating groundwater storage and forecasting groundwater levels (GWL) is crucial for sustainable water management, especially in areas facing rising water demands while under the pressure of population growth and climate variability [1]. As surface water sources decline due to droughts and overuse, groundwater has become vital in supplying nearly half of the world's drinking water and about 40% of irrigation needs [43, 82]. Yet, forecasting groundwater levels has remained challenging due to the complex interactions between hydrogeological factors, climate drivers, and human activities [52, 65, 80]. Factors like river–aquifer exchanges and geological differences can complicate traditional models.

Monitoring anomalies in groundwater levels is important for early warning systems, as these deviations can indicate droughts, overuse, or saltwater intrusion, particularly in coastal areas [19, 80, 87]. This makes identifying vulnerable areas a priority for policymakers. Modern forecasting combines machine learning models like LSTM and Random Forests, physical simulations such as MODFLOW, and remote sensing data (including GRACE satellite observations), improving accuracy and detail [38, 44, 77]. These tools aid in better water allocation, irrigation planning, and managed aquifer recharge (MAR), essential components for climate-resilient water management.

Improving groundwater forecasting is not only a scientific goal but also a socioeconomic necessity. Reliable prediction tools based on interdisciplinary data are vital for securing water resources in a changing climate. AI-driven methods can enhance forecasting by integrating real-time data and identifying risks, ultimately supporting sustainable development in vulnerable regions.

Traditional Statistical Methods

Forecasting groundwater levels is crucial for managing water sustainably, especially with tackling climate change, population growth, and agricultural expansion. Traditional statistical methods, such as linear regression, moving averages, ARIMA models, and the Mann-Kendall (MK) test, have been used to analyze groundwater trends and predict future changes [53, 68, 81]. These methods are appreciated for their simplicity and minimal data needs, making them suitable for areas with limited resources. For example, linear regression has effectively identified long-term GWL trends [68], while ARIMA models work well for short-term forecasting based on time series data [51, 53]. The MK test is often used alongside Sen's slope estimator to find consistent trends in hydrological data [81].

However, traditional statistical methods have their limitations. They typically assume linearity, stationarity, and independence, which are not common in real-world groundwater systems. Because methods such as linear regression, ARIMA, and the MK test are linear or monotonic by construction, they struggle to represent nonlinear GWL responses to factors such as rainfall, evapotranspiration, land use, and pumping rates [71]. This is a distinct limitation from that of physically-based numerical models (e.g., MODFLOW-type simulations), which explicitly solve nonlinear governing equations of groundwater flow but instead face a different constraint: calibrating them typically requires extensive, high-quality data on aquifer properties such as transmissivity, hydraulic conductivity, and recharge rates, and their accuracy is sensitive to initial conditions, which can compromise reliability, especially as climate conditions change [76]. In practice, machine learning approaches have been shown to outperform both families of traditional model on accuracy grounds [3, 10], though, as discussed below, they introduce their own data demands.

In Bangladesh, we see both the strengths and weaknesses of these methods. Groundwater is extensively used for dry-season boro rice cultivation, particularly in areas like the Barind Tract and Madhupur Tract, where aquifers are shallow and declining. Conversely, coastal regions face rising water tables due to tidal effects and saline intrusion [5, 7]. Using data from the Bangladesh Water Development Board, Shamsudduha et al. [68] applied linear regression to GWL records from 1985 to 2005 and noted widespread declines, especially in irrigated areas [68]. Mirdad [46] also observed significant drawdowns linked to unregulated pumping [46].

To enhance trend detection, methods like Seasonal and Trend decomposition using Loess (STL) have been developed. STL separates seasonal, trend, and residual components without assuming linearity, making it suitable for identifying changing recharge patterns, like reduced monsoon recharge in the Barind Tract [68]. However, STL is limited to univariate analysis and struggles to incorporate spatial variability or external climate influences such as El Niño-Southern Oscillation (ENSO). Most traditional models focus on individual points and fail to address the broader effects of monsoon variability, tidal dynamics, and human activities across Bangladesh. These challenges underscore the need for more advanced forecasting methods that are spatially aware and data-driven.

Machine Learning Approaches

Machine Learning is transforming groundwater forecasting by effectively modeling complex relationships among hydrological, climatic, and human factors, uncovering hidden patterns and adapting to changing conditions in ways that make it well-suited to long-term groundwater predictions. This capability comes with its own cost, however: ML models, and deep learning architectures in particular, typically require large volumes of training data to generalize well, a demand that in many respects mirrors the extensive calibration data physically-based numerical models require, and one that can be difficult to satisfy in regions with sparse or fragmented monitoring networks [33].

Ensemble learning techniques

Ensemble learning techniques, which aggregate predictions from multiple decision trees to reduce variance and improve generalization [9], have proven very effective for groundwater forecasting. Notable models like XGBoost, CatBoost, Gradient Boosting Machine (GBM), and LightGBM performed exceptionally well, achieving a Mean Standard Percentage Error (MSPE) of 32.33% in a three-month forecasting study. This was better than traditional Random Forest algorithms, as found by May-Lagunes et al. [45]. Additionally, a new model called Canonical Correlation Forests with Random Features (CCF-CRF) has also shown great promise, outperforming conventional methods like Least Squares Support Vector Regression (LS-SVR) in long-term forecasts, as noted in the research by Wang et al. [88]. Closer to this study's setting, Nury et al. [53] applied Random Forest alongside ARIMA and LSTM-RNN baselines to assess groundwater variability across northeastern Bangladesh, illustrating that ensemble tree-based methods remain competitive even in data-scarce South Asian aquifer systems.

Support Vector Machines (SVMs)

Support Vector Machines (SVMs) are effective for long-term forecasting because they can handle changes in input data and adapt to longer time periods [100]. This makes them a good option for applications requiring long-range predictions. On the other hand, Artificial Neural Networks (ANNs) work better for short-term groundwater level predictions since they can learn complex patterns in data. However, ANNs struggle with accurate long-term predictions, especially with complex, high-dimensional datasets [50, 51, 100]. This indicates that while ANNs are useful for immediate forecasts, their performance decreases over longer time frames and in more complex data situations.

Hybrid and Neuro-Fuzzy Systems

Hybrid models combine different techniques to improve predictive accuracy. For instance, Adaptive Neuro-Fuzzy Inference Systems (ANFIS) blend neural networks with fuzzy logic, allowing them to handle uncertainty effectively. As a result, ANFIS achieves a Root Mean Square Error (RMSE) of only 0.02 meters and a high coefficient of determination (R^2) of 0.96, outperforming traditional Artificial Neural Networks (ANNs) in predictive modeling [18]. Similarly, Genetic Programming (GP) refines forecasting by integrating data on precipitation and historical groundwater levels near weather stations. This approach enhances predictions and reveals complex relationships within the data, underscoring the benefits of hybrid methods in understanding environmental interactions [63]. Hybrid architectures extend beyond neuro-fuzzy and genetic approaches: Nourani, Ejlali, and Alami [50] coupled artificial neural networks with geostatistical interpolation to forecast spatiotemporal groundwater levels in a coastal aquifer, while Sahoo et al. [65] combined multiple soft-computing techniques for space–time groundwater forecasting. Both studies illustrate how blending statistical, geostatistical, and learning-based components can outperform any single method in isolation.

Deep Learning Architectures

Deep learning has significantly changed groundwater modeling by introducing new methods that provide valuable insights. Convolutional Neural Networks (CNNs) are known for quickly analyzing large datasets and extracting complex spatial features [95]. Temporal Convolutional Networks (TCNs) are particularly effective in studying long-term issues like coastal saltwater intrusion, as highlighted by Zhang et al. [104]. Recurrent Neural Networks (RNNs), especially Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM) networks, are essential for simulating changes in groundwater levels. Their ability to capture temporal dynamics is shown in the work of Jeong [28, 85]. Additionally, hybrid models like CNN-LSTM combinations have proven to enhance performance, especially when data is limited. This approach showcases the benefits of merging different network strengths, as discussed by Yang and Zhang [98].

Graph-Based Models

Recent advancements in groundwater forecasting have significantly improved with the use of Graph Neural Networks (GNNs). These advanced models treat monitoring wells as connected points in a graph, which helps understand their spatial relationships better. By using GNNs, researchers can effectively track interactions between wells and predict short-term changes in groundwater levels, greatly reducing error. This approach is transformative for the field [4], a trend consistent with broader surveys of spatio-temporal graph learning across domains [64]. Taccari et al. [79] demonstrated this potential directly on groundwater monitoring data, applying a spatial-temporal graph neural network to capture well-to-well dependencies without relying on a fixed grid structure, further supporting the case for graph-based representations of irregularly distributed monitoring networks like Bangladesh's.

Spatial-Temporal Deep Learning

Traditional AI models like Artificial Neural Networks (ANNs) and Long Short-Term Memory (LSTM) networks often fail to consider the spatial relationships between groundwater monitoring sites. This limitation hinders their ability to accurately reflect the heterogeneity of aquifers, particularly in complex systems like Bangladesh's Bengal Basin, where groundwater dynamics are affected by factors such as monsoon rainfall, intensive irrigation, river-aquifer interactions, and geological variations [4, 68, 70].

To overcome these challenges, Spatiotemporal Graph Neural Networks (ST-GNNs) have been introduced as a promising alternative. These models integrate graph convolutional layers to analyze spatial relationships among wells and temporal modules like RNNs or LSTMs to track changes in groundwater levels over time [15, 79]. ST-GNNs are particularly useful in areas with limited data since they can manage missing data and adapt to nonstationary patterns, making them highly applicable in rural Bangladesh. A significant advancement in ST-GNNs is positional encoding, which incorporates geographic coordinates and environmental variables (like land use, geology, and rainfall) into the node representations. This helps the model account for the spatial variability in aquifer responses, which is vital in Bangladesh where groundwater behavior differs significantly by depth and location [32].

Hybrid architectures can further enhance the performance of these models. For example, the GCN-LSTM model separates spatial learning (via Graph Convolutional Networks) from temporal learning (via LSTMs), improving both accuracy and interpretability [37]. Transformer-based models also outperform CNNs and LSTMs by effectively capturing long-range dependencies in space and time [78].

Emerging frameworks, like Spatio-Temporal Graph Transformer Networks (ST-GTNs) and Geographically and Temporally Neural Network Weighted Regression (GTNNWR), offer new strategies for handling nonstationarity. GTNNWR applies neural networks to estimate spatiotemporal weights for locally weighted regression, demonstrating high accuracy in air quality modeling and showing potential for groundwater applications [93, 99].

However, these advanced models are not without their limits, including high computational demands, sensitivity to data quality, and the need for extensive tuning [99]. These challenges are exacerbated by fragmented monitoring systems and inconsistent historical records. Notably, GTNNWR has yet to be applied to groundwater forecasting or compared with ST-GNNs or CNN-LSTM models in hydrological contexts [99].

Overall, spatial-temporal AI models signify a significant advance in groundwater forecasting, particularly in complex and data-scarce regions like Bangladesh. By capturing both spatial and temporal dynamics, these models provide valuable insights into aquifer behavior and facilitate better water management decisions. Future research should aim to tailor these models to local contexts, enhance data infrastructure, and explore hybrid approaches that balance accuracy with computational efficiency.

Research Gaps and Positioning of this Study

To our knowledge, no prior study has applied a graph transformer architecture to national-scale groundwater forecasting within a monsoon-driven deltaic aquifer system. Concerning relevant studies, Sun, Chang, and Chang [78] ran and evaluated a Transformer against CNN, LSTM, and MLP benchmarks over a 10-20 day horizon in Taiwan's Zhuoushui basin, whereas Taccari et. al [80] applied spatial-temporal graph neural networks to groundwater data without an attention-based temporal component. This thesis extends both lines of work to a seasonal forecasting horizon under Bangladesh's distinct monsoon-and-extraction-driven recharge regime.

Bangladesh has great potential to use spatiotemporal AI for improving water management, but there are several significant challenges to overcome. The country's network for monitoring groundwater is sparse and uneven, like a patchwork of data collection sites. Many of these sites do not have consistent, long-term records, leading to a fragmented understanding of groundwater resources. Advanced AI models, such as spatiotemporal graph neural networks (ST-GNNs), require high-quality, synchronized data, which is currently lacking at a national level.

Additional challenges include limited computing resources, which prevent local institutions from using complex models. There are also issues with understanding how these models work, making it hard to gain support for their use. This is especially problematic when unclear predictions from AI clash with the deep local knowledge of hydrogeology held by community experts.

To move forward, we need to focus on a few key areas. First, we must work urgently to expand and improve groundwater monitoring networks through partnerships between public and private sectors to fill data gaps. Second, we should develop simple AI models that can work even when data is hard to come by. Third, we can enhance our observations by using advanced remote sensing technologies, like GRACE and Sentinel satellites, along with data from citizens. It's also crucial to design forecasting tools in collaboration with local water managers, ensuring these tools work well and fit into local policies. Most importantly, any AI system must consider Bangladesh's specific geological features, complex cross-border water issues, and how groundwater access impacts rural communities. By taking a thoughtful and collaborative approach, we can effectively use AI to improve sustainable groundwater management in one of the world's most vulnerable deltas. These efforts can lead to a future where technology works alongside local knowledge to protect valuable water resources.

Study Area and Data

Study Site

Our study sites in the context of this research concerned the entire extent of Bangladesh as a whole. About 1,300 groundwater observation wells are distributed across the country and collectively organize the national monitoring network operated by the Bangladesh Water Development Board (BWDB), as highlighted in Figure 1.

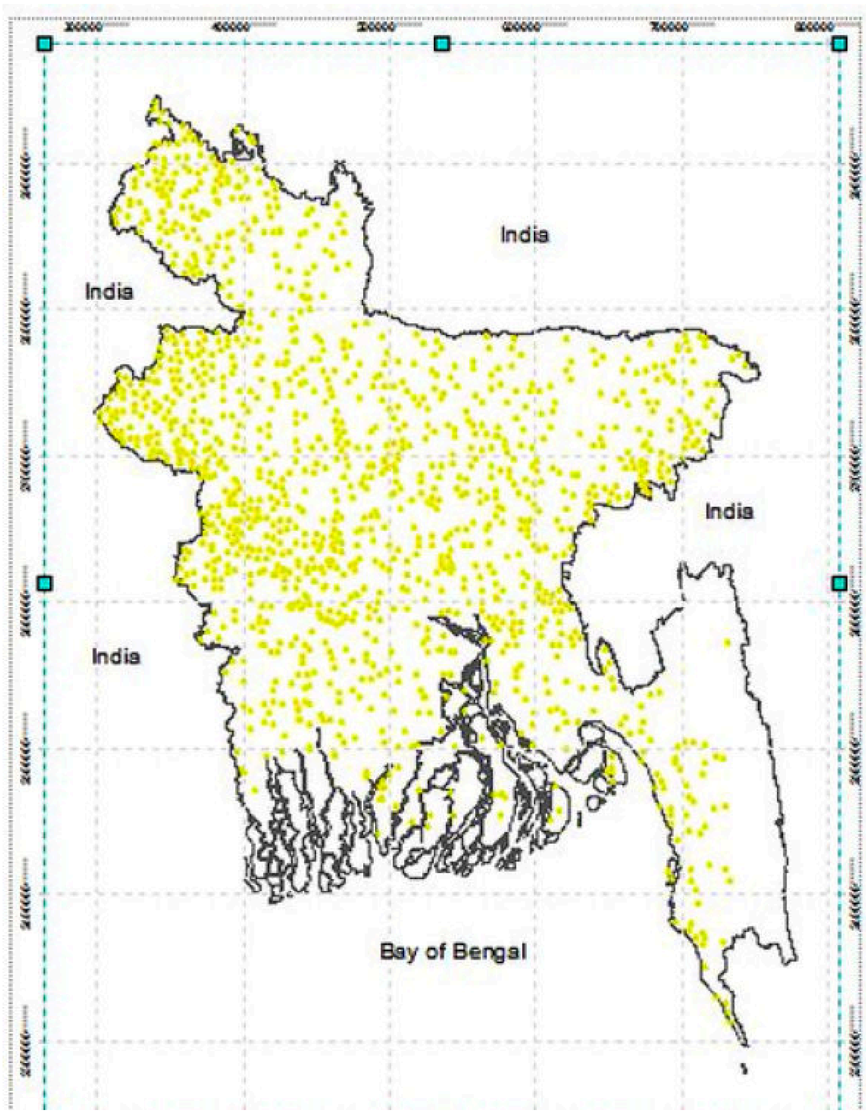


Figure 1: Spatial distribution of groundwater observation wells across Bangladesh (yellow dots; n ≈ 1,300). Well locations are sourced from the Bangladesh Water Development Board (BWDB) national monitoring network. The map extent spans approximately 88.0°-92.7° E and 20.6°-26.6° N (WGS 84). Neighbouring country labels (India) and the Bay of Bengal are shown for geographic reference. The dashed blue border demarcates the spatial domain used for GeoTIFF stack extraction and model training.

To offer multi-decadal records of groundwater level fluctuations substantial for water resource management, spatial coverage spans across all major hydro-geological components of the country; the Ganges, Brahmaputra, and Meghna floodplains included. Well density is more concentrated within the north-western and north-central regions: the Barind Tract and Sylhet Basin, both intensively irrigated. These hotspots reflect the increasing agricultural groundwater demand and historical research investment. In contrast, well coverage is sparser across the south-eastern Chittagong Hill tracts and the Sundarbans delta. This is due to the limitations of drill feasibility and long-term instrumentation hindered by hard-rock aquifers and the dangers posed by tidal flooding.

Primarily recharged by the South Asian monsoon (June-September), accounting for 70-80% of Bangladesh's annual precipitation, the alluvial aquifer system within the country is among the world's largest. Captured by the observation network, water table depth rises by about 2-6m during the monsoon period, facing reductions during the dry seasons (November - May) from irrigation pumping, primarily for Boro Rice cultivation.

Well records acquired from this network offer ground-truth labels used to train and validate the deep learning models described in this report (CNN-LSTM, ST-GNN, ST-GTN), spanning the years 2000-2020 to predict groundwater trends in Bangladesh.For graph-based models, each valid patch of the raster grid—rather than each individual well—is represented as a graph node; patch validity and node connectivity are determined by the target raster's spatial coverage and a fixed cardinal (4-neighbor) grid adjacency, described in full in the Groundwater Graph Dataset section below. Multi-year time series at each node are paired with coincident multi-channel remote-sensing predictors (I.E. NDVI, land-surface temperature, land use, etc.) for supervised spatio-temporal learning.

For pre-processing, every datapoint incorporated-including Groundwater datapoints-is processed and represented as raster image graphics, under the tagged image file format (.TIF or .TIFF). In order to prevent any inconsistencies across our training data, our criteria have it that every image in our created datasets for training should be of a resolution of approximately 1000m or 1km, while ensuring that the same coordinate reference system (CRS) is used across every image.

Rationale for Interpolation Method Selection

Two distinct interpolation methods were selected to match the structural properties of each input dataset. Groundwater observations are sparse, irregularly distributed point measurements (1,000–1,300 wells) exhibiting spatially heterogeneous covariance driven by underlying aquifer heterogeneity (e.g., the low-permeability Barind uplands versus the highly connected alluvial delta); ordinary kriging was therefore applied, as it is the geostatistically optimal (BLUP) estimator for such data, fits spatial autocorrelation directly from the observations via the variogram, and yields an explicit prediction-variance surface—a property particularly valuable given that groundwater depth serves as the target variable for all downstream models. Climate covariates, by contrast, originate from the ERA-Interim reanalysis, an already dense, physically continuous 12 km gridded product; the interpolation task here is one of resampling a smooth field onto the target grid rather than reconstructing a spatial surface from sparse, heterogeneous observations. Thin-plate spline interpolation was applied for this purpose, offering an exact, computationally efficient surface fit appropriate for smoothly-varying atmospheric fields, without requiring the variogram-fitting overhead of a geostatistical model whose core assumptions (sparse, spatially heterogeneous point data) do not hold for this input.

Data Sources

Water Table Data

The Bangladesh Water Development Board (BWDB) has monitored groundwater levels since the early 1970s through an ensemble of wells across the country. Depth measurements from wellheads to groundwater levels are referenced to the Public Works Datum (PWD), set at approximately mean sea level (MSL) with a vertical error of ±0.45 meters. Most monitoring wells were dug wells initially, but piezometers have replaced many. From 1961 to 2006, there were 2,154 monitoring wells: 735 dug wells and 1,419 piezometers, with depths ranging from 3.9 to 352 meters below ground level (average depth 24 mbgl). Dug wells have primarily been replaced by piezometers, and faulty piezometers have also been replaced over time. The database contains >1,200 unique well locations. We treat each replacement well as a separate monitoring station to prevent issues with spurious trends due to well substitution. We have acquired these data points from the BWDB and used them to create a 1000m x 1000m raster of GWL for every 16 days—with each interval marked by the date of the year—per year through Multi-Gaussian Kriging [23].

Groundwater datapoints from the BWDM were initially stored as Microsoft Access Database (.mbd) files. A Python script was created and run to extract the contents of the database file to be stored as a separate CSV for preprocessing, incorporate x, y and longitude, latitude coordinates in the process. Each datapoint is formatted as follows:

- DIST: Zila; The district to which the well is located.
- WELLID: The ID of the well.
- UPZ: Upzila; Subdistrict or small administrative unit in Bangladesh
- LAT: Latitude coordinates
- LONG: Longitude coordinates
- x: x coordinates
- y: y coordinates
- Date: Current date measured per datapoint
- YEAR: Current year
- MONTH: Current month
- DAYS: Current day
- JulianDate: Current day of the year measured
- WEEK: Current week measured based on date
- WaterTable: Water table value measured

Groundwater wells are distinguished based on their well id, district and upz names, and their locations, with their respective water table values varying by the date they were measured. These datapoints were formatted as CSVs separated by year, under the same format to ensure ease of consistency and automation. From the Julian Dates or day of the years (dotys) calculated for each year, mean water table values were calculated for every 16-day interval, stored as 16-day bins within a separate column with each bin or interval represented as a number indicating the start of each 16-day interval. The first bin is represented as the 1st doty, the 2nd bin with the 17th and so forth. The ”gstat” package [54] in an R statistical computing environment [57] has been used for all geostatistical analysis conducted on the water table datasets.

Normal Score Transformation (nscore) Prior to performing auto-kriging prediction on each 16-day interval calculated, water table values attributed to each groundwater well across specific time periods must be normalized to limit the chances of under and overestimation during model predictions. This is performed through normal score transformation, as normally distributed data is not only essential for making Gaussian assumptions, but also an important and required component for Digital Soil Mapping.

Nscore transformation utilizes a graphical transformation technique—such as a correspondence table between equal p quantiles—to normalize any given distribution. [22, 92]:

The first step has n datapoints be ranked from low to high, with datum $z^{(k)}$ be ranked by superscript k across all n datapoints [92].

$z^{(1)} \leq \ldots \leq z^{(k)} \ldots z^{(n)}$....... (1)

The normal score would transform $z^{(k)}$ to be the k/n'th quantile of the standard normal cdf, demonstrated in the 2nd equation below: [92].

$y^{(k)}= G^{-1}\left\{F\left[z^{k}\right]\right\} = G^{-1}(p^{k})$........... (2)

The H2O R package has been used to predict the outcomes of normal deviates where y*(u) has been back-transformed to their original units. In the third equation, F(z) is the cdf of the original [92].

$z^{*}(u)=F^{-1}(G\left(y^{*}(u)\right))$............ (3)

A separate r script—nscore.R—contains the necessary techniques needed to perform normal score transformation on the water table values across each 16-day interval.

Normal-Score (Multi-Gaussian) Ordinary Kriging. A subset of kriging used for interpolating groundwater datapoints into raster maps. A geostatistical interpolation technique that aims to give the best linear unbiased prediction (BLUP) of a spatially-correlated random field across unsampled locations. Not only does it models spatial autocorrelation through a variogram but it also provides prediction variance or uncertainty and is optimal under given assumptions of stationarity as well as through correct variogram specifications.

Kriging assumes that the spatial process observed at a given location s can be modeled through the following equation:

$$Z(s) = \mu + \varepsilon(s)$$

The variogram should depend only on the separation vector:

$$\gamma(h) = \frac{1}{2} Var\left[Z(si) - Z(sj)\right]$$

$$= \frac{1}{2} E[(Z(si) - Z(sj))^2]$$

The overall goal of kriging serves to estimate the value of Z(s_0) at some unsampled location s_0 as the weighted average of n neighboring observations:

$\hat{}Z(s_0) = \sum_{i=1}^{n} \lambda_i Z\left(s_i\right)$

Under the following constraints:

- Unbiasedness: $E[\hat{}Z(s0)\ -\ Z(s0)]\ =\ 0\ \Rightarrow\ \sum_{i=1}^{n} \lambda i$ = 1
- Minimize estimation variance: $\sigma_K^2 =\ Var[\hat{}Z(s0)\ -\ Z(s0)]$

For each lag distance h, the experimental variogram is calculated:

$$\gamma(h) = \frac{1}{2}N(h)\sum_{i=1}^{N(h)} [Z(si) - Z(si+h)]^2$$

With N(h) being the number of point pairs that are separated by the approximate distance h.

A valid variogram model γ(h; θ) is then chosen and parameterized by θ = $\{$nugget, sill, range$\}$. Common variogram models are as follows:

Spherical:

$$\gamma^h = \begin{cases} c0 + c\left[\frac{3h}{2r} - \frac{h^3}{2r^3}\right] & 0 < h \leq r \\ c0 + c & h > r \end{cases}$$

Exponential:

$$\gamma(h) \ = \ c0 \ + \ c(1 \ - \ e^{-\frac{h}{a}})$$

Gaussian:

$$\gamma(h) \ = \ c0 \ + \ c(1 - e^{-(\frac{h}{a})^2})$$

Matérn (Stein's parameterization):

$$\gamma(h) = c_0 + c\left[1 - \frac{1}{2^{\kappa-1}\Gamma(\kappa)}\left(\frac{h}{a}\right)^{\kappa} K_\kappa\left(\frac{h}{a}\right)\right]$$

where K_κ is the modified Bessel function of the second kind and κ is a smoothness parameter.

All four models (Spherical, Exponential, Gaussian, and Matérn) were supplied as candidates to automap's autoKrige function for each 16-day interval; the model minimizing the residual sum of squares between the sample and fitted variogram was automatically selected per interval.

Parameters across the given equations are fit by minimizing the residual sum of squares between the sample and fitted variogram, following the automated model-selection procedure implemented in automap.

To set up Ordinary Kriging as the final step, let γ_{ij} be $||si - sj||$, being the variogram between the data points i and j. Then, set up the variogram between data point i and the prediction location: $\gamma i0 = \ \gamma\left(||si\ 0\ s0||\right).$

The ordinary kriging system is represented by the given matrix-vector multiplication equation below, with μ being a Lagrange multiplier that enforces the unbiasedness constraint.

$$\begin{bmatrix} \gamma 11 & \gamma 12 & \cdots & \gamma 1n & 1 \\ \gamma 21 & \gamma 22 & & \gamma 2n & 1 \\ \vdots & \vdots & \ddots & & \vdots \\ \gamma n1 & \gamma n2 & \cdots & \gamma nn & 1 \\ 1 & 1 & & 1 & 0 \end{bmatrix} \begin{pmatrix} \lambda 1 \\ \lambda 2 \\ \vdots \\ \lambda n \\ \mu \end{pmatrix} = \begin{pmatrix} \gamma 10 \\ \gamma 20 \\ \vdots \\ \gamma n0 \\ 1 \end{pmatrix}$$

In compact form:

$$\begin{bmatrix} \Gamma & 1 \\ 1^\top & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma 0 \\ 1 \end{bmatrix}$$

Matrix inversion or the Cholesky decomposition is then performed to solve the full linear system:

$$\lambda = \Gamma^{-1}(\gamma 0 - \mu 1)$$

Finally, we compute the prediction and kriging variance values.

Prediction:

$$Z(s0) \ = \sum_{i=1}^{n} \lambda i Z(si)$$

Kriging variance or the measure of uncertainty:

$$\sigma_K^2(s0) = \lambda^\top \gamma 0 + \mu$$

Equivalently: $\sigma_K^2 = \lambda^\top \gamma 0 + \mu$

Ordinary kriging—via auto-kriging—has been applied across multiple groundwater data points for each year across Bangladesh in an automated pipeline, generating multiple raster maps covering each 16-day interval per datapoint across years 2000–2020, with each raster map generated covering a spatial resolution of 1 km x 1 km.

Groundwater Analysis

We then compiled and analyzed our monthly groundwater raster datasets to characterize the spatial and temporal variability of groundwater conditions across our study area. We harmonized every raster layer to a common grid geometry, including spatial extent, resolution, and coordinate reference system, to ensure pixel-wise comparability. The input rasters represented groundwater values for individual monthly observations, and all no-data cells were identified using the raster no-data value and converted to missing values to avoid bias during averaging. The raster stack was then organized chronologically using the corresponding observation dates.

To evaluate seasonal groundwater behavior, each raster was assigned to one of three hydrological seasons based on its acquisition date: premonsoon (15 March–15 July), monsoon (16 July–15 November), and winter (16 November–14 March). For each season, a seasonal mean groundwater raster was generated by averaging all valid monthly rasters within that seasonal group on a pixel-by-pixel basis using the arithmetic mean, all the while ignoring missing values. This produced one representative groundwater surface for each season, expressing the long-term mean groundwater condition for premonsoon, monsoon, and winter periods.

To compare groundwater conditions consistently across seasons and months, percentile-based classes were derived from the pooled distribution of finite pixel values. Specifically, the 25^{th} percentile ($P25$), 50^{th} percentile ($P50$), and 75^{th} percentile ($P75$) were calculated from each valid pixel of the seasonal or monthly mean groundwater maps. These thresholds were then used to classify groundwater conditions into relative depth zones. Since lower groundwater values represent a shallower water table and higher values indicate a deeper water table, pixels with values $<P25$ were interpreted as shallow groundwater zones, pixels between $P25$ and $P75$ as intermediate groundwater zones, and pixels $>P75$ as deep groundwater zones. In addition, a four-class summary ($<P25$, $P25$–$P50$, $P50$ - $P75$ and $>P75$) was also calculated for detailed seasonal comparison.

For each monthly and seasonal groundwater map, the areal proportion of each class was calculated as the percentage of total raster pixels occupied by that class. Class frequencies were computed using only valid pixels, while the percentage area was normalized by the full raster extent to allow direct comparison among months and seasons. A separate valid percentage was also calculated to quantify the proportion of pixels with finite groundwater values in each mean raster. This approach allowed for the identification of temporal shifts in the extent of shallow, moderate, and deep groundwater conditions throughout the annual cycle.

Spatial visualization was performed in Python using raster-based mapping. Seasonal and monthly mean groundwater rasters were displayed using a common color scale derived from the pooled distribution of groundwater values, with lower values shown in cooler colors and higher values in warmer colors. Shared map extents and a unified legend were used to facilitate visual comparison between panels. All raster processing, statistical summaries, and figure generation were performed using the Python libraries NumPy, Rasterio, and Matplotlib.

To characterize long-term trends in groundwater depth, a pixel-wise ordinary least-squares regression of mean groundwater depth against time (year) was performed for each monthly and seasonal mean raster stack spanning 2000–2020. For each pixel p, the model:

$$\overline{GW}_{p,t} = \beta_{0,p} + \beta_{1,p} \cdot t + \varepsilon_{p,t} \tag{1}$$

yields a slope estimate $\beta_{1,p}$ (m yr^{-1}) representing the annual rate of change in mean groundwater depth, and a coefficient of determination R_p^2 quantifying the linearity and consistency of that trend. Positive slope values indicate a progressive deepening of the water table over the study period. The resulting slope and R^2 rasters were visualised as combined two-row figures for both the monthly and seasonal temporal aggregations, allowing simultaneous assessment of trend magnitude and trend consistency across the study area.

Climate Data

Hourly weather data such as average, minimum, and maximum temperature, total rainfall, and photosynthetically active radiation has been extracted from the global atmospheric reanalysis ERA- Interim archive produced by the European Centre for Medium-Range Weather Forecasts (ECMWF). This gridded data product includes a large variety of 3-hourly surface parameters and 6-hourly upper-air parameters covering the troposphere [17]. I have processed gridded (12 km x 12 km) hourly data of total rainfall, mean, minimum, maximum temperature and evapotranspiration. Data download and processing has been done with several Python and R scripts.

Datapoints are as follows:

- 2 Meter Temperature (t2m)
- Surface latent heat flux (slhf)
- Total Runoff (ro)
- Total Precipitation (tp)
- Surface net thermal radiation (str)
- Evaporation (e)
- Minimum temperature at 2 meters (mn2t)
- Maximum temperature at 2 meters (mx2t)
- Surface Solar Radiation (ssr)

A broader set of ERA-Interim variables — including surface pressure (sp), sensible heat flux (sshf), 10m wind components (u10, v10), 2m dewpoint temperature (d2m), and photosynthetically active radiation (par) — was processed through the same interpolation pipeline during initial data exploration but was not retained in the final covariate stack, either due to redundancy with retained energy-balance variables or limited hydrological relevance to water-table prediction at a seasonal scale.

Data points are initially downloaded as netCDF files and loaded into Python as GPU-accelerated CUDA RAPIDs dataframe files, with each year attributing the aforementioned data points. Proper conversions are made depending on the data points covered, such as converting Kelvin to Celsius when concerning temperature and converting meters to millimeters. Through the extracted coordinates, a station grid is created in the form of station points, with each station represented by the location to which it is located, with data attributed to each station extracted and ordered by the time and date in 3-hour intervals. Then, daily averages are calculated from the given hourly or sub-hourly data extracted, with each data having their own daily average for a certain data point.

A Bangladesh boundary shapefile is loaded and used to clip the resulting dataset such that any datapoints with coordinates beyond the boundary will be filtered out. Afterwards, 16-day interval means are calculated in accordance to MODIS data reporting dates. Each interval covers a span of approximately 16-days and labeled by the day of the year (doty) marking the beginning of each interval I.E. the first interval is marked by the doty of 1, the 2nd marked by doty 17 and so forth. These steps have been fundamental in reorganizing the aforementioned datapoints before invoking Thine Spline Interpolation to map them across raster maps.

Thine Plate Spline Interpolation (TPS) A non-parametric, spline-based technique aimed to fit a smooth surface across scattered data in at least two spatial dimensions. It seeks the minimum possible bending energy among all differentiable functions that match the input data. such that the surface being fit remains smooth as is. In its 2D form, it would seek a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ passing through given data points while minimizing said thin-plate bending energy, naturally expanding to higher dimensions or higher-order splines.

Given N datapoints $(xi, yi, zi)Ni = 1$ where no two points conflict with one another, TPS seeks to find a function that interpolates said data:

$$f(x, y) = \alpha 0 + \alpha 1x + \alpha 2y + \sum_{i=1}^{N} wi\phi(\| (x, y) - (xi, yi) \|)$$

The global trend is given by parameters $\alpha 0,\ \alpha 1,\ and\ \alpha 2y$ forming a polynomial of degree 1 while wi represents the coefficients for radial basis.

The function $\phi(r)$—a radial basis function being the fundamental solution associated —is represented as:

$$\phi(r) = r^2 ln(r)$$

f(x,y) must satisfy the following interpolation conditions:

$$f(xi, yi) = zi, i = 1, \dots, N.$$

To ensure that the solution remains unique, without any degeneracies present, f(x,y) must also satisfy the following:

$$\sum_{i=1}^{N} wi = \sum_{i=1}^{N} wixi = \sum_{i=1}^{N} wiyi = 0$$

Thus, a linear system for the unknown parameters α0,α1,α2,w1,...,wN is established.

Thine plate splines are a result of minimizing the given bending energy function, subject to the interpolation constraints f(xi,yi)=zi:

$$J[f] = \iint (\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2 dxdy$$

TPS is formed from solving the Euler-Lagrange equations that are associated with energy minimization under interpolation condition constraints, yielding a solution to be a polynomial with a maximum degree of 1, summed with the weighted sum of the radial basis functions $\phi(r) = r^2 ln(r)$.

By substituting f(x,y) results in a system of N + 3 linear equations, solved to obtain the TPS coefficients:

$$[0 \quad P^\top \quad P \quad K][\alpha \quad w] = [0 \quad z]$$

- P is a N x 3 matrix of rows [1, xi, yi]
- K is the N x N matrix of entries $Kij \ = \ \phi(\| \ (xi, yi) - (xj, yj) \ \|)$
- z is a vector of all observed zi values.
- a is $[\alpha 0, \alpha 1, \alpha 2]^\top \ and \ w = [w1, \ldots, wN]^\top$.

In Python, Thin Spline Interpolation for each 16-day interval per datapoint is performed through SciPy's RBFInterpolater function, acting as its main kernel, with smoothing set to a value of 0.0 as it takes in an interpolation grid created based on the maximum x y coordinates of the Bangladesh boundary map, adjusted by the target raster dimensions. The resulting interpolation is then written as raster files through rasterio, of the Bangladesh Universal Transverse Mercator (BUTM 2010) coordinate reference system and masked with the boundary file. These steps are repeated in a batch processing pipeline for all climate datapoints, across all intervals for every year observed, from 2000 to 2018.

MODIS NDVI Data

Moderate Resolution Imaging Spectroradiometer (MODIS) 16-day time series of 250m gridded normalized difference vegetation index (NDVI) and Enhanced vegetation index (EVI) data has been acquired from the NASA Land Processes Distributed Active Archive Center, USGS/Earth Resources Observation and Science Center, Sioux Falls, South Dakota. Data download has been conducted through the Google Earth Engine (GEE) module via Google Colab, collecting images across 16-day intervals for each year from 2000 to 2020. Two granules of the MOD13Q1 compounds cover the study region; every image is clipped using a Bangladesh boundary file.

Raster images are saved at a scale of 80 and at an initial coordinate reference system of EPSG:4326, with each image designated by the beginning doty of the current 16-day, grouped by the chosen year. However, images initially remained at a high spatial resolution of 250 m. To remedy this, a separate Python script—using the Rasterio library—is created that resamples every MODIS file downloaded such that every image is of a spatial resolution of 1 km and bears the same coordinate reference system and extent with every other datapoint. It is done by reading a separate Raster file covering some climate datapoint generated from the prior preprocessing task, being a mx2t Raster file. The script would read the metadata of said Raster file-including bounds, transformation matrix, CRS-and use it to reproject and transform every NDVI raster image it reads, saving every image as separate files.

Landuse Data

Anthropogenic changes in land-use and land-cover on a global scale is an important driver of climate change [90] which in turn would impact groundwater distribution. HILDA+ (Historic Land Dynamics Assessment+) v.2b has processed and presented a global dataset of annual high-resolution land-use and land-cover (LULC) maps from 1960 to 2020 [89], useful for monitoring sustainability changes resulting from landscape patterns. The dataset was downloaded through the PANGAEA data publisher, more specifically, a resampled variant of the original HILDA+ dataset that incorporates nearest neighbor resampling to reproject the dataset to the WGS 84 (EPSG: 4326). However, resampling has also increased to the resolution to 0.01 degrees. To remedy this, after clipping the full maps within Bangladesh's boundaries, we would apply the same resampling scheme we used in projecting our MODIS NDVI imagery, not only resembling every landuse imagery to have a resolution of 1 km of reprojecting them to have the same CRS as every other datapoint used in this study. Figure 2 outlines the initial feature classes of our landuse maps for this study, through our first year. Each feature is identified by a particular numeric code, which is then gone into further detail through table 1.

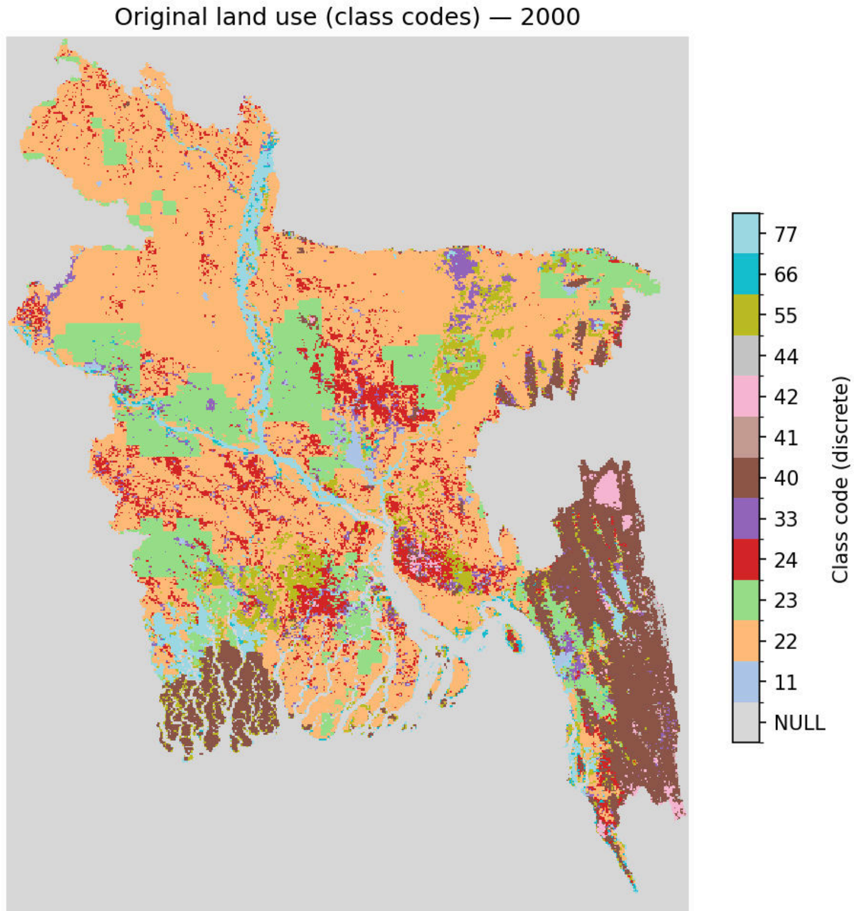


Figure 2: Initial landuse feature maps, highlighted for the year 2000.

However, the initial Hilda+ feature set contains several redundant categories, namely covering the forest category. These include multiple classes either completely absent from Bangladesh's own landscape, as highlighted in 2, or show only minimal hydrogeological distinction within the 1 km spatial resolution used in this study. To address this, we applied a reclassification scheme to reduce these categories to seven consolidated classes, from the original fourteen categories. Table 2 shows our new categorical values and how the old categories are mapped to them.

Our rationale for consolidation of these features grounds itself in hydrological relevance. Any subcategory that falls under crops—namely annual crops, tree crops, and agroforestry—is merged into single "Agriculture" category. From the perspective of groundwater recharge concerning agri-

Feature Code	Category
11	Urban Areas
22	Annual Crops
23	Tree Crops
24	Agroforestry
33	Pasture/Rangeland
40	Forest (Unknown/other)
41	Forest (Evergreen, needle leaf)
42	Forest (Evergreen, broad leaf)
43	Forest (Deciduous, needle leaf)
44	Forest (Deciduous, broad leaf)
45	Forest (Mixed)
55	Unmanaged Grass/Shrubland
66	Sparse/No Vegetation
77	Water

Table 1: Quantitative landuse features mapped to qualitative categories.

Feature Code	Reclassified Code	Category	Reclassified Category
11	1	Urban Areas	Urban Areas
22	2	Annual Crops	Agriculture
23	2	Tree Crops	Agriculture
24	2	Agroforestry	Agriculture
33	3	Pasture/Rangeland	Pasture/Rangeland
40	4	Forest (Unknown/other)	Forest
41	4	Forest (Evergreen, needle leaf)	Forest
42	4	Forest (Evergreen, broad leaf)	Forest
43	4	Forest (Deciduous, needle leaf)	Forest
44	4	Forest (Deciduous, broad leaf)	Forest
45	4	Forest (Mixed)	Forest
55	5	Unmanaged Grass/Shrubland	Unmanaged Grass/Shrubland
66	6	Sparse/No Vegetation	Sparse/No Vegetation
77	7	Water	Water

Table 2: Reclassified landuse features and their associated codes.

culture, what matters for groundwater recharge and extraction boils down to whether land is under cultivation, a significant driver for irrigation demand and suppression of surface infiltration. In other words, having multiple crop distinctions is ultimately less hydrologically meaningful and offers minimal predictive signal for water table dynamics than by simply having a cultivated/noncultivated distinction. Likewise, the five Hilda+ forest subcategories are all consolidated into a single Forest category. Having the initial needle-leaf vs. broad-leaf distinctions as opposed to the more deciduous distinctions are largely irrelevant to modeling recharge dynamics at only 1 km range in Bangladesh. Regardless of species composition, forest cover functions as a low-extraction, high-interception class. Every other category remains unchanged from the initial set.

In our model pipeline, our reclassified land-use categories are encoded as categorical integers across our annual raster maps, akin to the initial maps that we've downloaded. Stacked as a static or annually-varying channel within the covariate tensor X, alongside the time-varying climate and NDVI layers. As land-use changes at an annual rate as opposed to a bi-weekly timescale, each year's reclassified map is instead tiled across the 23 bi-weekly intervals within a particular year prior to stacking. This is to ensure consistent tensor dimensions across all our input channels.

Population Data

Monitoring the population overtime is an important metric in forecasting groundwater sustainability, especially in Bangladesh, being a densely-populated country with a population that would only continue to grow overtime. Effective policies in sustainable groundwater management, especially with a dense, growing population, are but a taxing endeavor. In this study we will use population metrics to see whether they have an influence on Groundwater prediction. To represent population metrics in this study, we would be using the GlobPOP dataset, a 33-year (1999-2022) gridded annual global population dataset of high spatial resolution generated and analyzed by Luling et al. [40] A dataset of high-precision raster imagery generated from cluster analysis and statistical learning approaches, GlobPOP serves as an essential product to help with implementing effective policies especially in sustainability development. The dataset itself is segmented into two categories: population count and population density. We would instead download imagery from the set monitoring global population density. Across each imagery, spatial resolution is especially high, of about 30 arcseconds, with a pixel size of approximately 0.01 degrees.

After cropping each annual image to be confined within Bangladesh's boundaries, we would apply the same resampling scheme that we've applied to our MODIS and Landuse datapoints to normalize our population datapoints to bear the same dimensions and datapoints. Harmonizing population density to the same 1 km grid and CRS as every other input layer ensures pixel-wise comparison with groundwater.

To obtain a temporally meaningful signal from 20-year population density time series (2000-2020), we fitted a pixel-wise linear regression to a single pixel time series, of population density against the respective year of each map. For each spatial pixel p per map, we fit the following model:

$$\text{PopDensity}_{p,t} = \beta_{0,p} + \beta_{1,p} \cdot t + \varepsilon_{p,t} \tag{2}$$

where t denotes the calendar year, $\beta_{1,p}$ is the estimated annual rate of population change (persons km^{-2} yr^{-1}), and the coefficient of determination R_p^2 quantifies how closely the pixel's population trajectory conforms to a linear trend over the study period. Then, we extracted the slope (β_1) and R^2 of our maps, and generated two summary raster maps. Our slope map captures the magnitude and direction of population change at each location, whereas our R^2 highlights temporal consistency or linearity of population change. Both metrics distinguish areas of rapid, sustained urbanization, culminating from high slope and high R^2, from areas experiencing volatile or non-linear population shifts, where slope is high with otherwise low R^2. The latter could possibly reflect real world displacement dynamics, such as flood-driven migration, or uneven rural-to-urban transitions that characterize Bangladesh's demographic landscape.

Data Preprocessing

Data Normalization

Upon wrapping up data collection, we transition to load our datapoints via Pytorch's data loaders in order to pass them into our models. To prepare our data loading, our covariates and targets are

Subset	Year Range
Training	2000 - 2016
Validation	2017 - 2018
Testing	2019
Forecasting	2020

Table 3: Training/testing sets, split across the time axis annually.

initially built as Dask arrays via the Python Dask library. For each given calendar pair (y, τ) (year, day-of-year index), covariate layers are read as delayed tasks and are stacked into the following format:

$$X_{dask} \in R^{T_{tot} \times C \times H \times W}, Y_{dask} \in R^{T_{tot} \times 1 \times H \times W}.$$

This would follow blocked lazy arrays and task scheduling [29].

Then for normalizing our dask arrays via channel-wise affine normalization on our inputs X, we let Ω_{train} index time slices that are used for statistics. For each channel c:

$$\mu_c = \frac{1}{|\Omega_{train}|\, HW} \sum_{t \in \Omega_{train}} \sum_{i,j} X_{t,c,i,j}, \sigma_c = Std_{t \in \Omega_{train}, i,j} \left(X_{t,c,i,j}\right),$$

With σ_c floored to avoid any blow-ups in any calculations made should a miniscule value be calculated for σ_c. Our inputs are normalized through the following equation:

$$X_{t,c,i,j} = \frac{X_{t,c,i,j} - \mu_c}{\sigma_c}.$$

Data Splitting and Loading

Once normalized, our dask arrays, both X and Y, are then evaluated as Numpy arrays and split into hold-out sets along the time axis, highlighted in table 3.

Our testing set is used to test how well our models can generalize to an unseen future year through data directly learned from the previous years.

All three models follow this same procedure in preparation for data-loading. However, being graph-based models, ST-GNN and ST-GTN both follow a different procedure from CNN-LSTM to format their Numpy arrays into a data loader. The following table notes the relevant symbology:

Symbol	Meaning
H, W	Raster height and width (pixels).
p	Patch size `PATCH_SIZE` (square $p \times p$).
s	Stride `PATCH_STRIDE` between patch origins.
T_{tot}	Total biweekly time slices in the stacked cube (all selected years).
C	Number of input channels after stacking covariates. (biweekly + yearly)
T	Input sequence length `SEQLEN` (history fed to the model).
H_{pred}	Forecast horizon `PREDHORIZON` (future steps supervised per sample).
B	Batch size.
N	Number of graph nodes (valid patches).
d	Model dimension.
$A \in \{0, 1\}^{N \times N}$	Binary adjacency with self-loops (4-neighbor patch grid).

Groundwater Graph Dataset To pass inputs into our ST-GNN and ST-GTN implementation, we formatted our Numpy arrays as a Graph. That is, a Patch Graph G = (V,E). As images stored within the numpy arrays per channel c are split into multiple patches of a fixed size, vertices V represent patch origins (i,j) across a stride-s grid (PATCHSTRIDE s), with a patch size of p x p (PATCHSIZE p), iff some target value in the patch is finite overtime. The edges of this patch graph represent cardinal neighbors to each patch, origins differing by +- s across row or column (identity included). This patch graph is represented as an adjacency matrix, where $A_{uv} = 1$ if $u = v$ or for patches that share an edge. Otherwise, $A_{uv} = 0$.

Traversing through our patch graph is done through supervision or sliding windows. For window start t and node v of origin (i,j):

$$x_{v,t:t+T-1} = PatchMean\left(X_{t:t+T-1,:,i:i+p,j:j+p}\right) \in R^{T\times C},$$

$$y_{v,t+T:t+T+H_{pred}-1} = PatchMean\left(Y_{t+T:t+T+H_{pred},1,i:i+p,j:j+p}\right).$$

Creating graph signals compatible with attention over N is done through pooling; spatial means collapsing each patch to one vector for each timestep. Following these procedures, our Groundwater Graph dataset returns tensors of the expected shape: (T, N, C) and $(H_{pred}, N, 1)$ for X and Y respectively. By passing these tensors into a PyTorch Dataloader, the batch dimension B is prepended.

Groundwater Dataset Unlike ST-GNN and ST-GTN, CNN-LSTM does not require its inputs to be formatted as a graph. It does, however, draw on the same $p \times p$ patch-extraction mechanism described above: each patch supplies a local spatial neighborhood over which the model's convolutional layers learn spatial features, while the temporal ordering of frames within each patch is passed to the LSTM component to learn the sequence dynamics. The key difference from the graph formulation is what happens to a patch once it is extracted. In the Groundwater Graph Dataset, each patch is spatially averaged (via PatchMean) into a single feature vector, so that it can be treated as one node of a shared graph reused across every training batch. Here, no such pooling or graph structure is imposed: the full $p \times p$ spatial grid of each patch is retained as-is, and every (t, i, j) triple—window start t and patch origin (i, j)—is treated as an independent training example, with no adjacency or shared structure enforced between patches.

A triple (t, i, j) is considered valid if the corresponding target patch (defined below) contains at least one finite, non-missing groundwater pixel across the forecast horizon H_{pred}; triples without any finite target pixel are excluded from the dataset. During training only, valid patches may optionally be flipped horizontally or vertically as a form of data augmentation; no augmentation is applied at validation, test, or forecast time.

For each valid triple (t, i, j) with patch origin (i, j) and window start t:

$$\mathbf{x} = \mathbf{X}_{t:t+T-1,\,:,\,i:i+p,\,j:j+p} \in \mathbb{R}^{T\times C\times p\times p},$$

$$\mathbf{y} = \mathbf{Y}_{t+T\,:\,t+T+H_{pred}-1,\,1,\,i:i+p,\,j:j+p} \in \mathbb{R}^{H_{pred}\times 1\times p\times p},$$

where $\mathbf{X}$ and $\mathbf{Y}$ are the channel-wise normalized covariate and target cubes defined in the Data Splitting and Loading section above. Passing these tensors into a PyTorch Dataloader prepends the batch dimension B, yielding batches of shape (B, T, C, p, p) for inputs and $(B, H_{pred}, 1, p, p)$ for targets—in contrast to the pooled, per-node (T, N, C) and $(H_{pred}, N, 1)$ representations used by the graph-based models.

Methodology

Problem Formulation

In this study, we formulate groundwater depth forecasting as a supervised spatiotemporal regression problem. First, we let $\mathbf{X} \in \mathbb{R}^{T_{tot} \times C \times H \times W}$ denote the stacked covariate cube as described in the Data Preprocessing section, each of the T_{tot} biweekly time slices carries C channels—climate, vegetation, land-use, and population covariates–over a raster of height H and width W that spans Bangladesh's spatial domain. We then let $\mathbf{Y} \in \mathbb{R}^{T_{tot} \times 1 \times H \times W}$ represent the corresponding water-table depth target defined over the same spatiotemporal grid; missing observations are encoded as non-finite values.

Given a historical window of $T = \mathrm{SEQLEN}$ consecutive time slices that end at time t, $X_{t:t+T-1,:,:,:}$, our forecasting task is to predict water-table depth $H_{pred} = \mathrm{PREDHORIZON}$ steps into the future, $Y_{t+T:t+T+H_{pred}-1,1,:,:}$, at every spatial location for which a finite target exists over the forecast window. We formally seek to identify a function f_θ that, at the patch level,

$$f_\theta : \mathbb{R}^{T \times C \times p \times p} \longrightarrow \mathbb{R}^{H_{pred} \times 1 \times p \times p},$$

or, under the graph formulation introduced below,

$$f_\theta : \mathbb{R}^{T \times N \times C} \longrightarrow \mathbb{R}^{H_{pred} \times N \times 1},$$

minimizes a training objective computed between predicted and observed water-table depth. Because the patch-based and graph-based architectures evaluated in this study handle the sparse, intermittent coverage of the BWDB monitoring network (Chapter 3) differently, we define two related but distinct objectives below.

For the patch-based CNN-LSTM architecture, f_θ minimizes the **masked mean-squared error** between predicted and observed water-table depth, computed only over valid (finite) target locations:

$$\mathcal{L}_{\text{CNN-LSTM}}(\theta) = \frac{1}{|\Omega|} \sum_{(b,h,i,j)\in\Omega} \left(\hat{Y}_{b,h,i,j} - Y_{b,h,i,j}\right)^2$$

where Ω denotes the set of batch, forecast-step, and spatial indices for which an observed target exists. Masking the loss to Ω ensures that no spurious gradient signal identified at unobserved locations is ever injected by the sparse, intermittent coverage of the BWDB network.

For the graph-based ST-GNN and ST-GTN architectures, f_θ instead minimizes a **Huber loss** (smooth L_1) between predicted and observed node-level water-table depth, computed over the patch nodes of the patch graph $G = (V, E)$ introduced in Chapter 3, together with an optional **spatial-diversity regularization** term that discourages degenerate, spatially-uniform forecasts across nodes:

$$\mathcal{L}_{\text{graph}}(\theta) = \frac{1}{|B|} \sum_{(b,h,n)\in B} L_\delta\left(\hat{Y}_{b,h,n} - Y_{b,h,n}\right) \; - \; \lambda_{\text{div}} \cdot \frac{1}{|B|} \sum_{b,h} \operatorname{std}_n\left(\hat{Y}_{b,h,\cdot}\right)$$

where

$$L_\delta(e) = \begin{cases} \frac{1}{2}e^2 & |e| \le \delta \\ \delta\left(|e| - \frac{1}{2}\delta\right) & |e| > \delta \end{cases}$$

is the Huber loss with $\delta = 1$ (`HuberLoss(delta=1.0)`), B indexes the batch, forecast-step, and node triples in the current mini-batch, $\operatorname{std}_n(\cdot)$ denotes the standard deviation taken across nodes at a fixed (b, h), and $\lambda_{\text{div}} = \text{SPATIALDIVWEIGHT}$ is a tunable regularization weight (setting $\lambda_{\text{div}} = 0$ recovers the unregularized Huber objective). Unlike $\mathcal{L}_{\text{CNN-LSTM}}$, $\mathcal{L}_{\text{graph}}$ is not explicitly masked to a finite-target index set Ω: the graph training loop assumes all targets supplied in a mini-batch are already finite, and treats a non-finite loss value as a hard failure rather than filtering it out.

Both objectives target the same underlying forecasting task described above. Where the three architectures differ is (i) how the spatial domain is represented—as independent $p \times p$ image patches for CNN-LSTM (Groundwater Dataset, above), or as nodes of a patch graph $G = (V, E)$ connected by a stride-s cardinal adjacency A for ST-GNN and ST-GTN (Groundwater Graph Dataset, above)—(ii) how each architecture parameterizes f_θ over the temporal and spatial dimensions, further outlined for each model in the subsequent sections, and (iii) the precise training objective used, as detailed above.

Model Architectures

CNN-LSTM (Baseline)

A **CNN-LSTM** is a hybrid deep learning architecture that pairs a convolutional feature extractor with a recurrent temporal model, and is a natural starting point for spatiotemporal grid data such as video, satellite imagery, or—as in this study—stacked raster covariates. The design rests on a simple division of labor: convolutions are well-suited to modeling local, translation-equivariant spatial structure within a single time slice [35], while recurrent units are well-suited to modeling how a representation evolves across a sequence of such slices [26]. Rather than fuse these two operations into a single recurrent cell as in the Convolutional LSTM of Shi et al. [72]—where convolutions replace the fully-connected gates of the LSTM itself so that spatial and temporal mixing happen jointly at every step—our implementation *decouples* the two stages: a weight-shared CNN first compresses each raster frame into a compact feature vector, and a separate LSTM then models the resulting feature sequence. This decoupling keeps the recurrent state low-dimensional and the model comparatively cheap to train on the $p \times p$ patches used here, at the cost of discarding fine-grained spatial detail before the temporal model ever sees it; the skip connections described below are one way of partially recovering that detail at the decoding stage. Because CNN-LSTM operates independently on each regular image patch rather than on an irregular graph of wells, it serves as the dense-grid counterpart to the graph-based ST-GNN and ST-GTN architectures introduced next, and is treated throughout this study as the baseline against which their added spatial structure is evaluated.

Per-frame CNN Encoder For each time index τ and batch element b, a **weight-shared** CNN maps

$$\mathbf{x}_{b,\tau} \in \mathbb{R}^{C\times p\times p} \;\mapsto\; \text{GAP}\left(\text{conv-pool stack}\left(\mathbf{x}_{b,\tau}\right)\right) = \mathbf{f}_{b,\tau} \in \mathbb{R}^{F},$$

where F is the last encoder channel count (`CNN_CHANNELS[-1]`). Successive conv-pool blocks widen the channel dimension while shrinking the spatial resolution, so that deeper layers aggregate progressively larger receptive fields over the patch; global average pooling (GAP) then collapses the final feature map to a single vector per frame, discarding absolute spatial position in favor of a compact summary of "what" is present in the patch at time τ. Weight-sharing across τ ensures the same spatial filters are applied to every frame in the window, so that the encoder's parameter count does not grow with sequence length T. Skip feature maps $\mathbf{s}^{(1)}, \mathbf{s}^{(2)}, \mathbf{s}^{(3)}$ and bottleneck are kept from the last input frame for decoder conditioning, preserving the spatial detail that GAP would otherwise discard for at least the most recent time step.

This implementation follows translation-equivariant feature extraction on grids through convolutions [35].

Bidirectional LSTM (Temporal) Let stack $\mathbf{f}_{b,1:T}$ be a sequence, our bi-directional LSTM, following Hochreiter and Schmidhuber [26] and [24], outputs:

$$\mathbf{H}_b = BiLSTM\left(\mathbf{f}_{b,1:T}\right) \in \mathbb{R}^{T\times 2R},$$

Where the hidden size R = LSTM_HIDDEN for each direction. The LSTM's gated cell state lets the model retain or discard information from earlier bi-weekly frames as needed, mitigating the vanishing-gradient behavior of plain RNNs over the length-T history window [26]. Running the recurrence in both temporal directions and concatenating the resulting hidden states, following Graves and Schmidhuber [24], allows each position in the encoded sequence to be informed by both earlier and later frames within the observed window, rather than only by the frames that precede it—useful here since the full history window is available at inference time and is not being decoded causally step by step.

Optional Temporal Self-Attention Should it be enabled, multi-head self-attention (Vaswani et al. [83]) is applied over the time dimension, with the module returning the last time index as context. Where the BiLSTM propagates information step by step through its recurrent state, self-attention instead lets every time step attend directly to every other time step in a single operation, which can make it easier for the model to weigh a distant but informative frame—for instance, a past recharge event several bi-weekly steps back—without that signal having to survive propagation through the entire intervening recurrence.

Spatial Decoder (Skip Connections) A vector $\mathbf{c}_b$ (context) is linearly expanded to a small bottleneck grid and concatenated with the last frame's bottleneck feature map, before being unsampled with conv transpose + skip fusion: an encoder-decoder pattern that closely relates to U-Net [60] and FCN-style decoders [39]. As in those architectures, the skip connections exist to counteract the information bottleneck created by compressing each frame down to a single feature vector: without them, the decoder would have to reconstruct fine spatial variation in water-table depth from $\mathbf{c}_b$ alone, whereas fusing in the last frame's higher-resolution feature maps lets it recover boundary and texture detail that GAP had discarded.

H_{pred} channels are emitted by the final 1 x 1 convolution, as outputs are passed through sigmoid x constant to then enforce a bounded, domain-specific physical range.

Our implementation follows the general intuition of CNN-LSTM. CNN handles local spatial patterns and multi-scale context on each patch, while LSTM (+ optional attention) handles temporal dependence across bi-weekly frames. The decoder then reconstructs spatial maps for each future step in the horizon.

CNN-LSTM-style models remain a widely used and actively developed baseline for hydrological time series, and several directions could extend the implementation used here. Benchmarking work comparing LSTM, CNN, and NARX architectures for groundwater level forecasting has found that recurrent and convolutional-recurrent models are broadly competitive with one another but that no single architecture dominates across all well sites [95], suggesting that the patch-level, weight-shared design adopted here is a reasonable but not uniquely optimal choice among CNN-LSTM variants. More tightly coupled alternatives are a natural next step: the Convolutional LSTM of Shi et al. [72] fuses spatial and temporal mixing inside a single recurrent cell rather than encoding each frame independently before recurrence, and combinations of temporal convolutional networks with LSTM layers have shown improved skill for coastal groundwater level forecasting by capturing multi-scale temporal patterns directly [104]. Physics-informed variants offer a further avenue: gated recurrent units constrained by physical relationships have improved groundwater level fluctuation prediction over purely data-driven baselines [28], and LSTM-based reconstruction of missing groundwater records [85] points toward using the same recurrent backbone for data-gap filling ahead of forecasting, which could reduce the reliance on masking sparse targets during training. Comparative studies applying LSTM-RNN, ARIMA, and random-forest baselines to groundwater variability specifically within Bangladesh [53], alongside broader evidence that recurrent architectures generalize well across catchments in hydrology more widely [34], support extending the present single-region model toward a regionally pooled CNN-LSTM trained jointly across multiple patches or basins. Finally, attention-first alternatives to the LSTM backbone—such as the transformer-based groundwater forecasting model of Sun, Chang, and Chang [78] and the temporal self-attention already available as an optional module here—suggest that, as done for the graph-based architectures compared in this study, an equivalent transformer-based grid model may offer a worthwhile point of comparison against the recurrent baseline in future work, particularly for longer forecast horizons where attention's constant path length between any two time steps could offer an advantage over the LSTM's sequential propagation of information.

ST-GNN

A **Spatial-Temporal Graph Neural Network (STGNN)** is a deep learning architecture designed to process data that varies over both **space** (relationships between entities) and **time** (evolution of those entities).

Standard neural networks struggle with this type of data:

- **CNNs** work well on grid-like spatial data (images) but not on irregular graphs.
- **RNNs/LSTMs** work well on time series but assume independent sequences, ignoring relationships between variables.
- **Standard GNNs** capture relationships between nodes but typically operate on a single static snapshot, ignoring time dynamics.

Recall from Chapter 3 that a node in our patch graph $G = (V, E)$ does not represent an individual BWDB monitoring well, but rather a fixed $p \times p$ spatial patch of the raster domain, retained as a vertex only if it contains at least one finite target observation over time; the node's covariate and target signals are the patch-mean pooled values of every pixel—and therefore every well falling—within that patch (Groundwater Graph Dataset, Chapter 3). The well-level intuition below is nonetheless a useful entry point, since each patch's pooled signal is still driven by the wells it contains; it should simply be read with "well" standing in for "patch (and the well or wells it aggregates)" throughout.

Spatial Dependency: The groundwater level (GWL) at one monitoring well often depends on the GWL at neighboring wells. For example, if well A observes a drop in groundwater level due to heavy pumping, nearby wells B and C may also experience a decrease, as groundwater flows between them. Thus, the state (GWL) of one well/node is influenced by its spatial neighbors—in our implementation, the cardinal, stride-s neighboring patches defined by the adjacency A in Chapter 3.

Temporal Dependency: The GWL at a monitoring well is also dependent on its own historical measurements. If the groundwater at well A was low last month, it's more likely to still be low this month unless some major recharge event occurred. The state of each node evolves over time based on its past states.

The STGNN mechanism thus operates in two intertwined flows when applied to the patch graph of groundwater wells:

Message Passing (Space): At every biweekly time step, each patch node aggregates information from its nearby patch nodes to update its groundwater level representation, considering spatial relationships such as distances, hydrogeology, or flow paths.

State Evolution (Time): The updated information at each patch node is then processed through a temporal model (like an RNN or temporal convolution) to capture how the groundwater level at that node changes from past to future time steps.

In this way, the STGNN can simultaneously model how changes at one patch node propagate through the network of patches (spatially) and how groundwater levels change at each node over time (temporally).

An STGNN architecture is typically composed of stacked layers that alternate or combine spatial and temporal operations. Below is the breakdown of a standard STGNN architecture (e.g., similar to **STGCN** or **Graph WaveNet**), instantiated on the patch graph $G = (V, E)$ and tensors (T, N, C) produced by the Groundwater Graph Dataset of Chapter 3.

The table below extends the notation introduced in Chapter 3 ($H, W, p, s, T_{tot}, C, T, H_{pred}, B, N, d, A$) with the additional symbols specific to the ST-GNN architecture:

Symbol	Meaning
$\widetilde{A}_{norm}$	Symmetric normalized adjacency, built from the patch adjacency A defined in Chapter 3 (see below).
G	Width of last graph conv hidden (`g_hidden`).
R	GRU hidden size (`rnn_hidden`).

Normalized Adjacency (GCN-Style) Chapter 3 defines the patch adjacency $A \in \{0,1\}^{N\times N}$ with self-loops already included by construction ($A_{uv} = 1$ if $u = v$ or if patches u and v share a cardinal edge). We reuse that same A here and write $\widetilde{A} = A$ purely to keep the graph-convolution formula below in its familiar, general form $\widetilde{A} = A + I$ [31]; because self-loops are already present, adding I a second time is not performed in our implementation. Define degree $\mathbf{d} = \widetilde{A}\mathbf{1}$. The symmetric normalization [31] is

$$\widetilde{A}_{norm} = D^{-1/2}\widetilde{A}D^{-1/2}, \quad D_{ii} = \sum_j \widetilde{A}_{ij}.$$

Our implementation lists this as `D_inv_sqrt @ A_hat @ D_inv_sqrt` with small ε to ensure numerical stability.

Graph Convolution Layer For each layer, applied independently at every time step t to the patch-node features $\mathbf{X}_{:,t,:,:} \in \mathbb{R}^{B\times N\times F}$ produced by the Groundwater Graph Dataset of Chapter 3, the model computes:

$$\mathbf{X}' = \sigma\left(\widetilde{A}_{norm}\mathbf{X}\mathbf{W} + \mathbf{b}\right),$$

applying ReLU and dropout in the full stack. Left-multiplying by $\widetilde{A}_{norm}$ replaces each patch node's features with a degree-normalized average of itself and its cardinal neighbors—the graph analogue of the convolution used per-frame in the CNN-LSTM encoder, but operating over the irregular stride-s patch adjacency A from Chapter 3 rather than a regular pixel grid. This incorporates message-passing or convolution on a graph view akin to Kipf and Welling's 2017 work, without applying Chebyshev polynomial parameterization present in certain spectral works.

Temporal GRU per node After spatial encoding is applied independently at each of the T bi-weekly time steps, the hidden $\mathbf{h}_{b,t,n} \in \mathbb{R}^G$ is arranged as a length-T sequence per node:

$$\mathbf{h}_{b,n,1:T} \mapsto GRU, \quad \mathbf{r}_{b,n} = \mathbf{h}_{b,n}^{last} \in \mathbb{R}^R.$$

GRU gating [11, 13] aims to capture temporal memory across biweekly steps separately for each patch node, playing the same role here that the BiLSTM plays over pooled CNN features in the CNN-LSTM baseline (Section 4.2.1), though as a single-direction, per-node recurrence rather than a bidirectional one, since spatial context is here already supplied by the preceding graph-convolution layers rather than by the recurrent module itself.

Forecast Head

$$\widehat{\mathbf{y}}_{b,\cdot,n} = \mathbf{W}_{out}\mathbf{r}_{b,n} \mapsto \mathbb{R}^{H_{pred} \times d_{out}},$$

which are reshaped to $\mathbb{R}^{H_{pred} \times N \times d_{out}}$ to match the $(H_{pred}, N, 1)$ target format of the Groundwater Graph Dataset established in Chapter 3.

Our implementation of ST-GNN combines graph convolution for spatial dependence and RNN for temporal dependence, a standard approach for typical ST-GNN implementations. In addition, our baseline incorporates explicit $\widetilde{A}_{norm}$ and GRU that would further set it apart from our more Transformer-style ST-GTN implementation.

Spatio-temporal graph neural networks (ST-GNNs) face several advanced challenges in real-world scenarios. First, the graph structure underlying the data can be dynamic—changing over time as in social networks or fluctuating road conditions—requiring models that can learn an evolving adjacency matrix rather than relying on a fixed one, as in the adaptive, self-learned adjacency of Graph WaveNet [94], in contrast to the fixed adjacency used by both ST-GNN and ST-GTN in this study. Second, capturing long-term dependencies for accurate predictions far into the future is difficult; to address this, approaches like Graph Transformers employ attention mechanisms to directly connect distant time steps, as in the spatial-temporal transformer of Xu et al. [96] and our own ST-GTN implementation below. Third, practical graph data are often heterogeneous, containing different node and edge types (such as highways versus local streets in a traffic network), necessitating heterogeneous GNN architectures that can flexibly process this diversity—a distinction our single patch-type, cardinal-adjacency graph (Chapter 3) does not currently draw. These innovations have made STGNNs highly adaptable, enabling successful applications in areas such as traffic forecasting using road sensor data and connectivity [101, 94], environmental monitoring with sensor or well networks—including groundwater level forecasting on graphs of wells or well clusters [4, 79]—action recognition from skeletal motion [97], and modeling infection spread across regions [30]. Ultimately, STGNNs work by modularly combining graph convolutions for spatial structure and temporal convolutions or recurrences for time dynamics, yielding models capable of forecasting complex, interconnected system behavior.

ST-GNNs use graph convolutions and RNNs/CNNs to aggregate information from **immediate neighbors** and recent time steps, making them efficient and well-suited for modeling local physical processes (like groundwater flow between adjacent patches and the wells they aggregate). **ST-GTNs** replace these with self-attention mechanisms, allowing every node to directly weigh relationships with **all other nodes** and time steps, capturing global dependencies at the cost of higher computational complexity. In short: ST-GNNs prioritize **local, physics-informed aggregation**, while ST-GTNs prioritize **global, data-driven attention**.

Several extensions to the ST-GNN implementation used here are suggested by recent work on graph-based groundwater and spatio-temporal forecasting more broadly. Graph neural networks applied directly to groundwater level forecasting have shown that graph structure can improve on purely temporal baselines when the underlying well network is represented explicitly [4], and spatial-temporal graph neural networks built specifically for groundwater data report similar gains from encoding well connectivity rather than treating wells as independent time series [79], both consistent with the motivation for evaluating ST-GNN and ST-GTN against the CNN-LSTM baseline in this study. Because the patch adjacency A used here is fixed, stride-s, and purely cardinal (Chapter 3), it encodes proximity on the raster grid rather than hydrogeologic connectivity directly; replacing it with a learned or adaptive adjacency, as in Graph WaveNet [94], or with a distance- or flow-informed graph construction would let the model discover or encode relationships—such as shared aquifer membership between non-adjacent patches—that a fixed cardinal neighborhood cannot represent. Positional encoder GNNs designed specifically for irregularly distributed geographic data [32] point toward an alternative to the regular patch lattice used here, potentially allowing the graph to be built directly over well locations rather than over raster patches. Spatiotemporal regression frameworks that explicitly model spatial and temporal nonstationarity [99] suggest a further direction: allowing the graph-convolution weights themselves to vary across Bangladesh's distinct hydrogeological zones (Chapter 3) rather than sharing a single set of weights over the full domain. Finally, broader surveys of spatio-temporal graph neural network architectures for time series forecasting [15, 64] catalogue many of the message-passing and temporal-modeling variants beyond the GCN-plus-GRU combination used here, offering a wider design space to draw on should future work seek to improve on the present ST-GNN baseline.

ST-GTN

Spatio-Temporal Graph Transformer Networks (ST-GTNs) are a class of deep learning architectures designed to forecast time series data situated on a graph, where multiple interconnected variables interact according to a spatial structure—such as sensors connected by roads (traffic flow), weather stations influenced by geography (weather), or nodes in an energy grid. Traditional models like RNNs/LSTMs are proficient in modeling temporal dependencies but falter at capturing spatial structures and struggle with long sequences, while CNNs/GCNs handle spatial patterns well but are limited in temporal scope, and Transformers excel at long-range dependencies by virtue of self-attention but are intrinsically designed for sequential (1D) rather than graph-structured (non-Euclidean) data. ST-GTNs address these limitations by combining the ability of Graph Neural Networks (GNNs) to model spatial dependencies with the Transformer's power to capture long-term temporal relationships via self-attention mechanisms, thus providing a unified approach for learning from spatio-temporal data on graphs.

Our ST-GTN implementation follows the Transformers paradigm [83], where the model is represented as stacked blocks of self-attention + a position-wise Feed-Forward Neural Network (FFN), bearing residuals and LayerNorm.

Model input to forward: $X \in R^{B \times T \times N \times C}$, the same (T, N, C)-shaped patch-node tensor produced by the Groundwater Graph Dataset of Chapter 3, batched over B; as with ST-GNN, each of the N nodes indexes a patch of the raster domain (Chapter 3) rather than an individual monitoring well.

Input projection and positional encoding For each tensor element (b, t, n):

$$h_{b,t,n}^{(0)} = W_{in}x_{b,t,n} + e_t^{(time)} + e_n^{(node)}, W_{in} \in R^{d \times C}, e_t^{(time)}, e_n^{(node)} \in R^d.$$

Giving us the learnable absolute positional encoding, distinct from the original sinusoidal PE present in the original transformer. Because node identity n here indexes a fixed patch position on the grid rather than an arbitrary, exchangeable graph vertex, the learnable node embedding $e_n^{(node)}$ doubles as a form of positional encoding over space, analogous to how $e_t^{(time)}$ positionally encodes the biweekly time index.

Scaled dot-product self-attention For the queries Q, keys K, and values V, where rows are positions:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

Through the scaled dot-product self-attention, multi-head attention concatenates h heads each with their own projections [83].

Spatial module (SpatialGraphAttention) For some fixed (b,t), nodes n = 1, .., N are treated as the sequence and allow $M \in R^{N \times N}$ be the additive mask from the same stride-s cardinal patch adjacency matrix A defined in Chapter 3 (the Groundwater Graph Dataset) and reused for ST-GNN above:

$$M_{uv} = \left\{\begin{matrix} 0 & A_{uv} = 1 \\ -\infty & A_{uv} = 0 \end{matrix}\right\} \text{(implemented as large negative finite)}.$$

Here, attention logits $L = QK^\top/\sqrt{d_k} + M$ have yielded zero weight on forbidden pairs after applying softmax, which is closely related to masked neighborhood attention in graph attention networks [84]. However, in our implementation, A is instead fixed and binary as opposed to it being learned edge attention weights. Restricting attention to A's cardinal neighborhood, rather than letting every node attend to every other node unmasked, keeps the spatial module physically grounded in the same patch adjacency used by ST-GNN, isolating the choice of aggregation mechanism—graph convolution versus masked attention—as the primary architectural difference between the two models.

Temporal Module For some fixed (b,n) and sequence length T, we get the causal mask $C \in \{0, -\infty\}^{T\times T}$ with $C_{ij} = -\infty$ if $j > i$. Therefore, position t appends only to $\leq t$, which is suitable for when treating history as an observed prefix before multi-step prediction from the final steps are applied.

Position-wise FFN As per Vaswani et al. [83], over the last dimension d:

$$FFN(z) = W_2\sigma\left(W_1 z + b_1\right) + b_2,$$

Where $\sigma = ReLU$ in our implementation.

Spatio-Temporal Block For each spatio-temporal block, we apply:

$$h \leftarrow LN\left(h + \mathit{Dropout}\left(SpatialMHA(h, M)\right)\right),$$

$$h \leftarrow LN\left(h + \mathit{Dropout}\left(TemporalMHA(h, C)\right)\right),$$

$$h \leftarrow LN\left(h + \mathit{Dropout}\left(FFN(h)\right)\right).$$

Where the block order is as follows: spatial -> temporal -> FFN. Fixing this order at every layer means each block first lets a patch node mix information with its Chapter-3-defined spatial neighbors at the current time step, then lets that already-spatially-mixed representation attend causally across the node's own history, so that by the final block a node's representation reflects both nearby-patch influence and its own temporal trajectory, mirroring at the level of attention what the graph-convolution-then-GRU ordering achieves in ST-GNN.

Decoder Head At the last input time index t = T – 1:

$$z_{b,n} = h_{b,T-1,n} \parallel x_{b,T-1,n} \text{if input skip enabled},$$

$$y_{b,\cdot,n} = W_{out} z_{b,n} \in R^{H_{pred} \cdot d_{out}},$$

and finally reshaped to $R^{H_{pred} \times N \times d_{out}}$.

These would ensure that the forecasts themselves are direct multi-horizon regression from the terminal encoding, following the procedure of multi-horizon sequence models, cf. Lim et al. [36]. Though both models follow different architectures respectively, they follow the same procedure of predicting H steps from the context goal.

Our ST-GTN implementation factorizes spatial mixing and temporal mixing across multiple separate attention stages, a standard in traffic and grid forecasting (e.g. [96]) and instantiates it with explicit patch-graph masking and causal temporal attention. Although it is conceptually related to another Transformer-style graph-time model: STAR [102], which has aimed to forecast pedestrian trajectory prediction through TGConv and external memory, our implementation differs greatly in its architecture.

ST-GTNs show promise for groundwater level forecasting because groundwater dynamics are inherently spatio-temporal: patches of wells influence one another through hydrogeologic connectivity (captured via graph-masked spatial attention), while groundwater responses to drivers such as climate and pumping show delayed and long-range temporal effects (captured via self-attention over time). By embedding graph-structured historical signals into a shared latent space and stacking **Graph-Masked Spatial Attention** with **Causal Temporal Attention**, the model can learn *where* influences propagate across the patch network and *when* they affect future levels. The main trade-off is that attention can be computationally expensive (scaling with both the number of nodes and the time window) and may require sufficient training data, but for complex groundwater systems this joint modeling can provide stronger forecasts than approaches that treat space or time in isolation.

A transformer-based architecture applied directly to groundwater level prediction has already demonstrated the general viability of attention-based forecasting for this variable [78], though without the explicit graph-masked spatial attention used here; combining the two—attention over both space and time, jointly conditioned on groundwater-specific covariates—is one direction the present ST-GTN implementation already moves toward. Because our spatial attention mask M is derived from the same fixed, cardinal patch adjacency used by ST-GNN (Chapter 3), it inherits the same limitation of not adapting to data or learning connections beyond immediate cardinal neighbors; Graph Transformer Networks that learn new graph structures end-to-end [103], alongside the adaptive-adjacency approach of Graph WaveNet already noted for ST-GNN [94], illustrate how the mask M could instead be learned or softened rather than fixed, letting the model discover longer-range hydrogeologic connections between non-adjacent patches at the cost of losing the direct interpretability of a physically motivated, distance-based adjacency. More broadly, recent surveys of spatio-temporal graph neural network and graph transformer variants for time series forecasting [15, 64] indicate that factorized spatial- and temporal-attention designs of the kind used here are now a common pattern across domains, suggesting that further gains for groundwater forecasting specifically are more likely to come from better graph construction and covariate design than from further architectural elaboration of the attention mechanism itself.

Experimental Design

Hyperparameter Fine-Tuning

Prior to proper training, hyperparameter fine-tuning was applied to optimize our models via Optuna, for a total of 20 trails across all three models at maximum. Both ST-GNN and ST-GTN were run for a total of 20 trials while CNN-LSTM was run for a total of 10 trials. For our graph-based models, we observed that it takes about at least 30 seconds for one epoch to run regardless of the hyperparameters set for each trial, compared to CNN-LSTM's varying epoch times that have each epoch take as long as 15 minutes at most. The best hyperparameters are determined based on which trial returns the lowest possible losses and error, or more specifically, the lowest possible Root Mean Squared Error (RMSE).

Model Evaluation Metrics

Multiple evaluation metrics were utilized in evaluating performance across all three of our models. The main evaluation metrics we used during training and testing are as follows:

Root Mean Squared Error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}$$

RMSE would measure the average magnitude of prediction errors, penalizing large error deviations by squaring such deviations made.

Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i - y_i|$$

Measures the absolute difference between the predictions made and the actual observations. Compared to RMSE, MAE is considered more robust to outliers.

Nash-Sutcliffe Efficiency (NSE):

$$\text{NSE} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

NSE would compare the model error variance calculated to the variance of the actual observations. An NSE value of 1 indicates a perfect fit while a negative value indicates the opposite, meaning that simply using an observed mean is a far better predictor than the model. At a value of 0, the model and observe mean predict no better than one or the other.

Mean Absolute Relative Error (MARE):

$$\text{MARE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|p_i - t_i|}{|t_i| + \varepsilon} \times 100\%$$

Symmetric Mean Absolute Percentage Error (SMAPE):

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{2 \cdot |p_i - t_i|}{|p_i| + |t_i| + \varepsilon}$$

Training and Validation

We have conducted training and validation after obtaining the best hyperparameters for each model and rebuilding our data loaders if necessary. For training, our set is split into 17 years-from 2000 to 2016-for all our input data points while our validation set is split to only two years: 2017 to 2018. The number of epochs run during training is dependent on what hyperparameters culminate in the lowest possible error. However, early stopping for each model would remain constant, for about 10 consecutive epochs monitoring validation loss to prevent over-fitting. Our goal here for each model is to observe how well our models read and train on relevant geospatial data and see how well they can predict groundwater depth.

Model Evaluation on Test and Forecast Sets

Our test and forecast sets consists of data covering our final two years: 2019 to 2020. Both sets aim to evaluate how well a model can forecast individual years that are most temporally displaced from the training set.

Forecasting and Quantile-Mapping Bias Correction

To tackle potential biases from our generated forecasts, we implemented Quantile Mapping (QM). A statistical technique intended to reduce biases across forecasts through adjusting outputs such that their distribution would better correlate with the actual observed data. For each forecasted value, the QM would calculate the quantile of said value within some forecast distribution from a given reference period in search for observed values that fit within this same quantile. Through a such a mapping, the adjusted forecasts are ensured to follow the same distributional shape as the observed data rather than by only having the same means.

Given each forecasted value:

$$\hat{y}^{\text{forecast}}$$

QM would perform the following mapping:

$$\hat{y}^{\text{corrected}} = F_{\text{obs}}^{-1}\left(F_{\text{pred}}\left(\hat{y}^{\text{forecast}}\right)\right)$$

Where:

F_{pred} represents the Cumulative Distribution function (CDF) of the model predictions from the test period.

F_{obs}^{-1} is the inverse CDF (quantile function) of the observations from the test period.

$\hat{y}^{\text{forecast}}$ is the forecast value to be bias-corrected (from the forecast period).

$\hat{y}^{\text{corrected}}$ is the bias-corrected forecast value (the output).

First, QM would compute CDFs from across the test period:

$$F\text{pred}(x) = P(\hat{y}^{\text{test}} \leq x)$$

$$F\text{obs}(x) = P(y^{\text{test}} \leq x)$$

Both of which are estimated empirically through sorting and ranking pixel values gathered from the test period raster images.

Next, each forecast is mapped to its respective quantile. In other words, for each forecast pixel value:

$$\hat{y}^{\text{forecast}}i$$

QM would find its matching quantile within the predicted distribution:

$$qi = F\text{pred}\left(\hat{y}^{\text{forecast}}i\right)$$

Thus, satisfying the claim that a forecast value sits at the qi-th percentile of what the model predicted during testing.

Finally, QM would apply the inverse observed CDF at the same quantile obtained:

$$\hat{y}^{\text{corrected}}i = F^{-1}\text{obs}(qi)$$

The key point of QM is that it assumes that the bias structure learned from the test period translates well to the forecast period. Should the forecast represent a climate state or an extreme event beyond what the test distribution has observed, the mapping would extrapolate.

Feature Importance Analysis

To assess the relative contribution of each input covariate to model predictions, we applied a post-hoc feature importance analysis to all three trained models after training was completed. Two complementary attribution methods were used, each capturing a distinct aspect of how the model depends on each input channel.

Permutation Importance

Permutation importance measures the increase in validation loss when a single input channel is randomly shuffled across the batch dimension, breaking the statistical linkage between that channel and the target while leaving all other channels intact [9]. Formally, for a frozen model f with baseline loss $\mathcal{L}_0$, the permutation importance of channel c is:

$$\Delta\mathcal{L}_c = \mathbb{E}_\pi \left[\mathcal{L}(f(x^{(\pi_c)}), y)\right] - \mathcal{L}_0$$

where $x^{(\pi_c)}$ denotes the input with channel c permuted across batch indices and π is drawn from random permutations. A larger $\Delta\mathcal{L}_c > 0$ indicates that the model relies more heavily on channel c; negative values indicate channels whose removal improves or does not change loss, suggesting limited or redundant contribution. Permutation is applied batch-wise across all validation batches and the mean $\Delta\mathcal{L}_c$ is recorded per channel. This method is model-agnostic and requires no retraining, making it applicable uniformly across our CNN-LSTM, ST-GNN, and ST-GTN architectures.

Gradient·Input Sensitivity

For the graph-based models (ST-GNN and ST-GTN), a complementary gradient-based attribution was additionally computed. Gradient·input sensitivity — also known as saliency-weighted attribution [75] — computes the element-wise product of the input tensor and the gradient of the loss with respect to the input:

$$\mathrm{GI}_c = \mathbb{E}\left[\left|\frac{\partial \mathcal{L}}{\partial x_c} \odot x_c\right|\right]$$

where the expectation is taken over spatial and temporal dimensions (B, T, N) for graph-format inputs. Unlike permutation importance, which reflects global reliance across the full validation set, gradient·input sensitivity provides a local measure of attribution at the evaluated batch, capturing the instantaneous sensitivity of the loss surface to each channel. For CNN-LSTM, gradient storage through the convolutional-recurrent backbone was not retained during inference, so gradient·input is reported only for the graph-based models.

Combined Rank and Cross-Model Aggregation

To produce a single unified importance score per channel, we compute a combined rank by first sum-normalising the non-negative values of both metrics independently (so each metric sums to 1 across channels), then averaging the two normalised scores:

$$\mathrm{score}_c = \frac{1}{2}\left(\frac{\max(\Delta\mathcal{L}_c,\, 0)}{\sum_j \max(\Delta\mathcal{L}_j,\, 0)} + \frac{\max(\mathrm{GI}_c,\, 0)}{\sum_j \max(\mathrm{GI}_j,\, 0)}\right)$$

For CNN-LSTM—where gradient·input is unavailable, the combined score reduces to the normalized permutation importance alone. Cross-model combined scores are then averaged across all three models to produce the final channel ranking, providing a consensus measure of feature importance that is not dominated by any single architecture's inductive biases.

Results

Exploratory Data Analysis of Groundwater Dynamics

Seasonal Analysis

Figure 3 presents the spatial distribution of seasonal mean groundwater conditions during premonsoon, monsoon, and winter across the study area, through using the interpolated water table depth data. Maps highlight marked seasonal contrasts in groundwater depth, where lower values (blue shades) indicate a shallow water table and higher values (yellow to red shades) indicate a deeper water table. Across all three seasons, the southern and south-central parts of the area are characterized by relatively shallow groundwater conditions, whereas the western to central zones consistently exhibit deeper groundwater levels. The monsoon season is dominated by lower mean values over much of the region, indicating an overall rise of the water table and broader spatial extent of shallow groundwater conditions, likely associated with seasonal recharge. In contrast, the premonsoon map shows more extensive high-value zones, especially in the western and central parts, suggesting deeper groundwater conditions prior to monsoonal recharge. The winter season displays an intermediate pattern of groundwater conditions being generally shallower than premonsoon in some areas but deeper than monsoon over much of the region. Overall, the figure highlights a clear seasonal response of groundwater levels, with the water table becoming shallowest during monsoon and deepest during premonsoon, while retaining persistent spatial heterogeneity across the landscape.

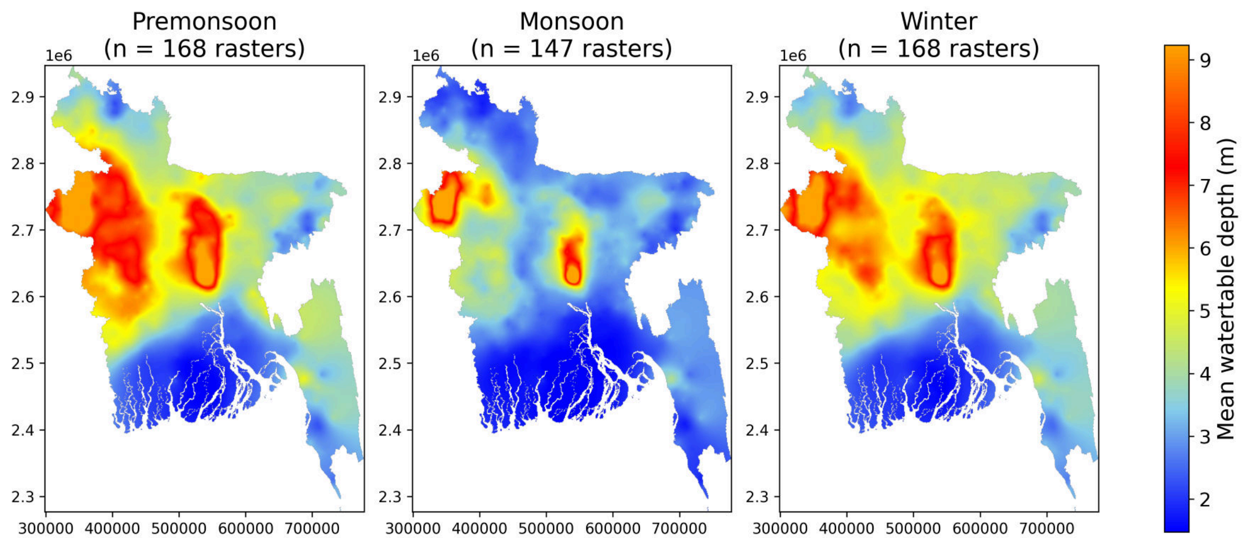


Figure 3: Spatial distribution of seasonal monsoon conditions from field data. The seasons outside of Monsoon showcases the greatest spatial distributions of high mean groundwater depth.

Figure 4 showcases a more detailed overlook of seasonal groundwater conditions, expanding it to a monthly distribution where each map represents groundwater conditions per month. The early months, indicative of the premonsoon season, showcases the highest mean groundwater depth especially for much of the study area. It isn't until June where Groundwater conditions seemingly improve with reductions in mean depth, as groundwater depth decreases for much of the country; the same period where monsoon season resumes. However, what remained consistent during the months was that not only the southern area was rich in shallow water table depths, but also that the Northwestern and Central areas of Bangladesh were dominated by high mean water table values that persist even during monsoon season. Areas of consistently deep groundwater levels.

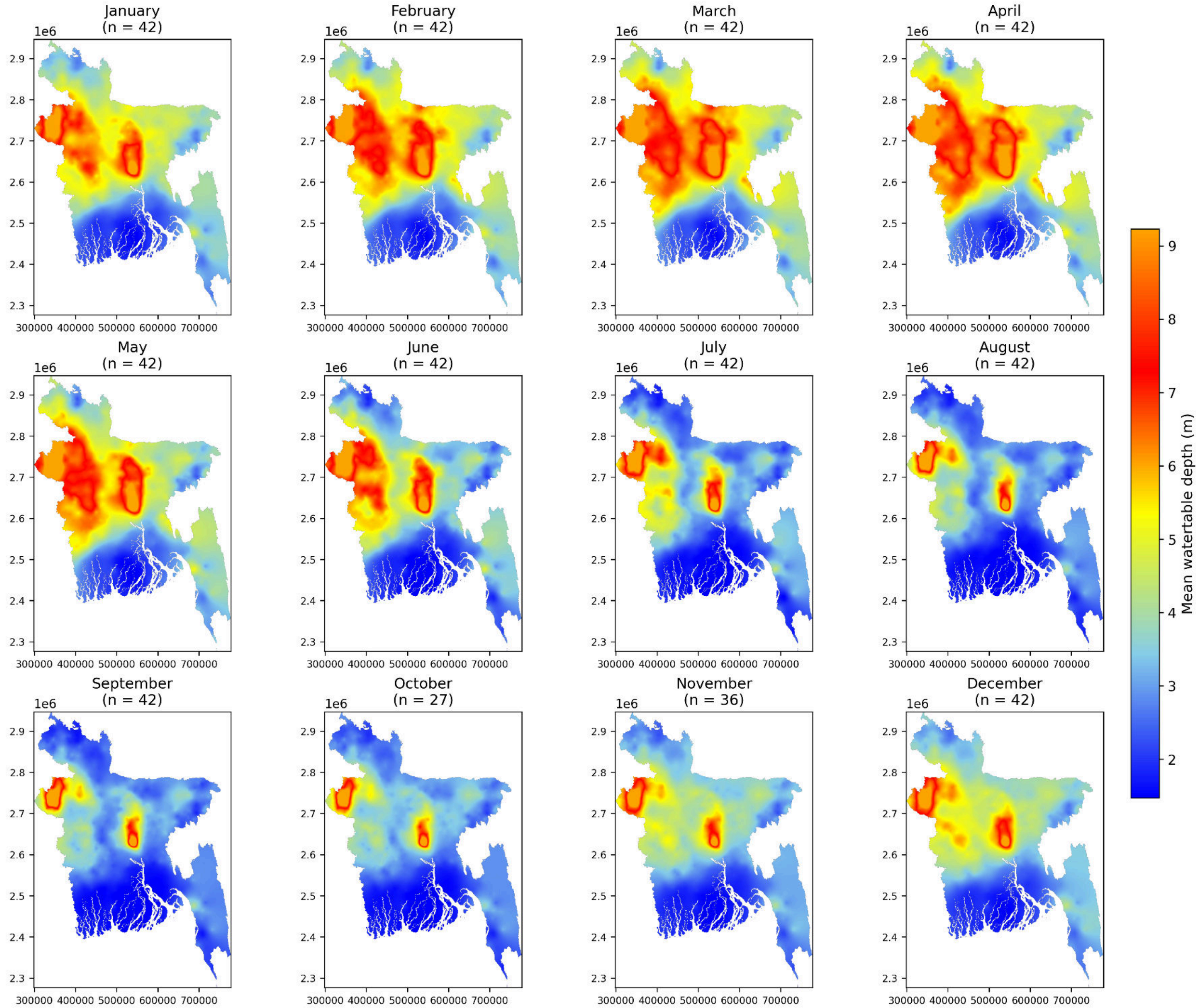


Figure 4: Monthly spatial distribution of seasonal monsoon conditions. Average groundwater depth would drastically decrease starting July.

The seasonal mean groundwater maps also reveal clear spatial and temporal variations in water table conditions across the study area, using our field data. Since lower groundwater values indicate a shallower water table and higher values indicate a deeper water table, pixels falling below the pooled $P25$ threshold represent relatively shallow groundwater conditions, whereas pixels above $P75$ indicate comparatively deep water table conditions (Table 4).

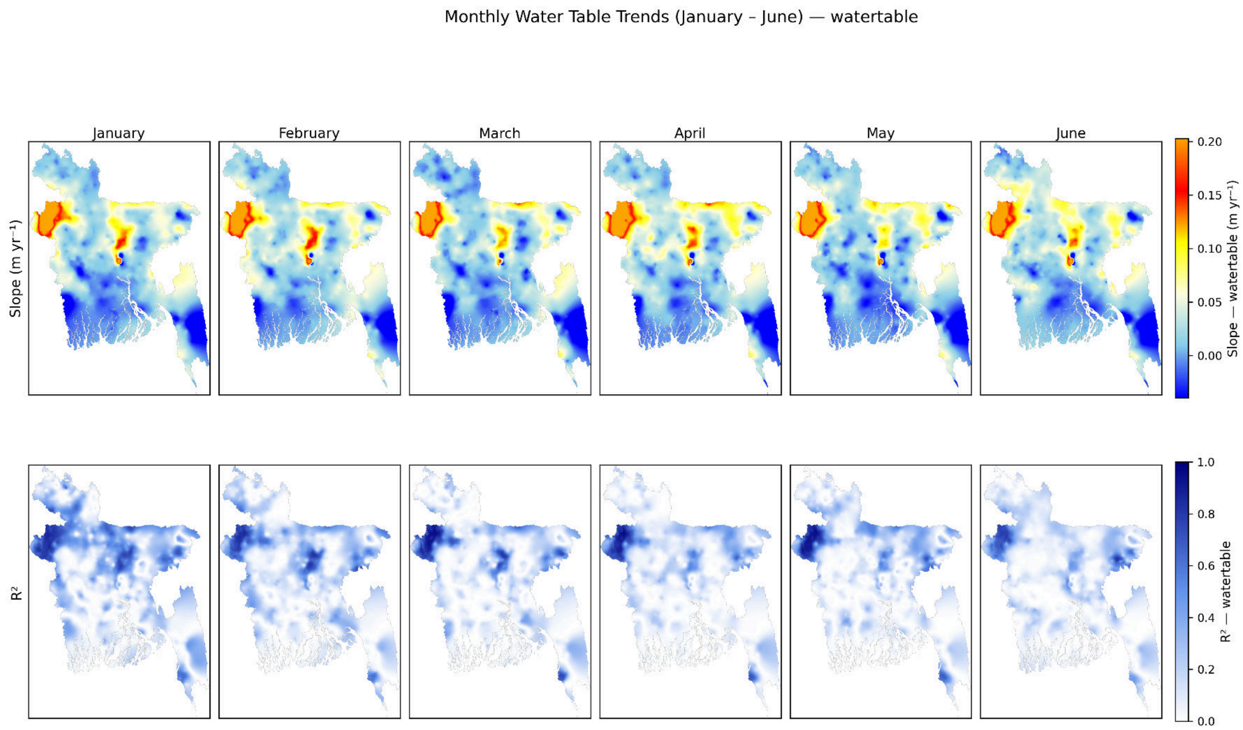
The percentile-based areal analysis has provided a useful basis for comparing seasonal groundwater status. In general, a greater proportion of area within the lower percentile class during a given season reflects a seasonal rise in the water table, while expansion of the upper percentile

Table 4: 4-class Groundwater summary, each value being watertable depth.

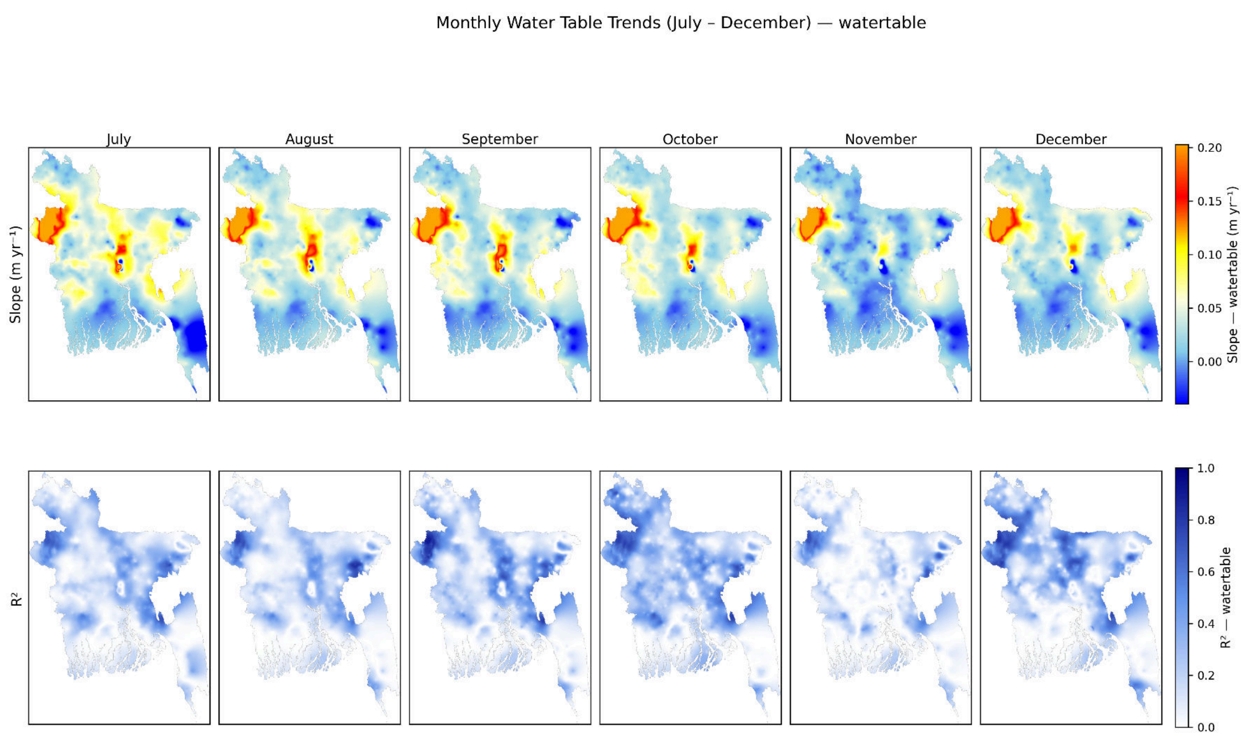
Month	<P25 (%)	P25 – P50 (%)	P50 – P75 (%)	>P75 (%)
January	7.32	7.54	14.85	14.03
February	6.38	5.27	12.95	19.15
March	5.33	4.91	11.67	21.82
April	4.97	5.76	11.99	21.01
May	6.38	6.29	14.89	16.18
June	10.55	11.13	11.06	10.99
July	16.21	14.87	6.57	6.08
August	18.52	15.04	6.66	3.52
September	19.86	15.55	5.39	2.94
October	15.88	18.56	6.57	2.72
November	10.97	14.56	13.39	4.81
December	8.83	11.72	15.22	7.95

class indicates groundwater recession and deepening of the water table. The intermediate class $P25$–$P75$ represents areas with moderate groundwater depth that remain within the central range of seasonal variability, the lower half highlighting greater seasonal rises in the water table across the monsoonal months and vice versa. Overall, the seasonal patterns highlight how groundwater conditions shift between shallow, moderate, and deep states over the annual cycle, with the mapped distributions and area percentages offering a concise summary of seasonal recharge and depletion behavior.

To assess whether the spatial patterns of groundwater depth identified in Figures 3 and 4 reflect stable long-term conditions or directional trends, pixel-wise linear regressions were applied to the seasonal and monthly mean groundwater raster stacks spanning 2000–2020. Figures 5a and 6 presents the monthly and seasonal water table trend maps, respectively, displaying the regression slope (top row) and R^2 (bottom row) for each temporal aggregation.



(a) January – June



(b) July – December

Figure 5: Monthly water table trends (slope and R^2) across Bangladesh for the 2000–2020 study period. Row 1 shows the annual rate of water table depth change (m yr^{-1}); row 2 shows the corresponding coefficient of determination (R^2) indicating trend linearity. Both half-year panels share a common colour scale to allow direct comparison.

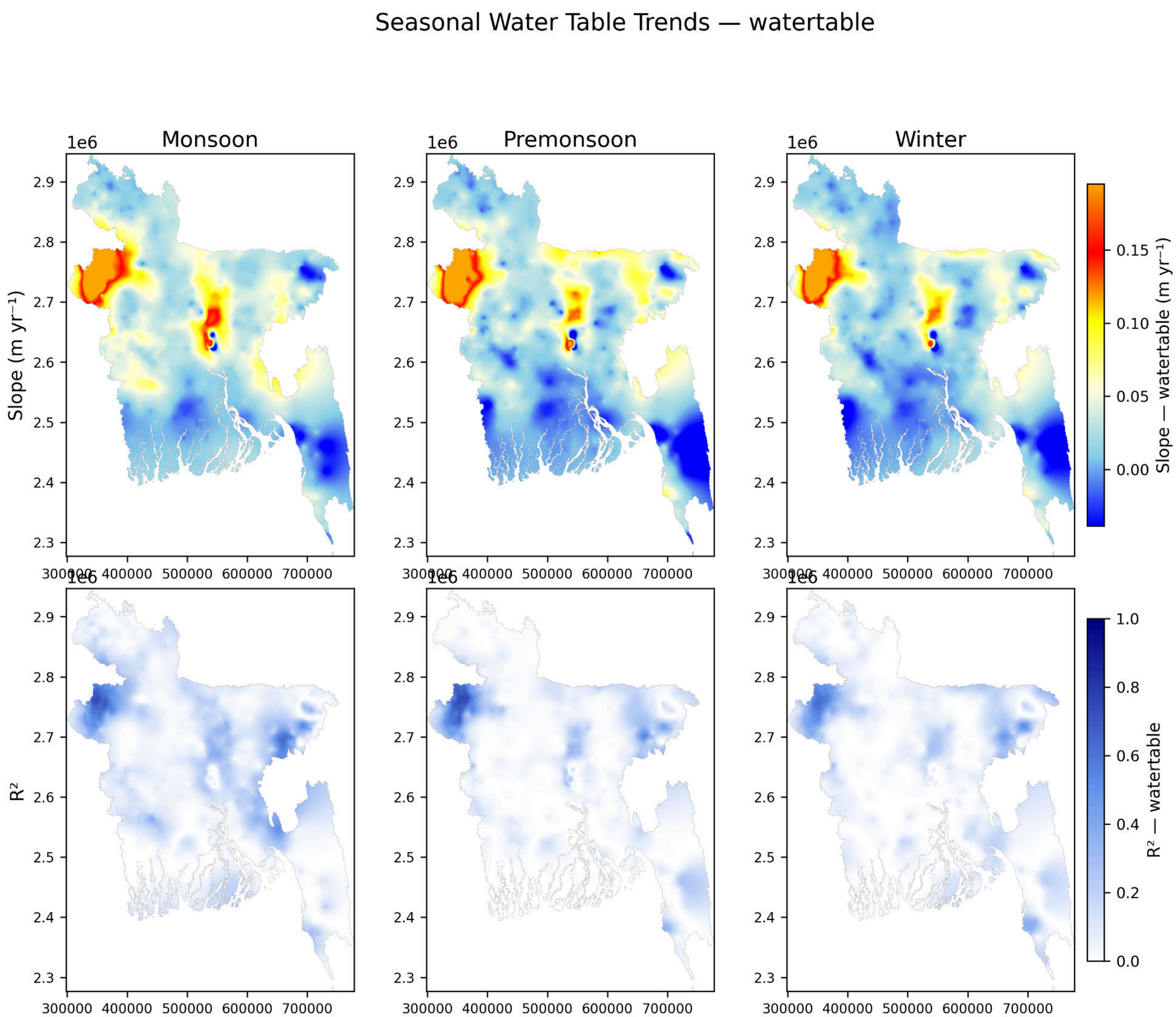


Figure 6: Seasonal water table trend maps (slope and R^2) across Bangladesh for the 2000–2020 study period.

Before interpreting these results, it is worth being explicit about how the two rows of each figure should be read together. The mapped variable is water table *depth* (distance below the ground surface), not elevation, so the sign of the slope is the opposite of what ”improving groundwater” intuitively suggests: a **positive** slope (yellow–red) means depth is increasing year over year, i.e., the water table is getting deeper (depletion), while a **negative** slope (blue) means depth is decreasing, i.e., the water table is rising toward the surface (recovery). The colour scale in both figures extends below 0.00 m yr^{-1}, so the darkest blue shading always represents a true negative slope rather than merely a small positive one. The R^2 row is not itself a trend; it reports, on a 0–1 scale, how much of a pixel's year-to-year variability is explained by a straight line, and should be read as a confidence qualifier for the slope value directly above it. A pixel with a large-magnitude slope but low R^2 has a noisy, non-monotonic trajectory that a line happens to fit loosely, whereas a large-magnitude slope paired with high R^2 (as in the Barind Tract) reflects a trend that is both strong and consistent year over year.

The slope maps expose a spatially heterogeneous pattern of long-term water table change across Bangladesh, dominated overall by deepening rather than recovery. The most pronounced deepening trend is concentrated in the Barind Tract in the northwest, where slope values reach approximately 0.15–0.20 m yr^{-1} across all twelve calendar months and all three seasons. The cross-seasonal consistency of this signal is particularly diagnostic, as even during the monsoon recharge period—when aquifer recovery would be expected if extraction pressure were moderate—the Barind Tract continues to show the highest rate of deepening in the country. This indicates that extraction demand persistently outpaces monsoon recharge in this region, corroborating the interpretation of structurally driven, anthropogenic depletion rather than climatically modulated variability.

Three distinct clusters of negative (blue-shaded) slope are visible and persist across nearly every monthly and seasonal panel: a small, tightly bounded cluster immediately adjacent to the secondary Barind-adjacent hotspot in north-central Bangladesh (around $x \approx 550{,}000$ m); an isolated, single-pixel-scale spot in the northeast (around $x \approx 700{,}000$ m, $y \approx 2.75 \times 10^6$ m); and a substantially larger, contiguous area spanning the southern coastal delta and Sundarbans fringe. The first two are small enough, and consistent enough with a handful of individual monitoring wells, to plausibly reflect either genuine localized recovery or a data artifact from well substitution or gaps in the observation record; neither is interpreted as a regionally significant recovery signal. The larger coastal cluster, however, is too spatially extensive to dismiss on the same grounds, and it is not a data artifact—but it is also not evidence of recharge-driven recovery. As the R^2 row directly below it shows, this coastal region has near-zero R^2 in every panel, meaning a straight line explains almost none of the pixel's year-to-year behaviour there. The apparent negative slope in this region is therefore better read as regression noise riding on a volatile, non-monotonic signal (discussed further below) than as a statistically reliable trend in either direction. In short: no region of Bangladesh shows a negative slope that is both *spatially sustained* and *statistically reliable* (high R^2)—the one large-area cluster that is spatially sustained is precisely the region where the linear model itself is least trustworthy, and should not be read as evidence of recharge-driven improvement.

The R^2 maps offer a complementary characterization of trend linearity, and directly explain the coastal negative-slope cluster flagged above. High R^2 values (0.6–0.9) are spatially concentrated in the northwestern region coinciding with the Barind Tract slope hotspot, confirming that the deepening trend there is not only large in magnitude but also temporally consistent and well-described by a linear model over the study period—a hallmark of sustained, structurally driven extraction rather than episodic or climatically forced fluctuation. In contrast, the southern delta, coastal fringe, and south-central lowlands exhibit near-zero to low R^2 values across all months and seasons, indicating that groundwater dynamics in these regions are more volatile and less amenable to linear characterization, likely reflecting the influence of inter-annual variability in monsoon intensity, tidal and riverine connectivity, and episodic flood recharge on shallow aquifer levels. This spatial dichotomy in trend consistency—high and linear in the irrigation- dominated northwest, low and non-linear in the monsoon- and flood-driven south—has direct implications for model performance, as the regions of low R^2 coincide spatially with areas of reduced forecasting skill identified in the model evaluation results.

Land-Use Reclassification and Change

After applying reclassification to our annual landuse maps using Table 2, Figure 7 highlights the new reclassified landuse maps of Bangladesh at five representative years, selected to illustrate the annual changes across the 2000–2020 period.

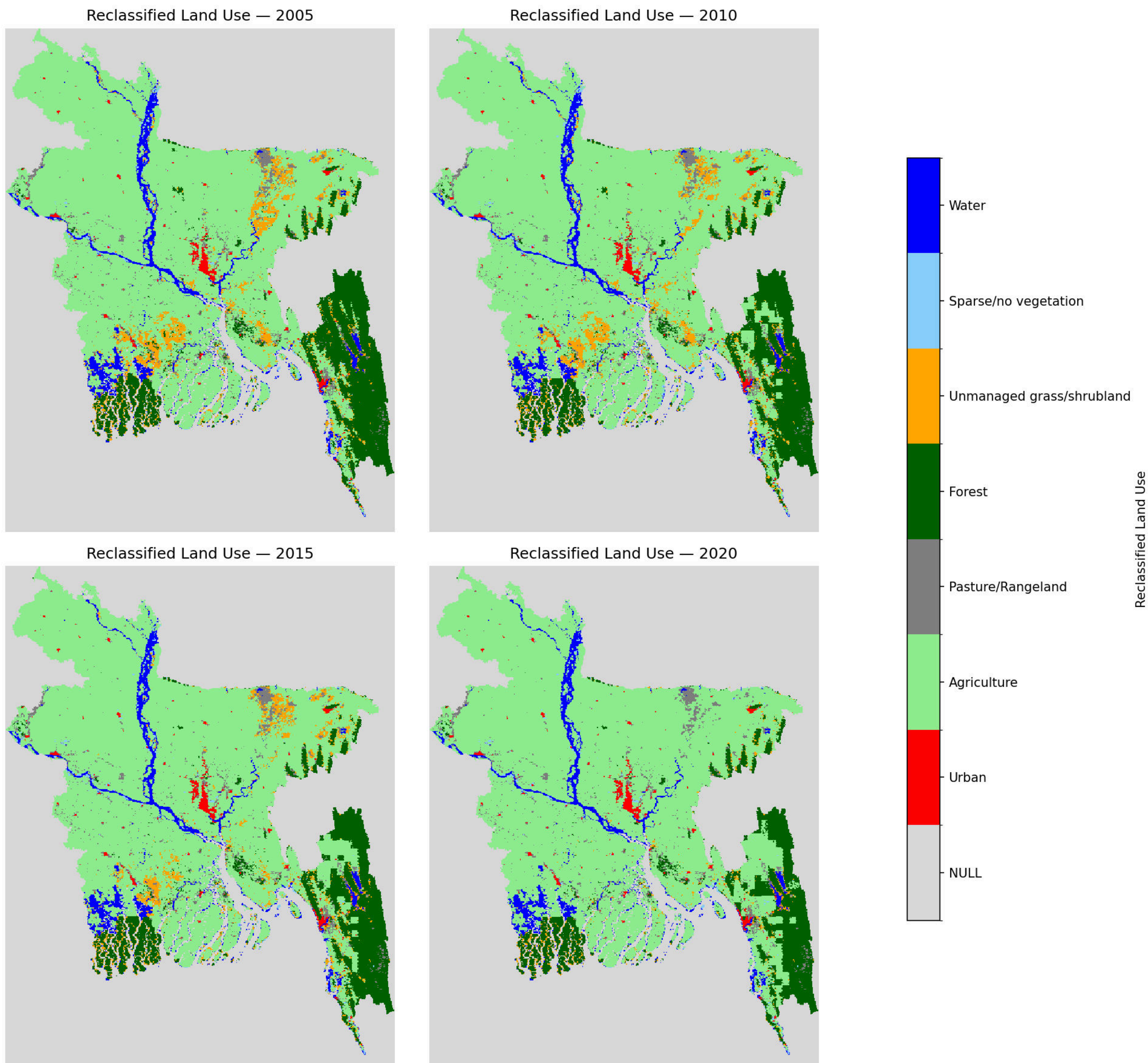


Figure 7: Reclassified landuse maps at five representative years (2000, 2005, 2010, 2015, 2020), drawn from the full annual 2000–2020 classification.

For all 20 years, Agriculture remains undoubtedly the most dominant class for the whole test period, spanning the majority of Bangladesh's arable land surface and has remained spatially stable in its broad distribution. This persistence aligns with Bangladesh's established agricultural economy, namely the long-term expansion of boro rice cultivation; the primary driver of dry-season groundwater extraction throughout the study period. Although comparatively small in fractional coverage, Urban Areas still show a discernible spatial expansion particularly concentrated within the Dhaka metropolitan region and around the Chittagong corridor, reflecting the rapid urbanization previously documented in the introduction. Much of Bangladesh's forest cover is primarily confined to the southeastern Chittagong Hill tracts and the southwestern Sundarbans coastal fringe. Forest coverage within the Chittagong Hill tracts is observed to have lessened significantly across the 20-year period, perhaps indicating potential large-scale deforestation from the expansion of agriculture. Water and Sparse/No Vegetation categories have remained spatially confined to the southern delta and the seasonal floodplains, with Unmanaged grass/shrubland sprinkled primarily across the Northeastern and southern regions of Bangladesh. The latter class visibly disappearing throughout the 20-year period.

To further highlight the annual transitions in land-use, Figure 8 isolates changes in land-use between 2010 and 2015.

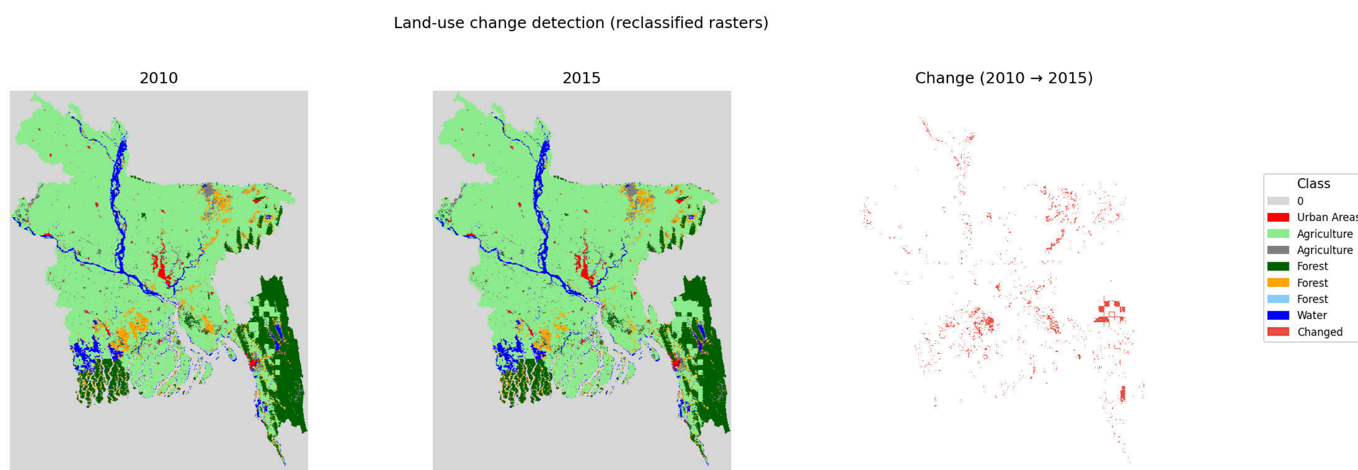


Figure 8: Land-use change detection from 2010 to 2015.

The most dominant transitions in land-use are observed to be Forest-to-Agriculture and Unmanaged Grass/Shrubland-to-Agriculture conversions, especially concentrated across the Chittagong Hill tracts and north-central Bangladesh. These conversions highlight significant expansion of Bangladesh's agricultural economy, as more, previously vegetated or semi-natural land-mass is being claimed for irrigation and cultivation as food supply demand from an expanding population rises. These are notably areas of usually mild or shallow water table depth, which sets a potential precedent for future groundwater consumption, as the replacement of forest canopy and semi-natural vegetation with irrigated cropland reduces interception and increases both surface runoff during the monsoon season and subsurface extraction demand during the dry season. Although present as a minor secondary transition concentrated around Dhaka and Chittagong, Urban expansion is comparatively limited in spatial extent over this 2010–2015 comparison window relative to the vegetation-to-agriculture conversions observed.

Population Trends

Figure 9 shows the results of applying pixel-wise linear trend analysis to the GlobPOP annual population density time series from 2000 to 2020, summarized through two derived raster maps: the regression slope (β_1, left panel) and the coefficient of determination (R^2, right panel).

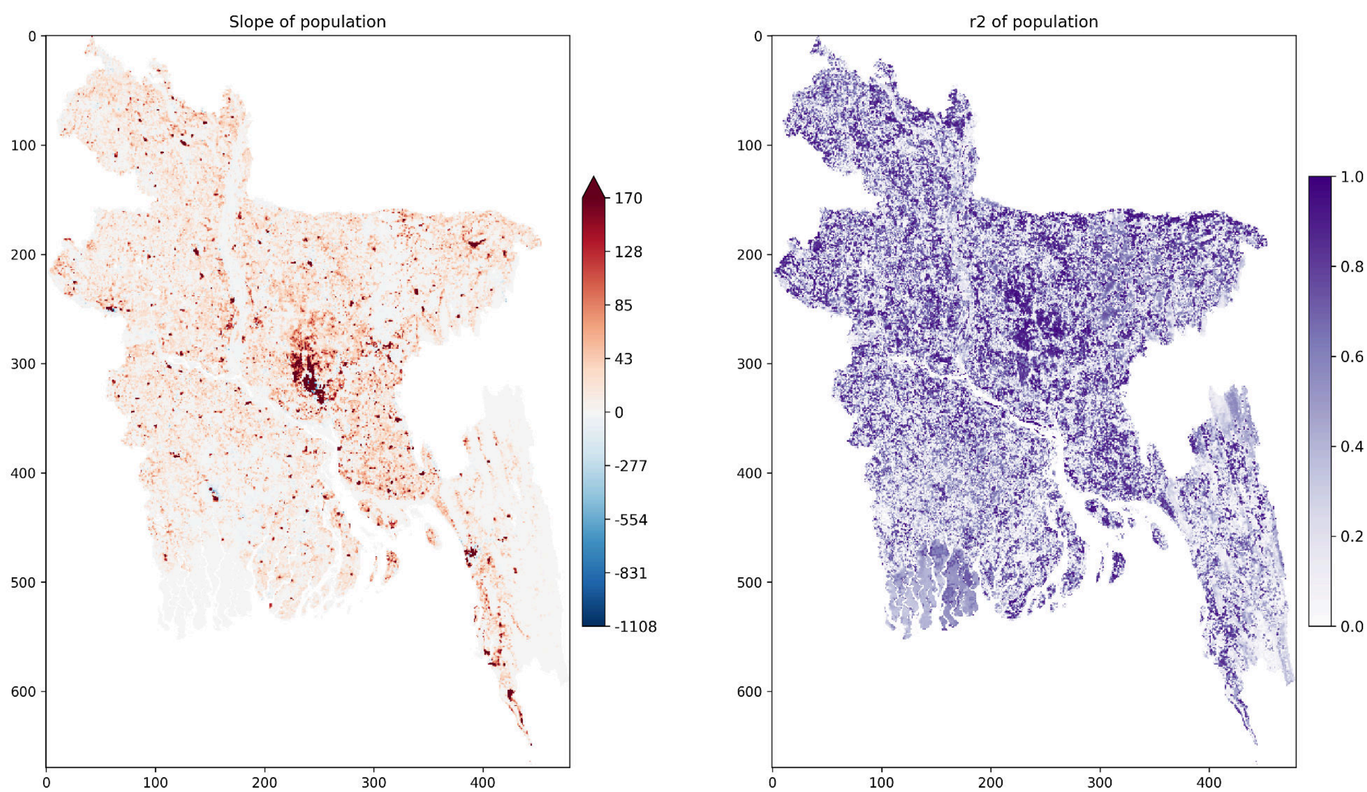


Figure 9: Population analysis through two metrics: slope and R^2.

The slope map reveals a spatially heterogeneous pattern of population change across

Bangladesh. Nationally, 58.3% of pixels show a positive slope (mean = 13.4, median = 4.3), indicating that modest population growth is widespread rather than confined to discrete hotspots. Growth is most concentrated within the Dhaka division and surrounding peri-urban zones, where the mean slope reaches 68.4 (median = 19.0; 99th percentile = 1558), consistent with Dhaka's documented growth to a metropolitan population of approximately 36.6 million by 2025. Sylhet shows a comparable, if smaller-magnitude, pattern of sustained growth (mean = 28.4, median = 17.0, 75% of pixels positive), reflecting Bangladesh's broader trend of multi-node urbanization. Chittagong, by contrast, presents a considerably more mixed signal: although its mean slope (26.6) appears similar to Sylhet's, its median is only 2.9, nearly half of its pixels (46.4%) are negative, and it exhibits the lowest R^2 of any region examined (0.27) — indicating that population change there is poorly described by a simple linear trend and is likely driven by more volatile, non-monotonic dynamics rather than sustained urbanization pressure. In comparison, the Barind Tract (mean = 8.5) and Madhupur Tract (mean = 15.7) both show positive but comparatively modest growth relative to Dhaka, consistent with continued, if slower, population increase alongside possible rural out-migration toward urban centers. Coastal regions in the southwest, near the Sundarbans fringe, show a genuinely negative trend — a mean slope of -0.7 and a median of -2.2, with 79.9% of pixels negative — indicating a real, spatially dominant pattern of population decline rather than a marginal or localized effect. This is consistent with population displacement associated with salinity intrusion, cyclone exposure, and land subsidence along the coastal fringe.

The R^2 map shows a complementary view of population trend linearity, as urban and peri-urban areas that exhibit high slope values also show high R^2 scores. Dhaka (R^2=0.48) and Madhupur (R^2=0.50) show the strongest linear fit, Chittagong the weakest (0.27). This indicates that population growth within these zones follows a consistent, approximate linear trajectory over the two-decade period, signifying sustained urbanization pressure as opposed to episodic fluctuation. In contrast, however, several coastal and low-lying areas display lower R^2 scores despite their respective non-negligible slope values. Within these areas, population dynamics may be far more volatile and less well described by only using a simple linear model, which may indicate influence of repeated flood events and climate-driven displacement on demographic patterns.

Through a groundwater perspective, this population trend analysis reinforces the spatial narrative established in the land-use and groundwater EDA results. By combining high agricultural land-use coverage, moderate yet sustained population density, and intensive dry-season irrigation demand in the Barind and Madhupur Tracts, deep water table anomalies in Figures 3 and 4 are able to be distinguished, even if these areas in particular may not exhibit the highest rates of population growth in absolute terms. The rapidly urbanizing Dhaka corridor—where population slopes are at their highest—is also associated with a distinct groundwater stress mechanism, being the concentrated extraction from deep tubewells to supply a growing city. Our EDA results indicate that groundwater depletion in Bangladesh is primarily driven by two partially distinct pressures operating across different regions: diffuse agricultural extraction concentrated in the northwest, and concentrated urban extraction in and around Dhaka. Both represent targets for the forecasting models we've developed in the subsequent sections.

Summary of Exploratory Findings

Taken together, the groundwater, land-use, and population exploratory analyses establish spatial and temporal context within which our deep learning models are trained and evaluated. Groundwater depth across Bangladesh follows a consistent seasonal cycle driven by monsoon recharge. And yet this seasonal signal is otherwise overlaid by persistent spatial anomalies—namely the deep water table hotspots of the Barind and Madhuper tracts—that are sustained year-round through anthropogenic extraction pressure. Land-use analysis confirms that the hotspots coincide with the most continuously agricultural regions of the country, whereas population trend analysis identifies Dhaka and its peri-urban fringe as a secondary zone of distinct extraction-driven stress. These findings motivate the inclusion of land-use and population density as covariates in our spatiotemporal models, providing a spatial reference frame against which model performance, namely the systematic underestimation of Barind Tract depths, is interpreted in the results and discussion that follow.

The trend analysis further reveals that water table deepening in the Barind Tract is not only spatially persistent but also temporally monotonic across the full 2000–2020 study period, with no month or season showing evidence of sustained recovery, underscoring the urgency of targeted groundwater management interventions in this region.

Data Loading

Data-loading performance and time has varied across the three models given the separate workspaces said models are run across individually, since each model is trained in its own workspace. Data-loading for our CNN-LSTM workspace took the least amount of time, at about 8.27 minutes total, contrary to data-loading within our ST-GTN workspace taking 8.88 minutes total. Interestingly, despite our dataset remaining constant across our data-loading, normalization for our CNN-LSTM workspace took the least amount of time to run, substantially less than our graph-based models. Likewise, as the Groundwater Dataset for CNN-LSTM doesn't require us to significantly transform our dataset format, only having to split our imagery into sliding-window patch samples, it only took 1.7 seconds to run, compared to the 2.5 seconds taken to convert our dataset into a Groundwater Graph structure of 211 nodes for both our Graph models. Even with the different times recorded for both our ST-GNN and ST-GTN workspaces, in ideal conditions, we expect them to run at the exact same time periods. The full stage-by-stage data-loading timing breakdown for all three models is reported in Table 7 of Appendix B.

Hyperparameter Fine-Tuning

Hyperparameter fine-tuning was applied to optimize our models via Optuna, for a total of 20 trials for both ST-GNN and ST-GTN, and 10 trials for CNN-LSTM. The best hyperparameters for each model were determined based on which trial returned the lowest validation Root Mean Squared Error (RMSE). The complete hyperparameter table, including the search space explored for each variable per model and the resulting best-performing values, is reported in Table 6 of Appendix A.

Model Training and Convergence

Training and validation has thus been conducted through our best hyperparameters for each of our three models. Compared to each Optuna trial during hyperparameter fine-tuning, actual training has consistently taken much longer for each epoch irrespective of what model is trained, which is especially reflected in CNN-LSTM. Figure 10 highlights the total training time across our three models.

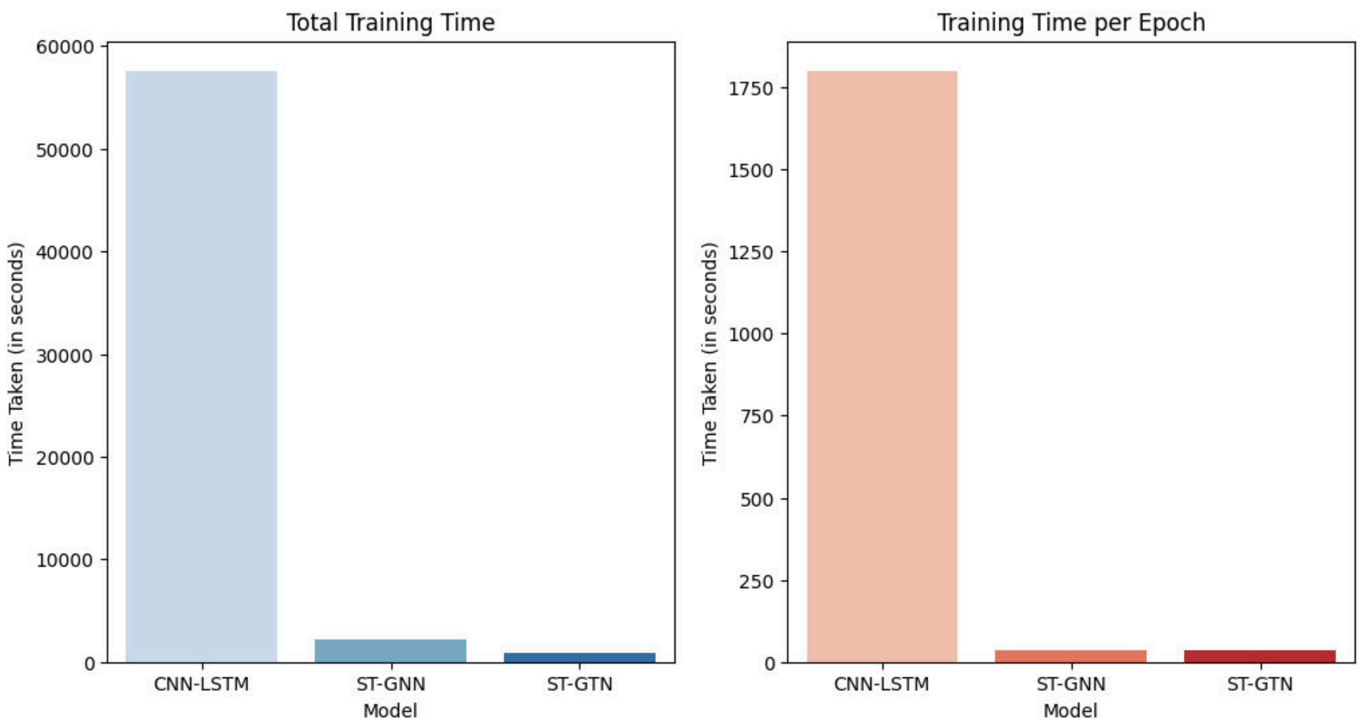


Figure 10: Total training time and training time per epoch for CNN-LSTM, ST-GNN, and ST-GTN. CNN-LSTM scores the highest total training time overall, almost reaching a total time of 16 hours at most, with an average epoch time of 30 minutes.

Compared to the longest recorded epoch time during the Optuna fine-tuning trials being 15 minutes at most, actual training per epoch of the CNN-LSTM model has taken an average of approximately 30 minutes, which, combined with the 33 epochs run culminates with a total approximate training time of 59400 seconds, or 16.5 hours. A dramatic step-up from the total training times recorded for ST-GNN and ST-GTN models, both models recording otherwise identical training times per epoch, at about approximately 35 seconds on average. This we can attribute to ST-GNN and ST-GTN initially recording 81,036 and 628,124 trainable parameters, respectively, whereas CNN-LSTM has initially recorded about 10,608,134 trainable parameters, as well as the Groundwater Graph dataset being lighter in scope in comparison to the Groundwater dataset.

After training has been conducted, history.npz files are generated covering the entire training history of each model, down to the number of epochs and the loss and error metrics calculated per epoch. Rather than presenting each model's training curves as separate figures, we generated a single combined figure overlaying all three models on each metric panel, allowing for direct epoch-by-epoch comparison of convergence behavior across architectures.

Reported RMSE values in the combined training-curve figure are computed on z-score normalized water table depth (standardized using the training-year mean and standard deviation, y_{std}), rather than on denormalized physical units (meters); the normalized RMSE is thus equivalent to the physical-unit RMSE divided by y_{std}.

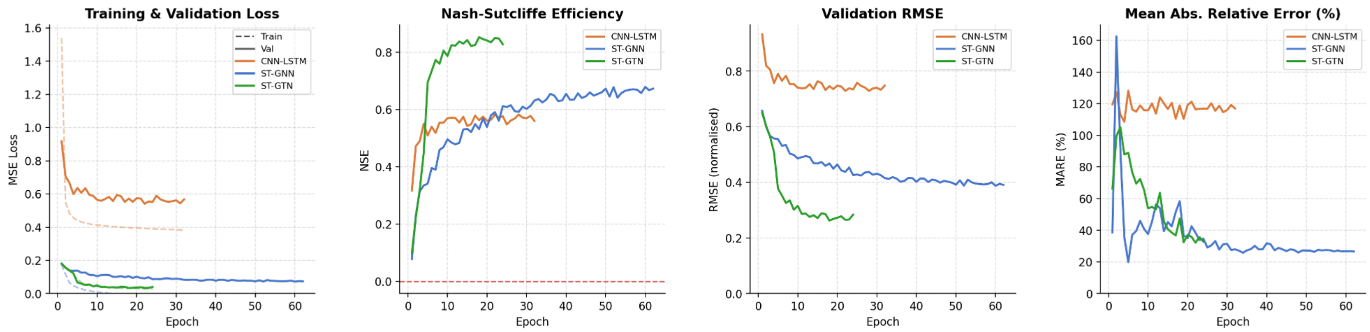


Figure 11: Training and validation metric curves for CNN-LSTM (orange, 33 epochs), ST-GNN (blue, 63 epochs), and ST-GTN (green, 23 epochs). Panel 1 (MSE Loss): dashed lines denote training loss and solid lines denote validation loss for each model; CNN-LSTM converges to the highest validation loss (0.57), while ST-GNN and ST-GTN converge to much lower and closely-tracked training/validation losses (0.07–0.08 and 0.04–0.05, respectively). Panel 2 (NSE): ST-GTN rises fastest and plateaus highest (0.84), followed by ST-GNN, which continues climbing gradually to 0.68 by epoch 63 without fully saturating, and CNN-LSTM, which plateaus earliest and lowest (0.57); a dashed red baseline (NSE = 0) is shown for reference. Panel 3 (Validation RMSE): CNN-LSTM starts highest (0.93) and decreases to 0.74 with persistent oscillation; ST-GNN decreases smoothly and monotonically from 0.65 to 0.40; ST-GTN decreases fastest and furthest, from 0.65 to 0.27 by epoch 23, the lowest final RMSE of the three. Panel 4 (MARE): CNN-LSTM oscillates within a narrow 110–130% band throughout training; ST-GNN spikes early (to 160%) before settling into a noisier decline toward 20–30%; ST-GTN spikes early (to 100%) then declines more steadily to a comparable 20–30% by its final epoch.

Figure 11 covers the training curves of all three models across their respective training runs. CNN-LSTM's training loss converges smoothly while its validation loss plateaus around 0.57 starting from an early epoch, and the persistent gap between the two suggests mild overfitting: the CNN encoder has learned patch-specific spatial textures that have not generalized well across unseen wells. Its NSE rises rapidly early in training before plateauing at its lowest value among the three models (0.57); while positive, and thus outperforming the simple mean-flow baseline, this confirms the model struggled to resolve fine-scale groundwater heterogeneity from patch-averaged CNN features alone. Its validation RMSE declines from 0.93 to 0.74 with persistent oscillation throughout, an instability consistent with sensitivity of the CNN-LSTM decoder to batch composition, especially for well data gathered from sparsely sampled Chittagong Hill tracts and coastal delta regions. Its MARE remains within a comparatively narrow 110–130% band throughout training, and the lack of any decreasing trend after the first few epochs confirms the model has reached its relative-error capacity given its current architecture and training regime. Among the three models, CNN-LSTM attains the smallest NSE and the highest normalized RMSE, making it our weakest baseline; although its pixel-level spatial decoding is well-suited for spatially-dense predictions, its lack of an explicit graph structure has limited neighborhood information sharing, which both ST-GNN and ST-GTN address via adjacency-based aggregation.

ST-GNN, by comparison, exhibits a far smaller train-validation MSE gap throughout its training than CNN-LSTM; validation loss tracks training loss closely as both decrease smoothly, suggesting that graph-based neighborhood aggregation generalizes well across the node set rather than memorizing patterns specific to individual patches. Its NSE climbs steeply early in training before continuing a slow, unsaturated improvement toward 0.68 by epoch 63, outperforming the CNN-LSTM baseline substantially and confirming that graph convolution over the well adjacency structure effectively captures spatial correlations in groundwater levels across Bangladesh. Its normalized RMSE decreases steadily from 0.65 to 0.40 without the marked oscillation seen in CNN-LSTM, reflecting the stability conferred by the symmetric normalized adjacency, which damps gradient variance by averaging messages from neighboring wells at each timestep. Its MARE spikes earliest and highest among the three models before settling into a noisier decline, consistent with large relative errors at wells bearing near-zero groundwater anomalies while the GRU hidden state warms up early in training; its comparatively rapid stabilization confirms the GRU has learned a useful hidden state representation within a modest number of epochs. Because ST-GNN's NSE and RMSE curves show no clear sign of saturation by epoch 63, additional training or scheduling may yield further gains, suggesting that its graph-convolution stack retains unused capacity for feature extraction.

ST-GTN converges fastest and most tightly of the three: both its training and validation MSE losses converge to within 0.04–0.05 of one another with near-perfect alignment throughout, and its NSE rises steeply within the first several epochs to plateau at 0.84, the highest of any model. This rapid, high-efficiency convergence is attributable to the transformer's attention mechanism, which simultaneously attends to all spatially connected nodes and all T timesteps jointly, capturing long-range dependencies that neither CNN-LSTM nor ST-GNN can model as efficiently, and avoiding the sequential bottleneck of purely recurrent architectures. This tight train-validation alignment is consistent with a well-regularized architecture rather than one memorizing training-set noise; it should, however, be read precisely: every well location remains present in the graph throughout training, with only future timesteps withheld for validation, so this alignment demonstrates strong temporal extrapolation at already-known locations rather than spatial generalization to held-out well locations. Section 6.6 returns to this distinction and to a proposed regional hold-out test that would probe spatial generalization directly. Its normalized RMSE reaches 0.27 by epoch 23, roughly a third lower than ST-GNN's final RMSE, driven by the dual-attention architecture in which spatial attention enforces graph topology while temporal attention captures the autoregressive dynamics of seasonal groundwater fluctuations across all timesteps simultaneously. Its MARE spikes early before declining to a level comparable to ST-GNN's by the end of training, with brief transient oscillations that confirm the optimizer's capacity to re-stabilize without permanently degrading the loss landscape.

Across all four metrics, ST-GTN consistently outperforms CNN-LSTM and ST-GNN despite training for the fewest epochs (23, compared to 32 for CNN-LSTM and 63 for ST-GNN), establishing it as the recommended architecture for operational groundwater forecasting from remote-sensing stacks. By combining graph-masked spatial attention, causal temporal attention, and learnable positional encodings, ST-GTN provides a principled framework for capturing both the spatial graph structure and the temporal autoregressive dynamics of Bangladesh's groundwater system.

Model Performance: Validation and Testing

After running training on our three models and load them, we then ran them against our two-year validation and one-year test sets for evaluation. Table 5 displays our ground truth values across these two periods for our three models.

Table 5: Performance evaluation of spatio-temporal deep learning models for Watertable Prediction across Validation (2017-2018) and Test (2019) periods

	Validation (2017-2018)			Test (2019)		
Metrics	CNN-LSTM	ST-GNN	ST-GTN	CNN-LSTM	ST-GNN	ST-GTN
RMSE (m)	0.7289	0.338	0.2626	0.9776	0.5796	0.3766
MAE (m)	0.5992	0.261	0.1775	0.8984	0.3989	0.2635
NSE	0.5825	0.6782	0.8526	-0.1565	0.034	0.5922
$R\hat{2}$	0.7413	0.7443	0.8537	0.7416	0.0594	0.6865
PBIAS	-816.32%	66.88%	50.58%	-295,66%	-64.69%	-113.70%
SMAPE	157.76%	85.42%	173.73%	173.79%	150.74%	96.36%
MARE	116.41%	2635.55%	4742.13%	237.01%	1973.59%	3611.55%
ACC	4.43%	10.06%	16.21%	1.60%	1.30%	10.19%

We see that our ST-GTN model has consistently performed the best across most our core evaluation metrics for both evaluation periods. Achieving the lowest RMSE (0.263) and MAE (0.178), as well as the highest NSE (0.853) and $R\hat{2}$ (0.852) suggests strong predictive accuracy and explained variance for the model. These advantages persist even into the test period, ST-GTN leading in RMSE (0.377), MAE (0.264), NSE (0.592), SMAPE, (96%) and ACC (10.2%).

CNN-LSTM has performed the weakest overall in comparison, as it consistently registers the highest error values. Not only did it observe the worst RMSE and MAE in both periods, but it has also attained a negative NSE in the test set (-0.157). Having a negative NSE suggests that predictions made are less than favorable than by simply using a mean, indicating poor generalization. In addition, its PBIAS for both periods is also observed to be significantly inflated (-816.3% and -295.7% in testing), which indicates systematic underprediction or bias.

On the other hand, ST-GNN has performed moderately, outperforming CNN-LSTM for most error metrics, yet falling behind to ST-GTN. However, it has achieved the best PBIAS percentage during the test period (-64.7%) and has observed a competitive SMAPE during the validation period (85.4%). Yet, its $R\hat{2}$ value dramatically declines on the test set (0.059), which raises concerns about its generalization.

Interestingly, the MARE metric displays an anomaly. Here, the CNN-LSTM model has performed the best on this metric (116.4% / 237.0%), while our best model, ST-GTN, has lagged behind (4742.1% / 3611.6%). This discrepancy may be a result of graph-based models being disproportionately affected by near-zero observed values, which highlights the issue of using relative error metrics in low-flow conditions.

Overall, compared to our other two models, ST-GTN has demonstrated the most robust and balanced period across both periods, and thus, sets itself apart as the most suitable model for our prediction tasks. Yet, the model does warrant further fine-tuning especially in low-flow predictions, considering the high MARE values it calculated.

Predictions on Test Year (2019)

We have accumulated and generated maps covering the monthly and seasonal distributions from our model predictions to highlight spatial differences between our results and what was already observed.

CNN-LSTM

Figure 12 displays both the observed and predicted groundwater watertable depths–number of meters beneath the surface–across three hydrological seasons throughout our validation and test sets. In the form of an 18-panel matrix covering spatially interpolated groundwater depth maps representing Bangladesh and organized by hydrological season (Pre-monsoon, Monsoon, Winter/Dry). The color scale spans from deep blue, representing shallow water table depth around 2m, through green-yellow-orange to a red hue, representing deep water table depth at around 9m. Warmer tones of color indicate deeper groundwater depth beneath the surface of Bangladesh.

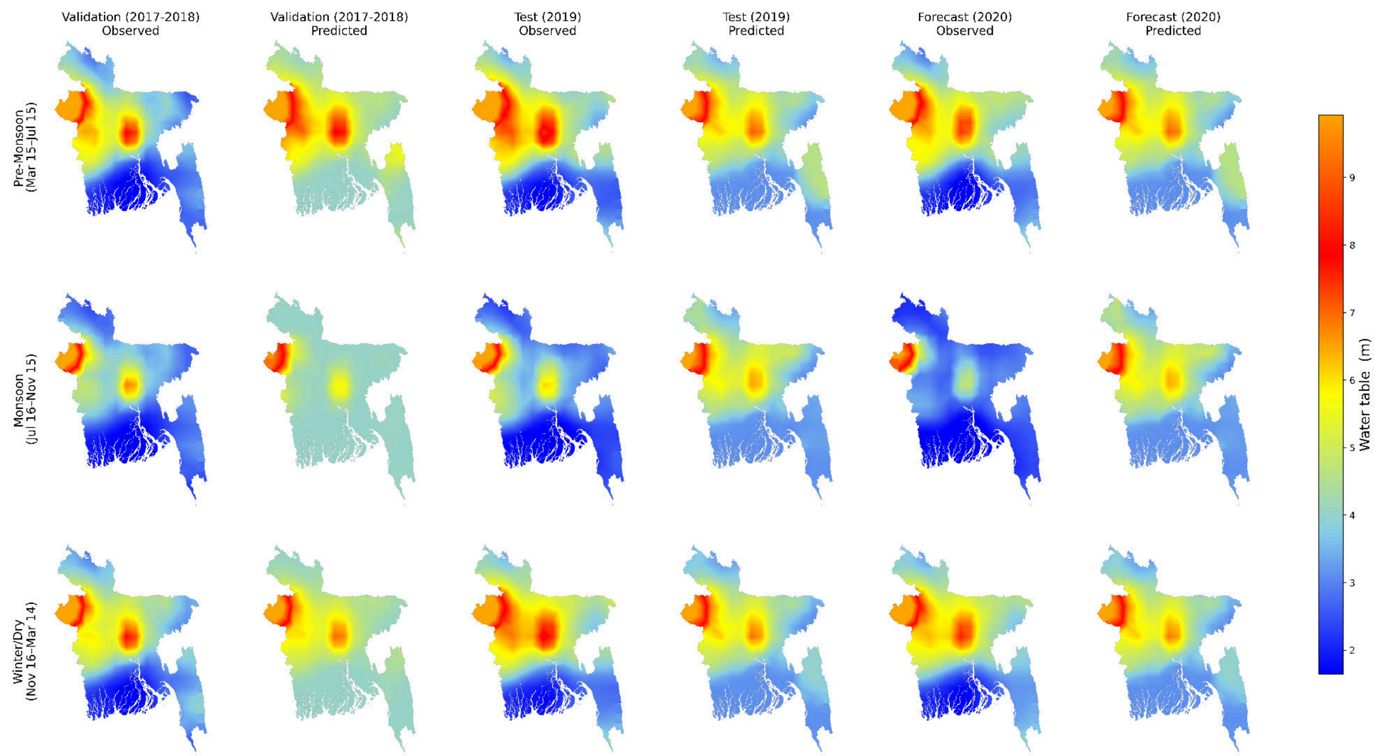


Figure 12: Spatially distributed seasonal groundwater depth (m) predicted by our CNN-LSTM model across the validation (2017-2018) , test (2019), and forecast (2020) periods. Maps reflect seasonal changes across the Pre-Monsoon (March 15 – July 15), Monsoon (July 16 – November 15), and Winter/Dry (November 16 – March 14) seasons and Bangladesh, each with their respective observed and predicted water table depth maps.

Deep watertable anomalies remain persistent across every season and the years they span. Deep watertables (red/orange hue, at 10-12m) are concentrated mainly within the northwestern region of Bangladesh, corresponding the Barind-tract, known to be an intensively irrigated high-land region. A hotspot for excessive groundwater extraction for Boro rice cultivation, driving water table depths to exceed 8-9m on average. A secondary hotspot of deep water table depth is also observed within north-central Bangladesh, associated with the Madhupur tract near the Dhaka city region.

Shallow water tables (blue hue, 2-4m) are prevalent across the southern and southeastern lowlands, which encompass the Ganges-Brahmaputra-Magna delta and the Sundarbans coastal fringe. Their spatial distribution is as such that they exhibit qualities of regional aquifer characteristics and recharge gradients especially given their consistently shallow water tables of about 2-3m. These observations indicate high recharge rates and a low topographic relate.

The pre-monsoon period (March-July) has observed the deepest water table depth across Bangladesh, notably within the north-west region, perpetuated by excessive dry-season irrigation pumping depleting unconfined aquifers ahead of the monsoonal recharge period. The monsoon period (July-November) displays widespread shallowing of watertable depth (I.E. increased blue hue coverage) from intense monsoonal precipitation, recharging the alluvial aquifer. The Winter/Dry (November-March) seasonal periods exhibits intermediate depths, hotspots re-intensifying from post-monsoon irrigation for winter crops. This reinstates the characteristic spatial gradient between the irrigated north-west and the far wetter south.

Figure 13 showcases how well the CNN-LSTM model reproduces the dominant observed spatial patterns during the validation period (2017-2018). The model was able to correctly localise the north-western Barind Tract hotspot and the shallow southern delta scale, albeit existing spatial smoothing and the underestimation of peak depths. These qualitative observations are taken and transformed into more spatially explicit quantitative metrics.

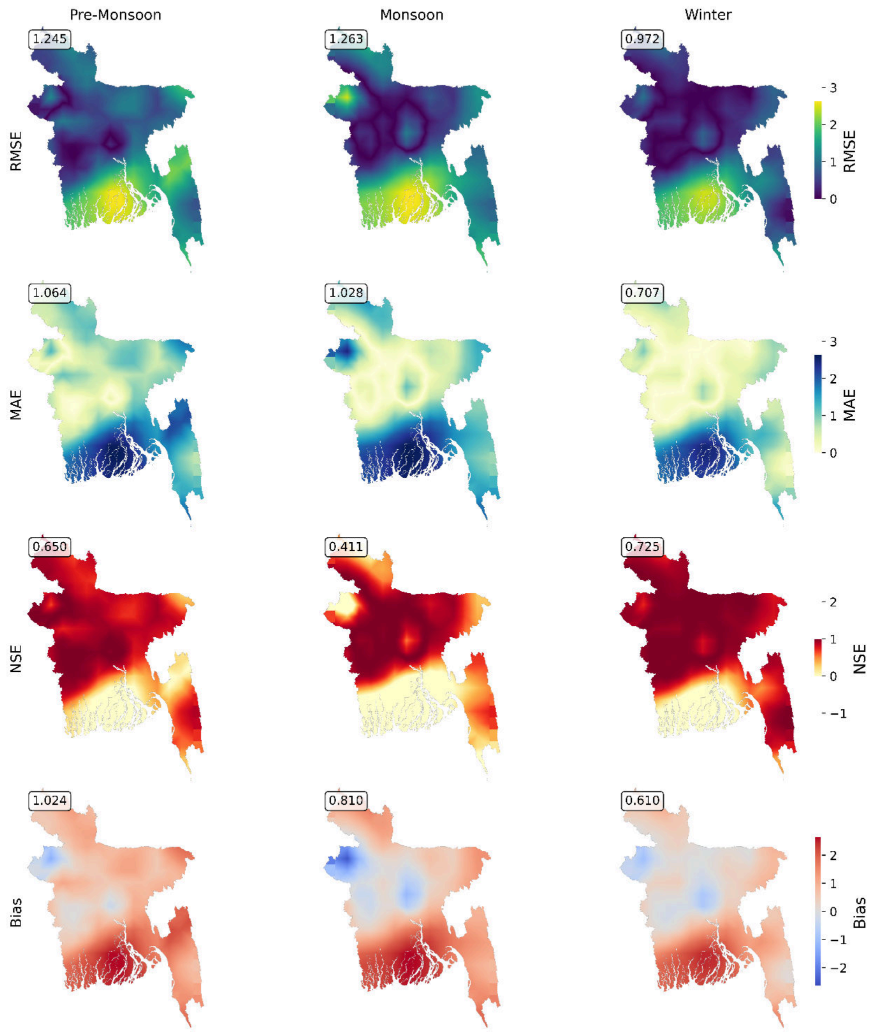


Figure 13: Spatially distributed seasonal validation performance metrics (RMSE, MAE, NSE, and Bias) of our CNN-LSTM model across three hydrological seasons evaluated over the validation period. Spatially averaged metric values are annotated over each panel. Color scales: RMSE and MAE (0-3 m, yellow-green = high error, dark blue=low error); NSE (dark red = high skill approaching 1, yellow-white = near-zero skill, dark maroon = negative skill); Bias (red = positive/over-prediction, blue = negative/underprediction).

Elevated validation RMSE values of 1.245 m for Pre-monsoon, 1.263 m for Monsoon, and 0.972 m for Winter/Dry periods are observed to be spatially concentrated within the south-eastern coastal zone and the transition zones adjacent to the Barind tract hotspot. In particular, the regions showing the highest discrepancies between observed and predicted color fields as denoted by Figure 12.

The concentration of error along the coast is, at first glance, counter-intuitive: a purely climate-driven system might be expected to show comparatively low variability near a large, moderating water body, making it an easier target to fit. Three factors likely combine to produce the opposite result here. First, monitoring well density is markedly lower along the Sundarbans coastal fringe and the Chittagong Hill tracts than elsewhere in the BWDB network (Chapter 3), so the kriged ground-truth surface that CNN-LSTM is trained against is itself less spatially constrained in this region, carrying more interpolation uncertainty into the training target than in the densely-monitored northwest. Second, coastal groundwater dynamics are governed by tidal exchange, salinity intrusion, and riverine connectivity rather than the monsoon-recharge and irrigation-extraction processes that dominate elsewhere in Bangladesh and that our covariate set (land use, temperature, NDVI, precipitation) is best suited to represent; the long-term trend analysis in Section 5.1 (Exploratory Data Analysis) independently corroborates this, showing that the coastal fringe is exactly where the pixel-wise linear regression achieves its lowest R^2 nationally—i.e., where groundwater behaviour is least amenable to being explained by any smooth, low-order model, let alone a covariate set built around monsoon and irrigation drivers. Third, and specific to CNN-LSTM's architecture, each patch is decoded independently with no explicit mechanism for sharing information across neighbouring patches (Chapter 3); in a well-sparse region this leaves the model comparatively little local signal to draw on within a given training sample, unlike the graph-based models discussed in the ST-GNN results below, which propagate information from better-observed neighbouring nodes through their adjacency structure.

The maps covering MAE (1.064, 1.028, and 0.707 m respectively) show the highest absolute errors within the central plains during the Monsoon period, which align with the over-smoothed recharge-driven gradient visible in Figure 12. Overwhelmingly positive bias (mean values of 1.024, 0.810, and 0.610 m across the three seasons) persists over the Bias maps, indicating systematic over-prediction implied in Figure 12, but not measurable through just Fig. 12 alone. Through what the NSE maps have measured (0.650, 0.411, and 0.725), model skill is at its lowest during the Monsoon season, the most difficult season for CNN-LSTM to reliably resolve the sharp north-south depth gradient between the recharging central plains and the otherwise deep north-western irrigation zones.

Figure 14 transitions from the validation period to then evaluate the independent test set (2019). Evaluation on said test has revealed a significant, yet counter-intuitive result, that the CNN-LSTM model generalizes better to 2019 than to the validation period (2017-2018).

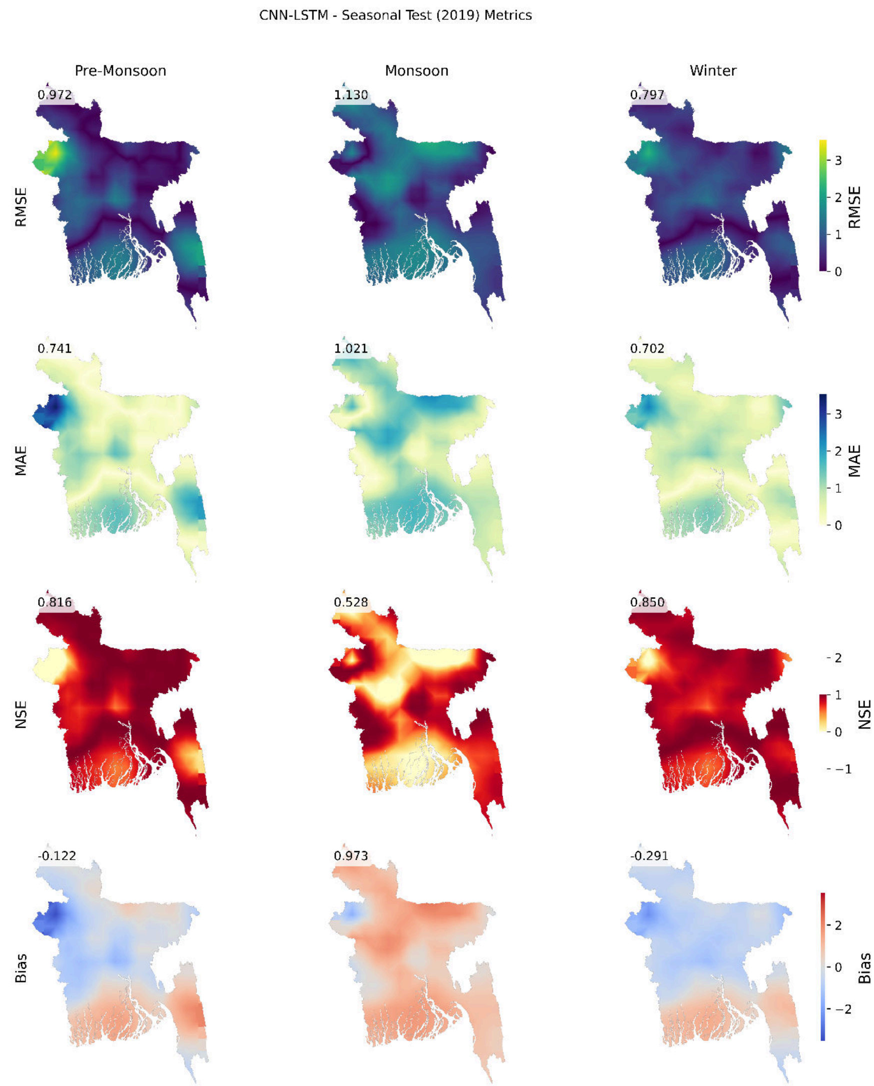


Figure 14: Spatially distributed seasonal test performance metrics (RMSE, MAE, NSE, and Bias) of the CNN-LSTM model across three hydrological seasons — Pre-Monsoon, Monsoon, and Winter — evaluated on the independent test dataset (2019), with spatially averaged metric values annotated in each panel.

Test RMSE values calculated (0.972, 1.130, and 0.797 m) are at most comparable or lower than the validation RMSE, while NSE scores have improved substantially, to values 0.816, 0.528, and 0.850. Given the visual evidence shown in columns 3-4 in Figure 12 showcasing somewhat sharper spatial gradients and better-resolved peak depths compared to validation predictions, these substantial improvements are ultimately in-line with what was previously observed. Yet, Bias maps show a critical shift in bias structure compared to figures 12 and 13. Here, Pre-monsoon nd Winter biases for the test period turn negative (-0.122 and -0.291 m), possibly revealing systematic under-prediction during dry seasons for 2019. Even then, bias within the Monsoon season has remained strongly positive (0.973 m). Through this seasonal asymmetry observed, spatial pattern discrepancies within the Monsoon period in Figure 12 become more apparent, as predicted maps grossly underestimate the spatial extent of shallow-depth zones across the central floodplains.

Notably, the elevated coastal RMSE identified in the validation period (Fig. 13) is substantially reduced in the test period (Fig. 14). Given that the underlying coastal hydrology and covariate representation are unchanged between the two periods, the most plausible explanation is that 2019 was, along the coast specifically, a comparatively less anomalous year—climatically and tidally—than 2017–2018, so the kriged ground-truth surface for this single test year happens to more closely resemble the smoother, lower-variability conditions the model implicitly learned to expect during training. This is consistent with the broader pattern noted above, that CNN-LSTM generalizes better to the single test year than to the two validation years overall, and it underscores a point returned to in the Limitations (Section 6.6): with only a single held-out year available for either validation or test, coastal skill estimates in particular are vulnerable to being disproportionately shaped by that one year's idiosyncratic conditions rather than reflecting a stable, reproducible level of model skill in this region.

Finally, Figure 15 extends evaluation further to the out-of-sample forecast period (2020), displayed in the rightmost columns of Figure 12, thus quantifying error accumulation accompanying multi-year extrapolation.

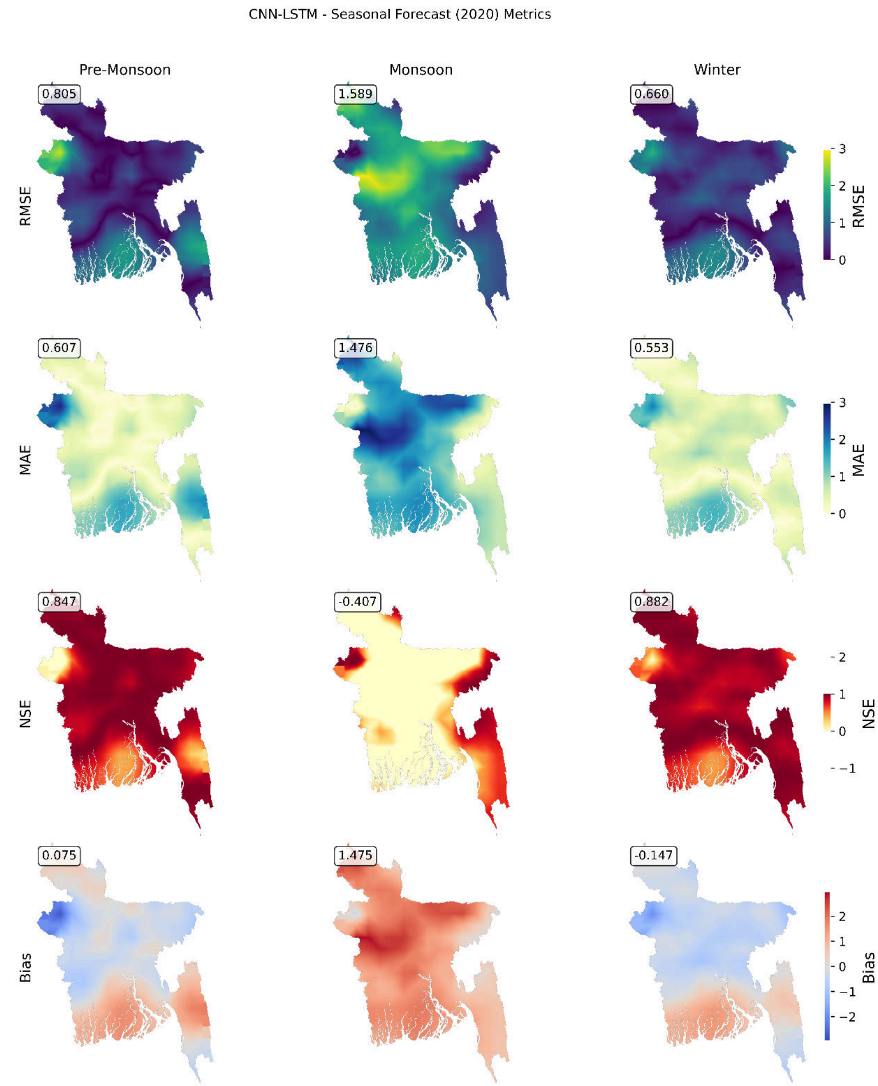


Figure 15: Spatially distributed seasonal forecast performance metrics (RMSE, MAE, NSE, and Bias) of the CNN-LSTM model across three hydrological seasons — Pre-Monsoon, Monsoon, and Winter — evaluated against observed groundwater levels for the forecast year (2020), with spatially averaged metric values annotated in each panel.

Monsoon forecasting across 2020 through CNN-LSTM has deteriorated dramatically. Not only NSE has collapsed from 0.528 in the 2019 test set (Fig. 14) to -0.407 in the 2020 forecast period 14), but RMSE and MAE has also sharply increased to 1.598 m and 1.476 m, respectively. The highest errors evaluated across all periods and seasons. Such high errors are deeply concentrated within central Bangladesh, represented as a broad yellow-green high-error zone in the RMSE map, aligning with the region shown in Figure 12 where the forecast column displays the greatest visual smoothing of the monsoon recharge signal. In comparison, Pre-Monsoon and the Winter forecast performance has remained relatively robust, with NSE equaling 0.847 and 0.882 respectively, aligning with a reasonably well-reproduced deep watertable hotspot in the north-western region covered in Figure 12. Given the strongly positive Monsoon Bias (1.475 m) in Figure 15, we can conclude that the model systematically over-predicts water table depth during the 2020 Monsoonal period and has failed to capture the full scale of aquifer recharge. Signs of this limitation, although visible in Figure 12's color field comparisons, are not quantifiable through said comparisons alone.

To directly address these systematic biases exposed in Figures 13, 14, and 15, we applied quantile mapping (QM) post-processing to the 2020 raw forecast maps as displayed in Figure 12's rightmost predicted column. Figure 16 models these corrections to the systematic biases diagnosed.

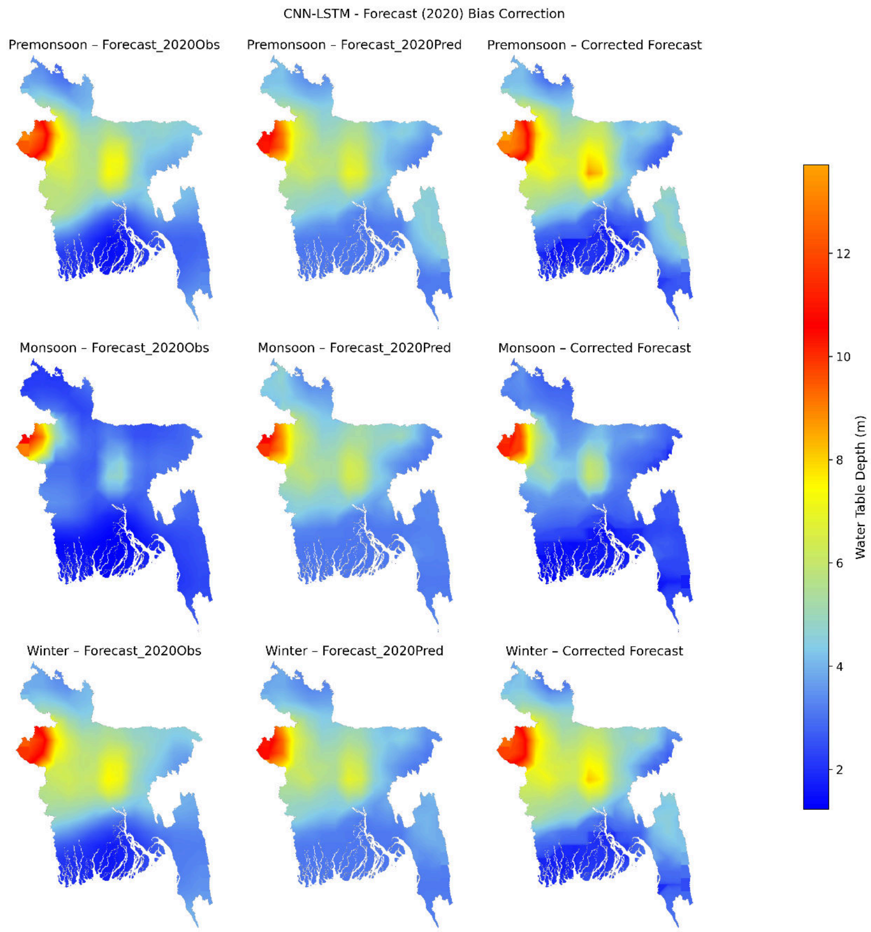


Figure 16: Quantile mapping (QM) bias-corrected seasonal water table depth forecasts (m) from the CNN-LSTM model for the forecast year 2020, presenting observed (left column), raw predicted (centre column), and bias-corrected predicted (right column) water table depth maps across three hydrological seasons: Pre-Monsoon, Monsoon, and Winter. The colour scale ranges from 2 m (blue, shallow water table) to $>$12 m (orange–red, deep water table).

Spatial smoothing and depth underestimation observed throughout figures 13-15 are all but confirmed and extended by comparisons made between the raw predicted maps (center column of Fig. 16) with the observed maps (left column). Not only is the Barind Tract hotspot spatially contracted throughout all three hydrological seasons, but the secondary north-central anomaly is also poorly resolved. Figure 16 shows that the QM-corrected forecasts have recovered substantially sharper spatial gradients, covering more realistic peak depth magnitudes measure across all three seasons. Not only is the north-western hotspot expanded towards its observed spatial extent, but the secondary anomaly has become more clearly delineated. In addition, the shallow southern coastal zone is better preserved. These corrections are more pronounced during the Pre-Monsoon and Winter seasons, both seasons bearing the largest positive systematic bias as shown in Figure 13, and in areas where forecasting skill has remained relatively high (NSE >0.84) as measured in Figure 15. Residual overestimation of central Bangladesh's watertable depths is, however, still prevalent during the Monsoon seasons even after QM-correction. While quantile mapping can effectively correct distributional bias present, it cannot recover the fundamental spatial structural errors from CNN-LSTM's global average pooling and the model's inability to resolve sharp-recharge driven lateral gradients without explicit graph-structured spatial reasoning, visible in every Monsoon predicted panel covered in Figure 12.

ST-GNN

Following the same 18-panel matrix format and color scale as Fig. 12, Figure 17 displays our observed and predicted water table depths across three hydrological seasons throughout our validation and test sets for our ST-GNN model.

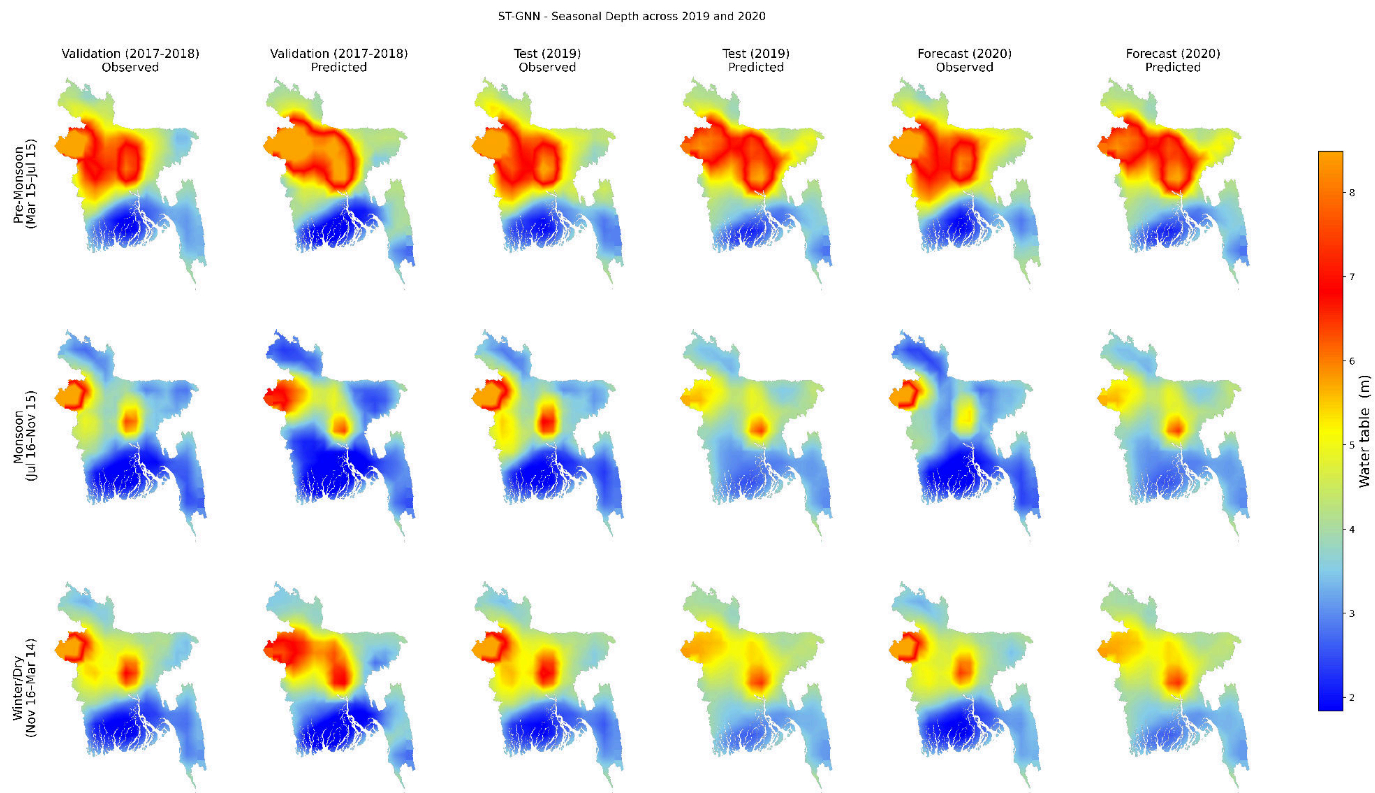


Figure 17: Spatially distributed seasonal water table depth (m) predicted by our ST-GNN implementation for model for validation (2017–2018), test (2019), and forecast (2020) periods across three hydrological seasons in Bangladesh: Pre-Monsoon (March 15 – July 15), Monsoon (July 16 – November 15), and Winter/Dry (November 16 – March 14). Observed and predicted water table depth maps shown for each period and season.

Similar to what was observed in Fig. 12, the Figure presents similar deep watertable anomalies dominating our observed water table maps for all seasons and years, identifying the same hotspots surrounding the Barind Tract and the Madhupur tract. Likewise, the southern lowlands remain to exhibit shallow watertables whereas the eastern Chittagong region has remained moderately deep. Different approaches were taken to recreate our observed maps from the two different types of datasets we created. As a result, our observed maps for our graph-based models may appear different from our observed maps for our CNN-LSTM model due to the differing colors scales used across our visualizations.

Figure 17 captures the ST-GNN's capabilities to similarly model spatial patterns during the validation period (2017-2018) like CNN-LSTM has. However, the model is able to correctly discern the shallower watertable zones across Bangladesh at a far more robust scale than the CNN-LSTM model, further evident with its substantially improved ground truth metrics in comparison.

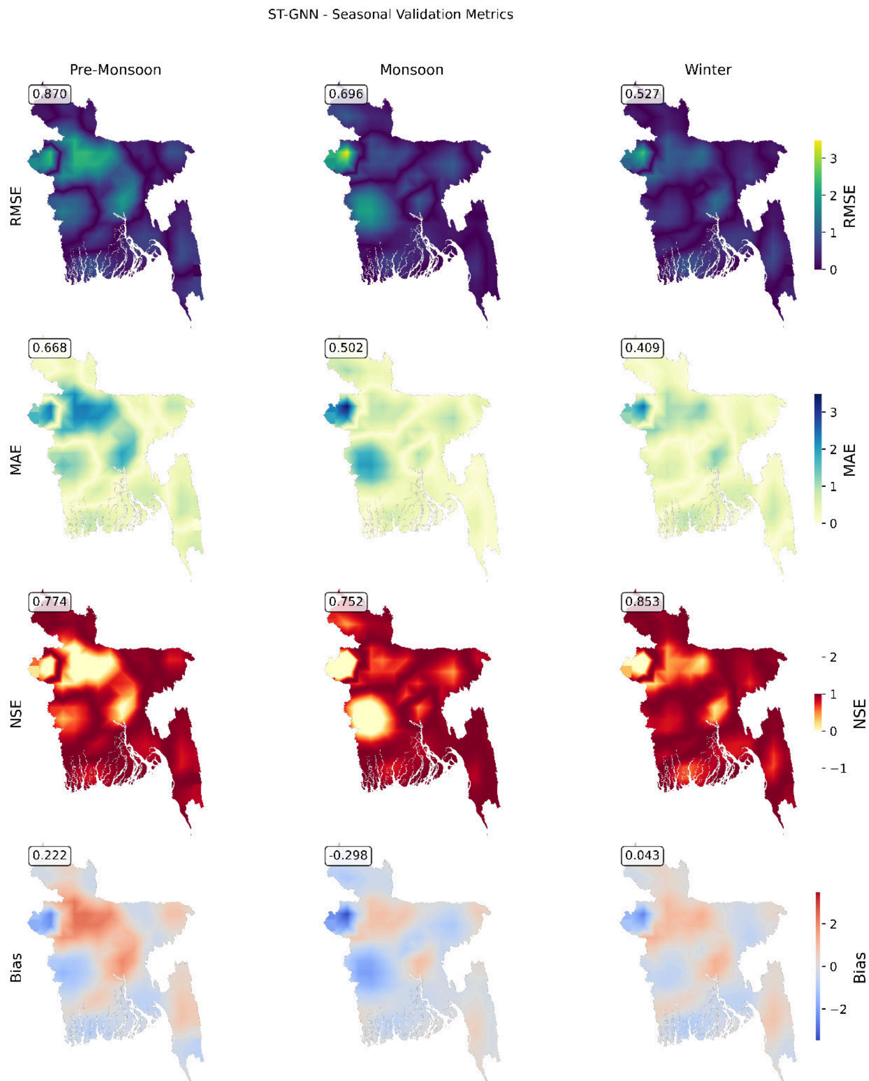


Figure 18: Spatially distributed seasonal validation performance metrics (RMSE, MAE, NSE, and Bias) of our ST-GNN model across three hydrological seasons. Spatially averaged metrics annotated across each panel. Colour scales indicate: RMSE and MAE (0–3 m, yellow–green = high error, dark blue = low error); NSE (dark red = high skill approaching 1, yellow–white = near-zero skill, dark maroon = negative skill); Bias (red = positive/over-prediction, blue = negative/under-prediction).

Elevated validation RMSE values of 0.870 m (Pre-Monsoon), 1.263 m (Monsoon), and 0.972 m (Winter) are shown to be significant improvements to what was observed in Fig.13 for the CNN-LSTM, furthermore with MAE (0.668, 0.502, and 0.409), NSE (0.774, 0.752, 0.853), and Bias

(0.222, -0.298, and 0.043). Unlike what was previously observed for CNN-LSTM, RMSE metrics for ST-GNN are reduced greatly within the south-eastern coastal zone, aligning with what was observed in Fig. 17 with the model correctly assessing shallower watertable depth within the region. This improvement is a direct, plausible benefit of ST-GNN's explicit graph adjacency: where CNN-LSTM decodes each patch independently and therefore has little to draw on locally in the well-sparse coastal fringe (see the CNN-LSTM results above), ST-GNN's message passing allows a coastal node to incorporate information aggregated from its better-observed neighbouring nodes at every layer, partially compensating for the lower local well density that limits CNN-LSTM in this same region. However, transition zones flanking the Barind Tract hotspot still exhibit disproportionately high RMSE concentrations. The MAE maps further brings light to this discrepancy with absolute errors at its highest within the Barind tract and the surrounding areas across Bangladesh's central plains, especially during the Pre-monsoon and Monsoon periods. NSE maps reveal that though the model has excelled in capturing the shallow watertable depth, it has struggled to capture aquifer dynamics within the Barind tract and its surroundings. Interestingly enough, NSE further highlights another hotspot of high error located south of the Barind Tract. Though Bias has been predominantly stable, the areas covering the Barind tract and the central plains exhibit zones of negative bias, indicating systematic underprediction for these particular areas, aligning with what was observed in 17 with multiple zones of underprediction. ST-GNN, although a substantial improvement over CNN-LSTM, faces problems with underprediction especially across the Barind Tract.

Transitioning to the test period maps covered in Figure 19 reveal that ST-GNN overall generalizes worse to the test period than it did towards the validation period, with RMSE (0.868, 0.990, 0.811), MAE (0.621, 0.683, 0.547), and NSE (0.806, 0.577, 0.684) being comparable to or worse than what was measured in the validation period, in addition to a higher bias concentration (-0.088, -0.002, 0.019). The opposite effect to what was observed in the ground-truth maps for CNN-LSTM. Interestingly enough, only the Premonsoon season has shown substantial improvements in ground truth metrics for the test period compared to the validation period.

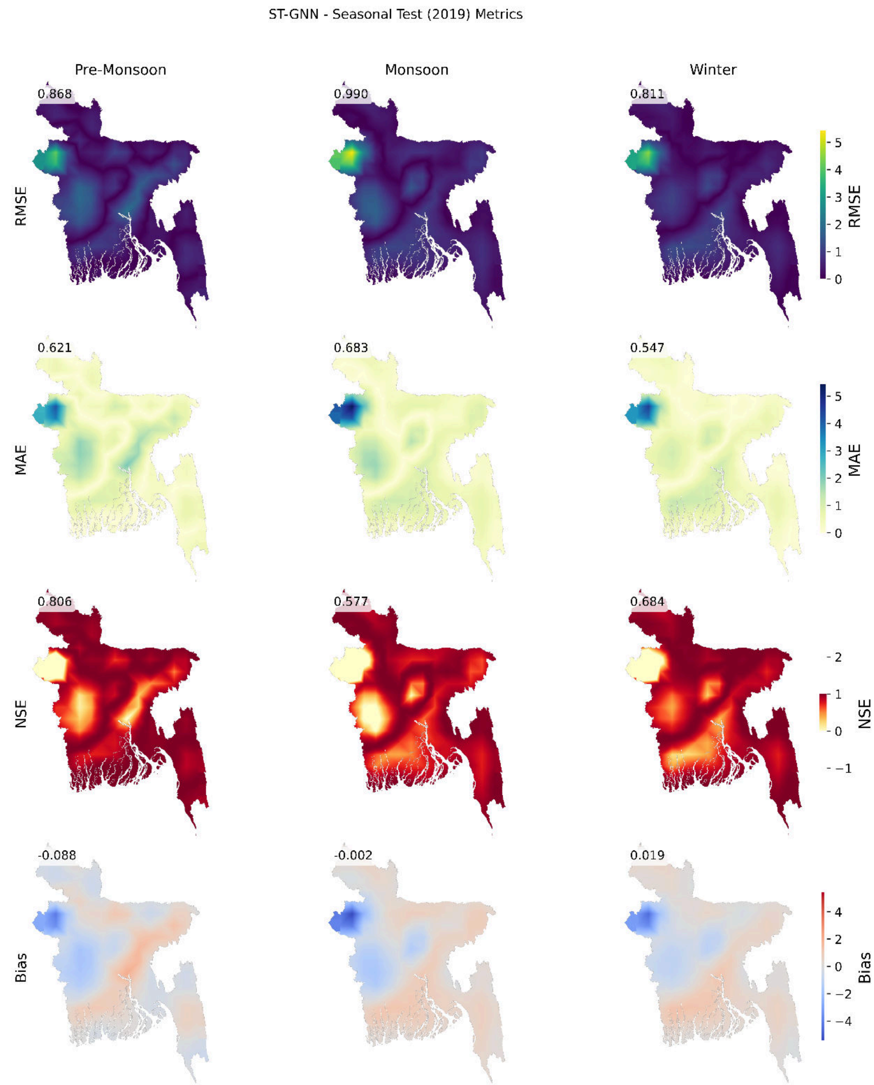


Figure 19: Spatially distributed seasonal test performance metrics (RMSE, MAE, NSE, and Bias) of the ST-GNN model across three hydrological seasons — Pre-Monsoon, Monsoon, and Winter — evaluated on the independent test dataset (2019), with spatially averaged metric values annotated in each panel.

Even then, this remains consistent with the visual evidence shown in Fig. 17, as ST-GNN was able to resolve peak depths in the Premonsoon season for the test period better than it has for the validation period, but has struggled with overprediction of shallow watertable zones and underprediction of groundwater hotspots, further evident with the NSE maps in Fig. 19 exposing a greater concentration of low NSE values across the southern coastal zone and for areas surrounding both hotspots of deep watertable depth. Bias structure has also been shifted slightly more positively across three seasons, yet systematic underprediction for the Barind Tract in particular has worsened.

Figure 20 extends this further to the forecast (2020) period, and here, ST-GNN's skill to forecast the Monsoonal period continues to collapse further, like how it has collapsed for CNN-LSTM.

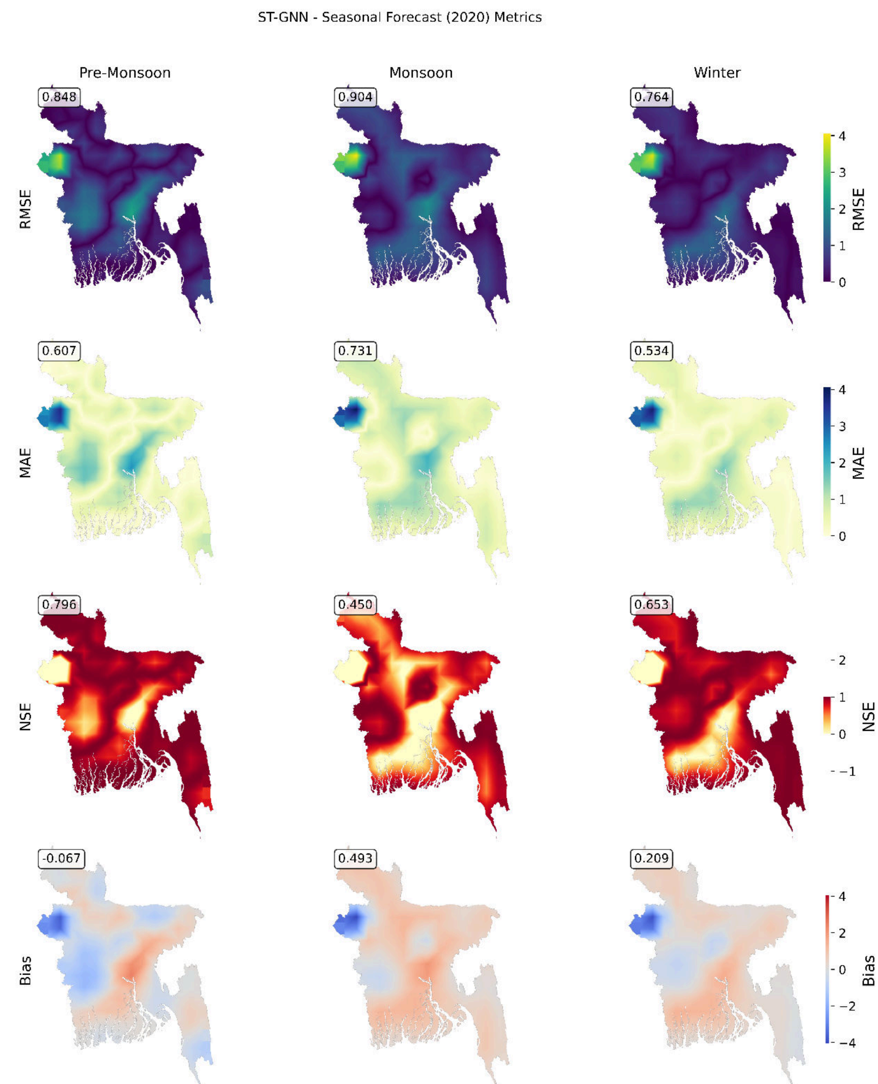


Figure 20: Spatially distributed seasonal forecast performance metrics (RMSE, MAE, NSE, and Bias) of the ST-GNN model across three hydrological seasons — Pre-Monsoon, Monsoon, and Winter — evaluated against observed groundwater levels for the forecast year (2020), with spatially averaged metric values annotated in each panel.

Initially starting at an NSE value (0.752) comparable to the other seasons across the validation period, only for it to collapse remarkably to a value of 0.450, compared to values of 0.796 and

0.653 for the Premonsoon and Winter seasons for the forecast period respectively. Similar to what was observed in 15, Monsoonal errors are largely concentrated within central Bangladesh, aligning with what was previously observed in Fig. 17 where overestimations persist across areas where shallower watertable depth is expected for all seasons. These overestimations are further evident in the bias maps showcasing a greater concentration of positive bias throughout Bangladesh.

We then apply quantile mapping (QM) post-processing to our 2020 raw forecast maps covering ST-GNN, results showcased in Figure 21.

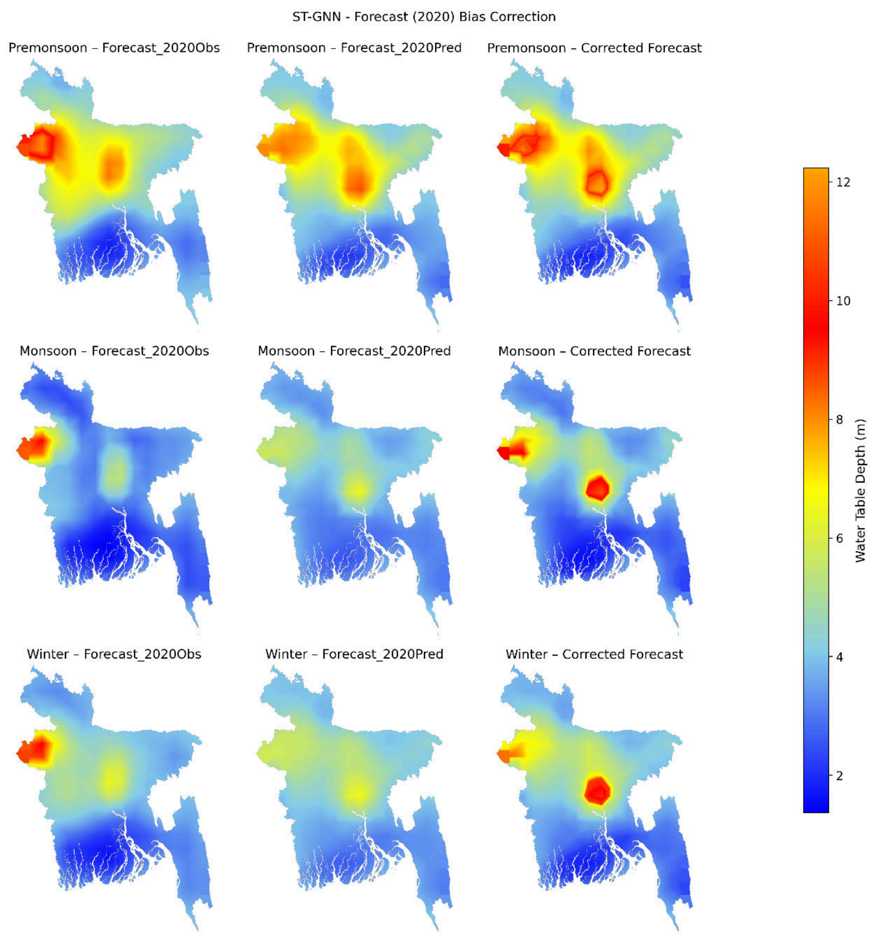


Figure 21: Quantile mapping (QM) bias-corrected seasonal water table depth forecasts (m) from the ST-GNN model for the forecast year 2020, presenting observed (left column), raw predicted (centre column), and bias-corrected predicted (right column) water table depth maps across three hydrological seasons: Pre-Monsoon, Monsoon, and Winter. The colour scale ranges from 2 m (blue, shallow water table) to >12 m (orange–red, deep water table).

Comparing the raw predicted maps with the observed maps, the model has struggled to capture watertable depth across the Barind and Madhuper tracts, both hotpots for deep watertable values, with only the Premonsoon season bearing raw predictions closest to what is observed. Though QM corrections were able to successfully amplify Barind Tract signals especially across the Monsoon and Premonsoon periods, it consequently overestimates watertable depth significantly within the 2nd hotspot, where watertable depth isn't as pronounced across the two seasons for the observed maps. Although QM mapping has been able to effectively correct distributional bias for certain regions, it is still prone to overestimate errors across other locations of distinct distributional characteristics.

ST-GTN

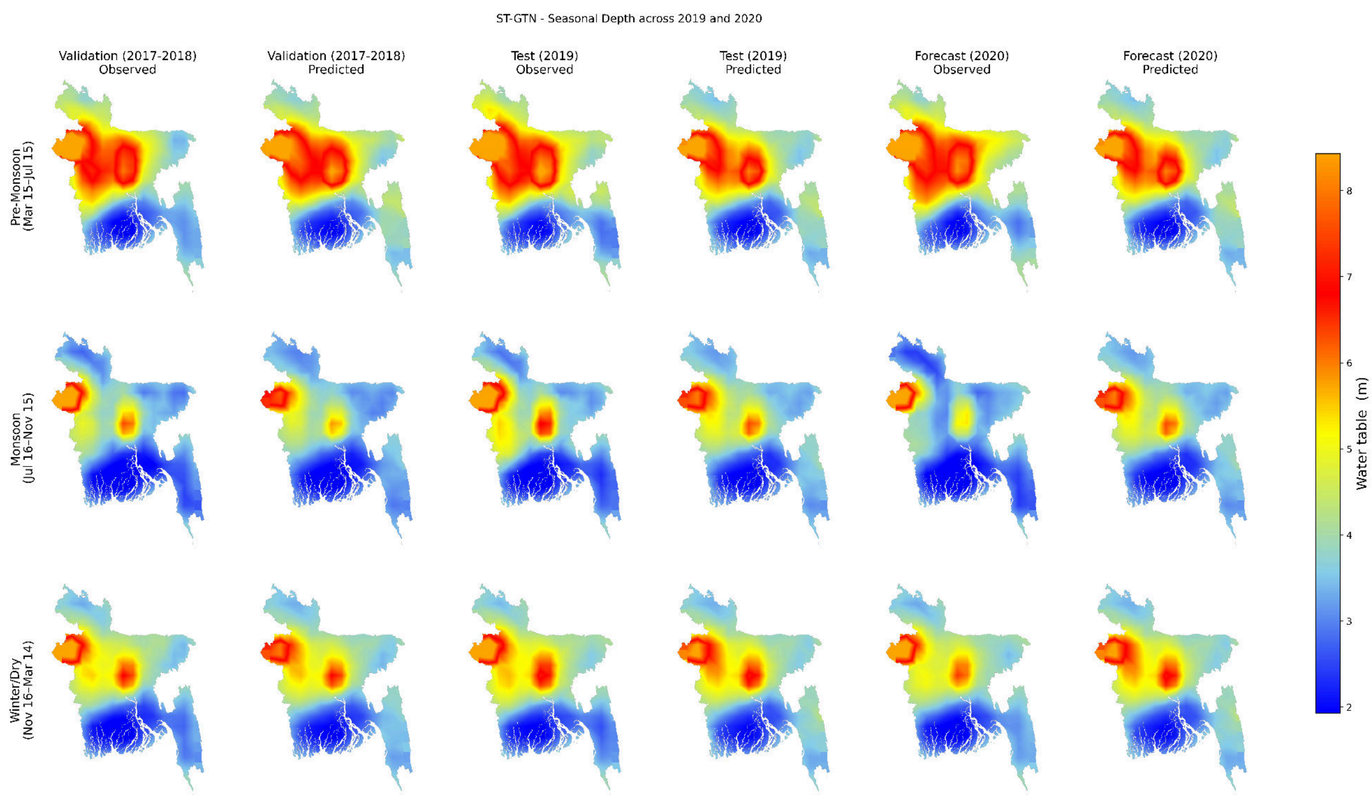


Figure 22: Spatially distributed seasonal water table depth (m) predicted by the ST-GTN model for validation (2017–2018), test (2019), and forecast (2020) periods across three hydrological seasons in Bangladesh: Pre-Monsoon (March 15 – July 15), Monsoon (July 16 – November 15), and Winter/Dry (November 16 – March 14), with observed and predicted water table depth maps shown for each period and season.

Figure 22 showcases the observed and ST-GTN predicted water table depths across three hydrological seasons, throughout the validation (2017-2018), test (2019), and forecast (2020) periods, under the same 18-panel matrix format and color as was displayed for the prior two models. Though the model was able to recognize the same deep water table anomalies as the prior models and identify the hotspots concentrated around the Barind and Madhupur tracts, visualizations show how effective the model is at capturing water table distribution. Of the three models, the ST-GTN model is the most effective at modeling the multiple groundwater zones concentrated across Bangladesh. Not only able to effectively predict the deep water table hotspots but it also accurately captured the shallower water table dynamics across the southern lowlands and the eastern Chittagong region.

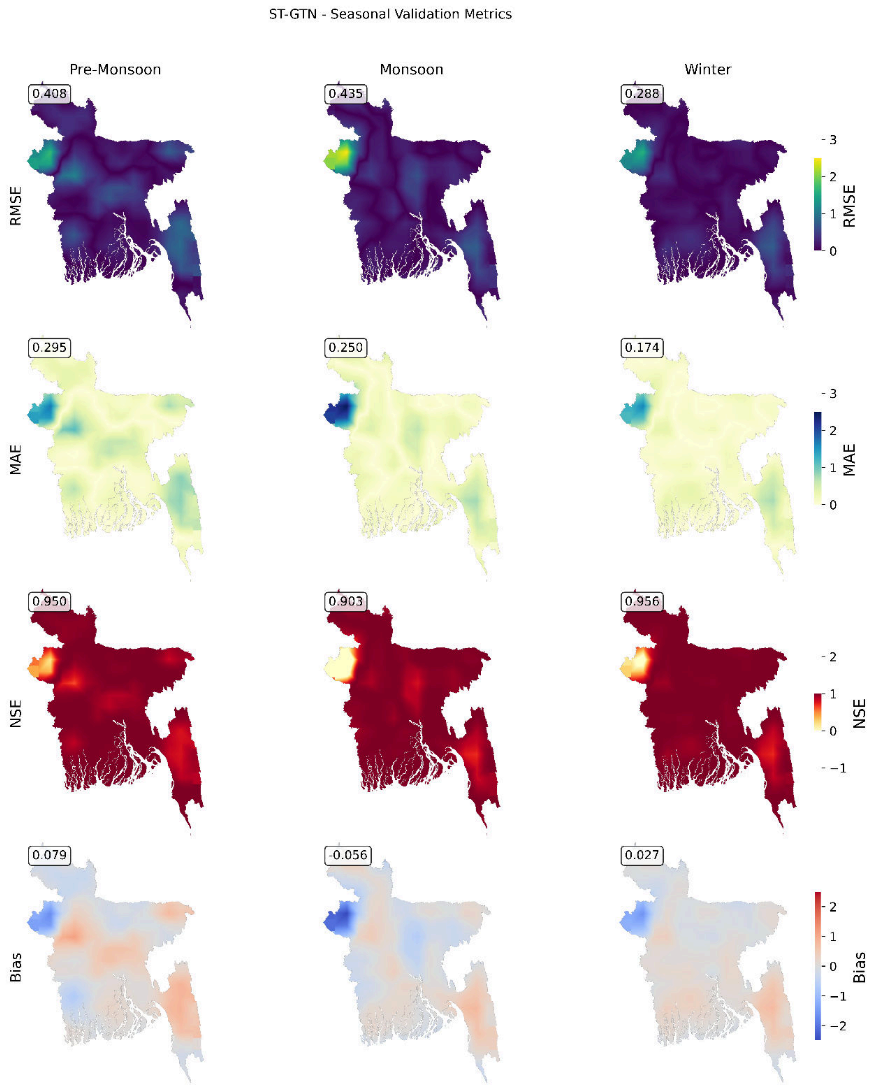


Figure 23: Spatially distributed seasonal validation performance metrics (RMSE, MAE, NSE, and Bias) of the ST-GTN model across three hydrological seasons — Pre-Monsoon (March 15 – July 15), Monsoon (July 16 – November 15), and Winter/Dry (November 16 – March 14) — evaluated over the validation period (2017–2018). Spatially averaged metric values are annotated in each panel. Colour scales indicate: RMSE and MAE (0–3 m, yellow–green = high error, dark blue = low error); NSE (dark red = high skill approaching 1, yellow–white = near-zero skill, dark maroon = negative skill); Bias (red = positive/over-prediction, blue = negative/under-prediction).

This is further reinforced in Figure 23, highlighting the best validation ground truth metrics for all three models. Scalar RMSE values of 0.408 m (Pre-monsoon), 0.435 m (Monsoon), and 0.288 m (Winter) are significant improvements over even the corresponding ST-GNN metrics calculated in Figure 18, which in of itself improve upon the metrics reported in CNN-LSTM in Fig. 13. The same conclusions are met with MAE (0.295, 0.250, 0.174), NSE (0.950, 0.903, 0.956), and Bias (0.079, -0.056, 0.027). Across all three season, however, the ST-GTN still appears to struggle with accurately predicting water table depth concentrated around the Barind Tract region, evident with higher RMSE and MAE concentrated within the same region, further reinforced with higher NSE and substantially lower Bias. Similar results were reported with ST-GNN's predictions in Fig. 18, where both models consistently under-predict water table depth surrounding the Barind Tract area despite recognizing it.

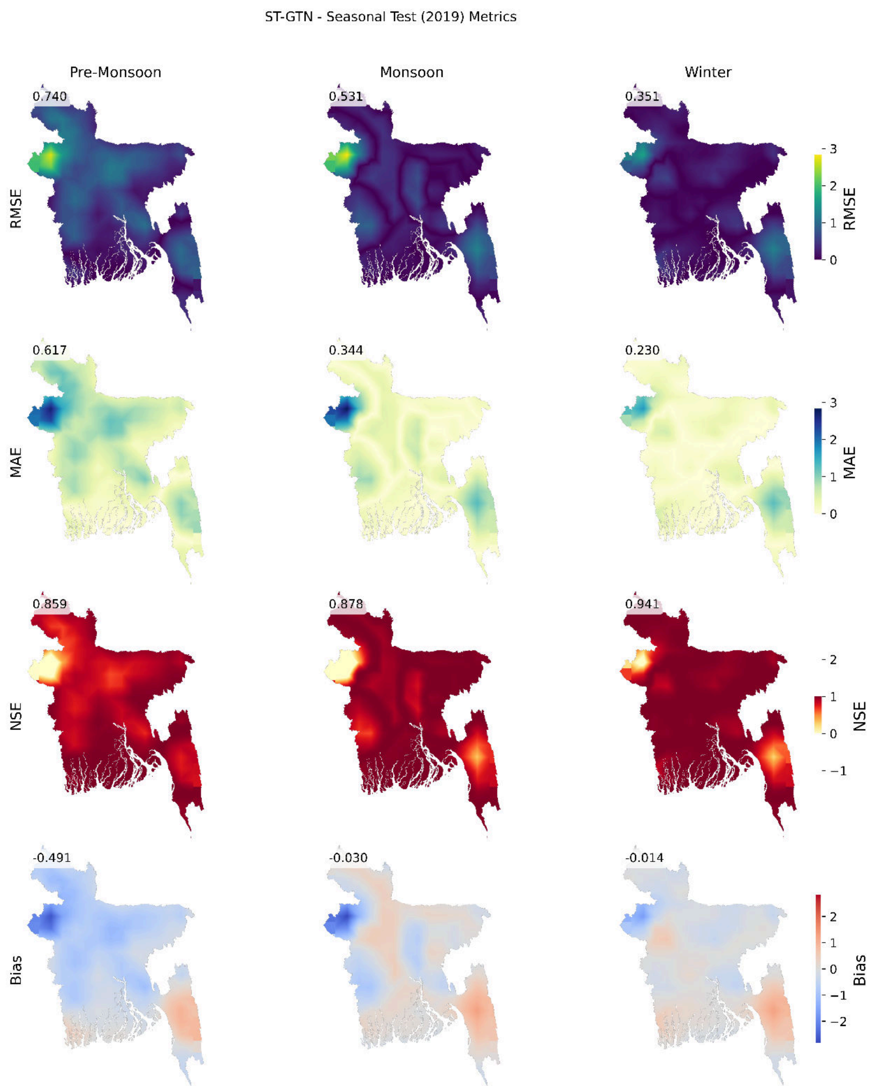


Figure 24: Spatially distributed seasonal test performance metrics (RMSE, MAE, NSE, and Bias) of the ST-GTN model across three hydrological seasons — Pre-Monsoon, Monsoon, and Winter — evaluated on the independent test dataset (2019), with spatially averaged metric values annotated in each panel.

Transitioning to the test period for Figure 24, ground truth deteriorate from the validation metrics as what we expected, yet still outperform what was previously reported in the prior two models. RMSE (0.740, 0.531, 0.351), MAE (0.617, 0.344, 0.230), and NSE (0.859, 0.878, 0.941) being notable downgrades from the prior validation predictions observed in Fig. 23, coupled with more widespread concentrations of negative bias (-0.491,-0.030,-0.014), indicating greater areas of underestimation especially surrounding the Barind Tract. Of the three seasons based on quantitative measures, we see that the Pre-Monsoon period showcases the worst performance, with greater systematic underestimation observed compared to the other three seasons. The inverse to what was previously observed in Fig. 19 with the prior ST-GNN model. Nonetheless, the test matrix has demonstrated consistent seasonal improvement in model accuracy given our RMSE and MSE progressively declining as the seasons pass, achieving optimal performance in the Winter period. Though the central and southern regions display strong predictive skill, the model struggles with capturing the complex groundwater dynamics highlighted in the northwestern and southeastern regions especially. Overall, this steady reduction in error-magnitude indicates the model's robust generalization to the independent test set covering 2019, yet the Pre-Monsoon period highlights the model's difficulty to accurately capture transitional recharge and extraction patterns.

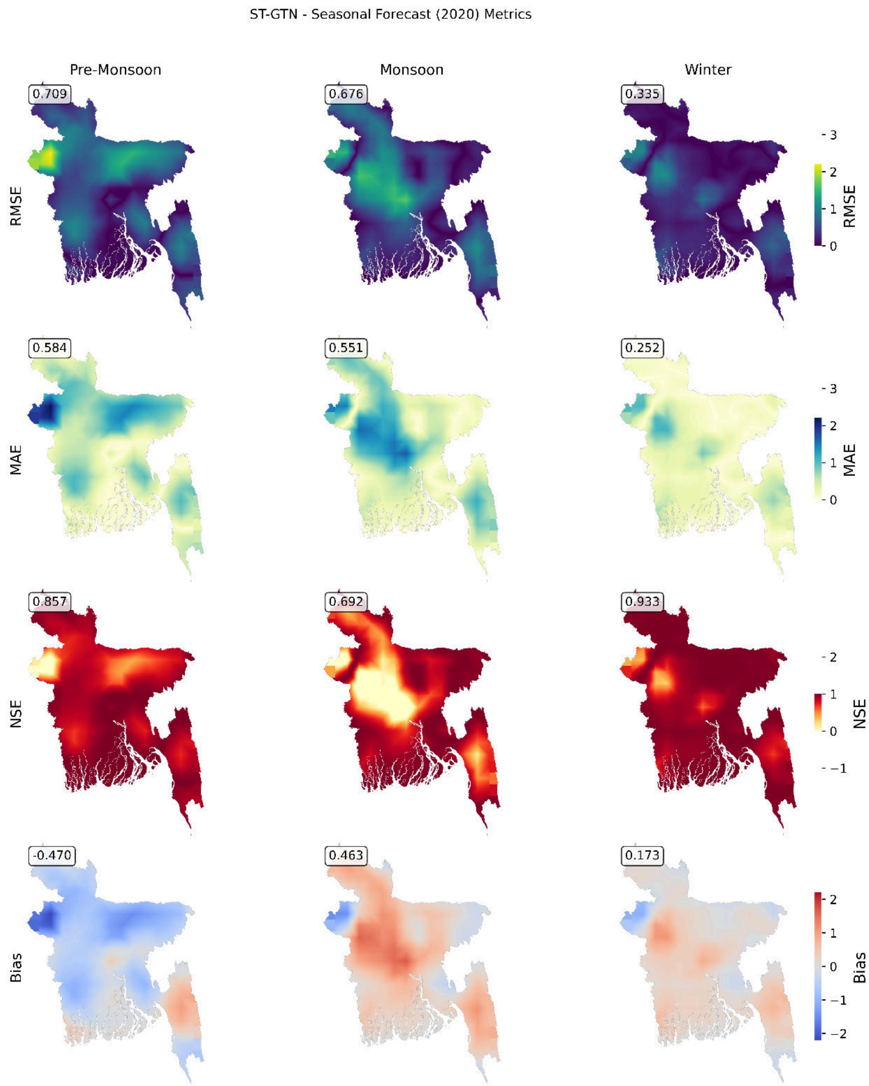


Figure 25: Spatially distributed seasonal forecast performance metrics (RMSE, MAE, NSE, and Bias) of the ST-GTN model across three hydrological seasons - Pre-Monsoon, Monsoon, and Winter - evaluated against observed groundwater levels for the forecast year (2020), with spatially averaged metric values annotated in each panel.

Extending to the Forecast period (2020) in Figure 25 displays the ST-GTN model performing at its worst, showcasing the worst RMSE (0.709, 0.676, 0.335), MAE (0.584, 0.551, 0.252), and NSE metrics (0.857, 0.692, 0.933), especially highlighted across the Monsoon period. Notably, the Pre-Monsoon season has foreseen incremental improvements across the Forecast period, but the model's predictive power over the Monsoon season has collapsed drastically, with greater overestimation of the central and southeastern Bangladesh regions highlighted by a greater concentration of positive bias (0.463 m compared to the previous value of -0.030 m in the test period) and a drastic decline in NSE encompassing these regions.

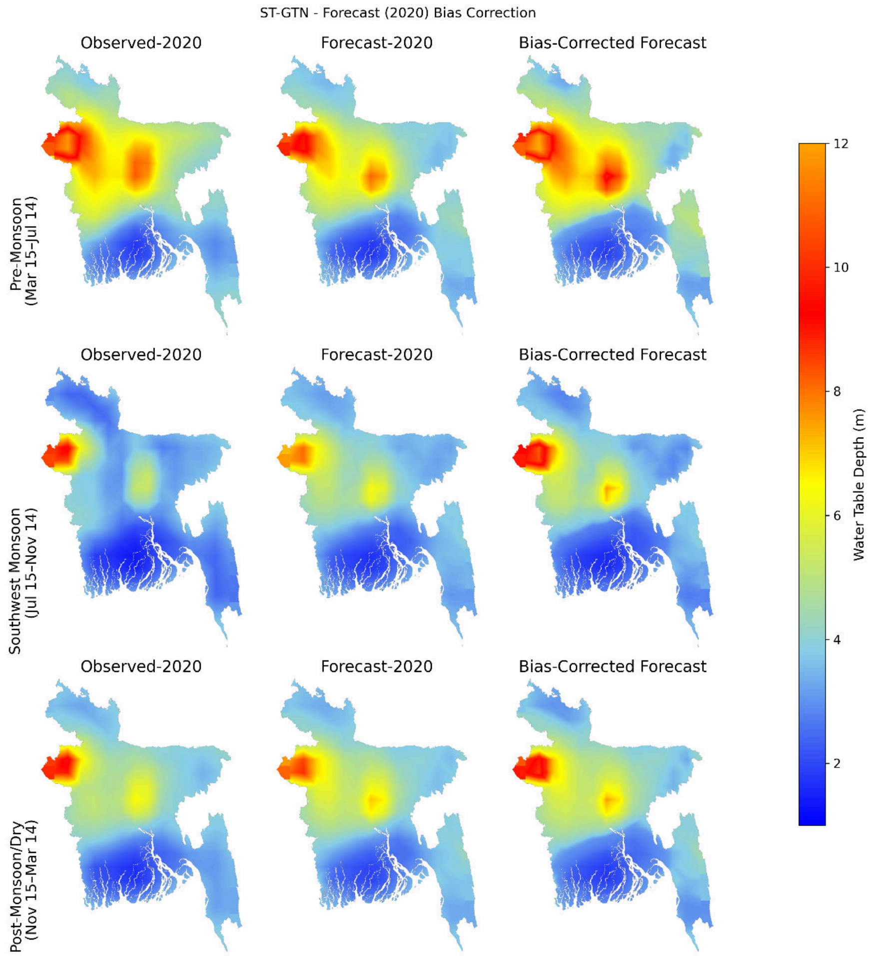


Figure 26: Quantile mapping (QM) bias-corrected seasonal water table depth forecasts (m) from the ST-GTN model for the forecast year 2020, presenting observed (left column), raw predicted (centre column), and bias-corrected predicted (right column) water table depth maps across three hydrological seasons: Pre-Monsoon, Monsoon, and Winter. The colour scale ranges from 2 m (blue, shallow water table) to $>$12 m (orange–red, deep water table).

We then apply quantile-mapping bias correction, seasonal results highlighted in Figure 26. Like what the previous ST-GNN model has shown in Fig. 21, applying quantile mapping successfully amplifies the Barind Tract signals to be closer to what is measured in the observed. However, we still run into this trade-off of QM inadvertently overestimating the central Bangladesh region, especially encompassing the Madhupur tract, our secondary hotspot.

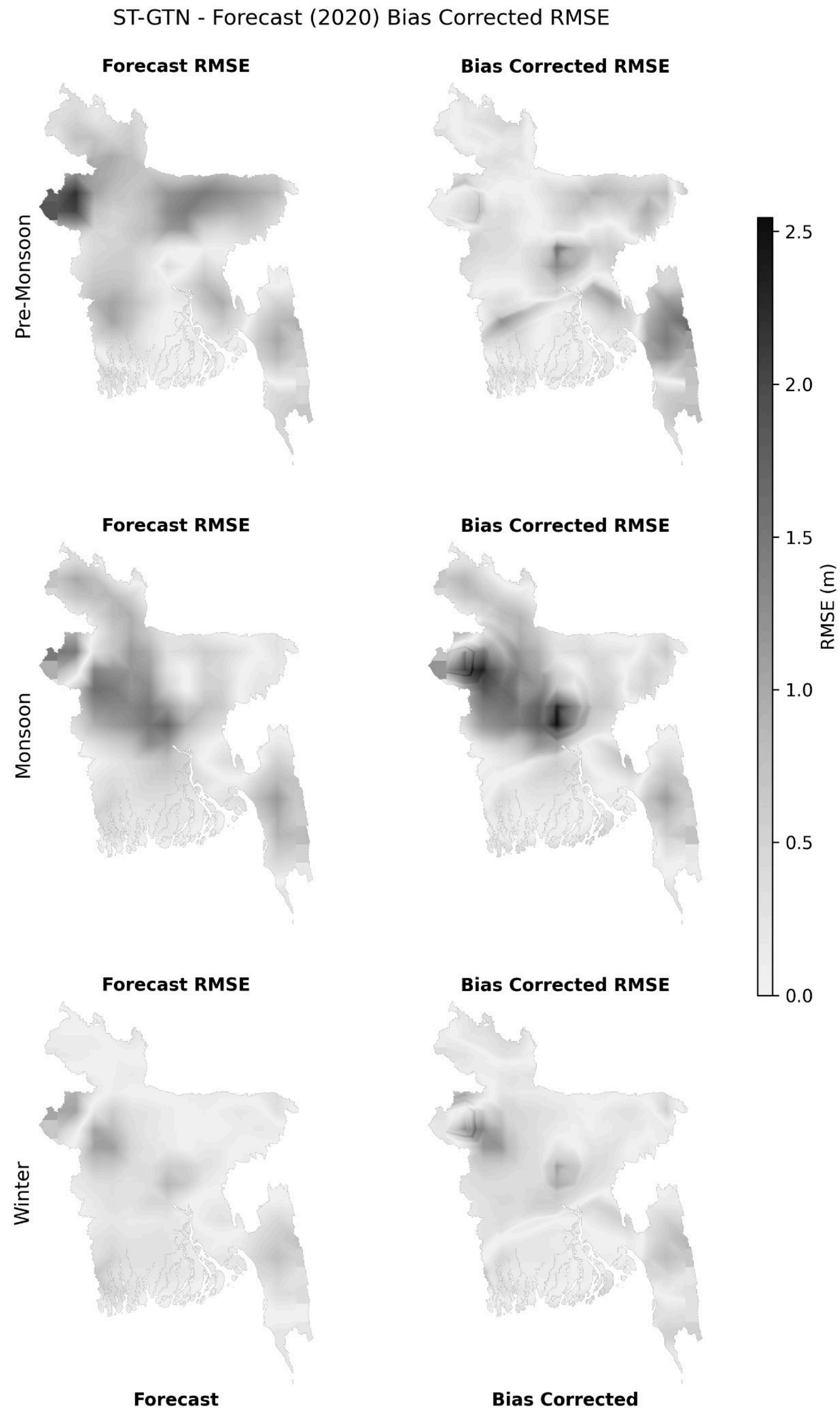


Figure 27: Seasonal RMSE maps comparing raw ST-GTN forecast (2020) predictions to bias-corrected predictions. Darker values indicate higher RMSE concentrated.

Figure 27 highlights this tradeoff further especially in the Monsoon period. Though slight spatial uniformity is evident as some regions improve in RMSE, the central-northwest and north-central zones still bear the highest errors, which seem to have worsened after applying QM. This suggests that Monsoon recharge introduces high variability that is otherwise difficult to capture even with bias correction applied, not fully resolving dynamic recharge processes that occur during periods of heavy rainfall. For the Pre-Monsoon period, a noticeable improvement is observed, with a more uniform error distribution that highlights reduction in systematic over/under-prediction. However, errors still persist in the south-eastern region of Bangladesh. The winter season shows the best model performance overall, holding generally the lowest RMSE errors, likely due to stable-dry season conditions harboring minimal recharge that make groundwater levels predictable.

Feature Importance Analysis

Figure 28 presents the combined feature importance analysis across all three models, organised into three panels: permutation importance, gradient·input sensitivity (ST-GNN and ST-GTN only), and the cross-model combined rank.

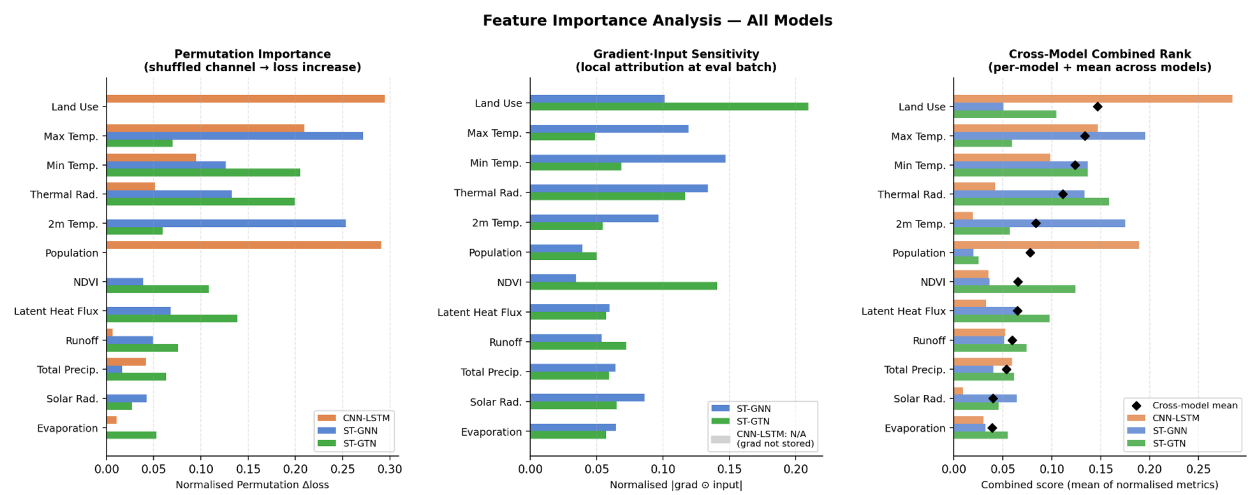


Figure 28: Feature importance analysis across all three models. *Left:* Normalised permutation importance (Δloss when channel is shuffled across the batch), shown for CNN-LSTM, ST-GNN, and ST-GTN. *Centre:* Normalised gradient·input sensitivity ($|\nabla_{x_c}\mathcal{L} \odot x_c|$), computed for ST-GNN and ST-GTN only (CNN-LSTM gradient not retained). *Right:* Cross-model combined rank (mean of normalised permutation and gradient·input scores per model, then averaged across models); black diamonds indicate the cross-model mean score. Features are sorted from most to least important by cross-model mean (top to bottom).

Permutation Importance

Across all three models, Land Use and Maximum 2-metre Temperature (Max Temp.) emerge as the two most consistently influential channels under the permutation criterion. Land Use registers the highest permutation Δloss in CNN-LSTM and ranks first overall in the cross-model mean, reflecting the strong spatial structure that reclassified land-use categories—particularly the persistent Agriculture class across the Barind and Madhupur Tracts—impose on groundwater depth. Max Temp. similarly scores highly across all models, consistent with its role as a proxy for dry-season evapotranspiration demand and irrigation timing, both of which directly modulate extraction pressure.

Population density achieves a notably high permutation importance in CNN-LSTM (second only to Land Use) but contributes comparatively little in the graph-based models. This divergence may reflect architectural differences: the CNN-LSTM processes population as a spatially varying raster channel within its convolutional encoder, which can exploit the sharp urban density gradient around Dhaka directly as a spatial feature; the graph models, operating on node-level aggregates, may absorb this signal through correlated channels such as land use.

Minimum Temperature, Thermal Radiation (str), and 2-metre Temperature each register moderate and broadly consistent permutation importance across all three models, forming a coherent temperature-radiation cluster that reflects the seasonal energy balance driving evapotranspiration and, by extension, net recharge.

Gradient·Input Sensitivity

The gradient·input results for ST-GNN and ST-GTN broadly corroborate the permutation findings but reveal additional structure. Minimum Temperature and Thermal Radiation rank highest in both models under this metric, indicating that not only does removing these channels increase loss (permutation), but the model's loss surface is also locally sensitive to their values at the evaluated batches — a stronger attribution signal than permutation alone.

Latent Heat Flux (slhf) and NDVI achieve higher gradient·input ranks than their permutation ranks would suggest, particularly for ST-GTN. This may reflect that these channels carry high-frequency or spatially localised signals that the transformer's attention mechanism has learned to exploit in specific spatiotemporal contexts, even when their global permutation impact is modest.

Cross-Model Combined Rank and Implications

The cross-model combined rank (right panel of Figure 28) confirms that Land Use, Maximum Temperature, Minimum Temperature, and Thermal Radiation are the four most consistently important features across all three architectures. The consensus across architecturally distinct models — a convolutional-recurrent hybrid, a graph neural network, and a graph transformer — strengthens the reliability of these rankings as genuine reflections of physical drivers rather than artefacts of any single model's inductive biases.

By contrast, Total Precipitation (tp), Surface Solar Radiation (ssr), and Evaporation (e) consistently rank in the lower half across all models. This counterintuitive result is consistent with the interpretation developed in the Discussion: in the monsoon-dominated hydrology of Bangladesh, precipitation's influence on the water table is temporally lagged by the unsaturated zone transit time, meaning that concurrent precipitation values carry limited explanatory power for same-period groundwater depth. The model may instead be capturing this seasonal recharge signal indirectly through temperature and radiation channels, which vary more smoothly and correlate with the hydrological season without the lag-mismatch problem. Additionally, in irrigation-dominated regions such as the Barind Tract, extraction demand — proxied partly by land use and temperature — may overwhelm the climatic recharge signal entirely, further depressing precipitation's apparent importance. Similarly, NDVI's low ranking despite its ecological relevance is consistent with vegetation response lagging groundwater change, reducing its value as a concurrent predictor.

These findings motivate future work to incorporate lagged precipitation and vegetation features, as well as direct irrigation extraction metrics, to assess whether the explanatory contribution of these variables improves when their temporal offset relative to groundwater response is explicitly represented.

A Methodological Caveat: Redundancy and Collinearity in Permutation Importance

The interpretation above treats Land Use's dominance and NDVI's comparatively low rank as reflecting genuine differences in their physical relevance to groundwater dynamics. That interpretation, however, competes with a purely methodological explanation that the permutation-importance criterion cannot by itself rule out: Land Use and NDVI are conceptually, and in practice statistically, correlated channels, since agricultural, urban, and forested land-cover classes are expected to carry systematically different vegetation signatures. Permutation importance measures the loss increase from shuffling a single channel while leaving every other channel, including any correlated ones, intact. If Land Use already encodes much of the same information that NDVI would otherwise supply, the model can fall back on the un-shuffled Land Use channel when NDVI is permuted, so the measured loss increase for NDVI understates its true relevance rather than reflecting a genuine absence of predictive value. This is the same redundancy effect noted in Chapter 6 for the temperature–radiation–precipitation cluster, and it applies with equal force here: NDVI's low permutation score is consistent with either (i) NDVI genuinely carrying little independent predictive signal once vadose-zone lag and extraction dominance are accounted for, or (ii) NDVI carrying real signal that is simply redundant with, and therefore masked by, the retained Land Use channel. Under the marginal, one-channel-at-a-time criterion used here, these two explanations cannot be fully disambiguated.

The gradient·input results offer partial, though not conclusive, evidence toward the first explanation: NDVI's gradient·input rank exceeds its permutation rank for ST-GTN (Section 5.6.2, Gradient·Input Sensitivity, above), indicating that the model's loss surface is locally sensitive to NDVI values even where globally shuffling the channel does comparatively little damage—a pattern more consistent with genuine, if secondary, predictive signal than with pure redundancy. Disentangling these effects with confidence would nonetheless require a conditional or Shapley-based importance measure that accounts for feature correlation directly, rather than the marginal, one-channel-at-a-time criterion used throughout this analysis. Until that analysis is carried out, the claim that land use functions as an indirect proxy for irrigation-driven extraction should be read as the interpretation most consistent with the available evidence, rather than one the feature importance results alone confirm outright; NDVI's low rank is at least as plausibly a symptom of how permutation importance handles correlated inputs as it is direct evidence about vegetation's physical (un)importance to groundwater forecasting. temporal offset relative to groundwater response is explicitly represented.

Discussion

Principal Findings

This research has aimed to ascertain the potential of spatiotemporal deep learning architectures to be able to produce skillful, seasonally resolved forecasts of groundwater depth across Bangladesh, using a multi-source covariate of climate, vegetation, land-use, and population data. From this research, we derived three principal findings.

For our first finding, our exploratory analysis reveals Bangladesh's groundwater system to operate under two somewhat distinct hydrological regimes. Amidst the southern and south-central lowlands, seasonal monsoonal recharge is what primarily governs water-table dynamics, resulting in shallow (approximately 2-3m) and spatially uniform water table depth that recharges annually. This contrasts with the groundwater dynamics observed in the northwestern Barind Tract and the north-central Madhupur Tract, where both regions exhibit persistently deep water table depth, of approximately 8-12 m on average. These depths deepen monotonically at rates of 0.20 m yr^{-1} across all seasons, which includes the monsoon recharge window. This cross-seasonal, monotonic deepening we ascertain to signify structurally driven anthropogenic over-extraction as opposed to it being a consequence of climatically modulated variability, consistent with the long line of observational evidence spanning these regions. Shamsudduha et al. [68] revealed that, from national monitoring records that, even if seasonality dominates observed variance in Bangladesh's water table depth, much of the declining trends that exceed 1 m yr^{-1} are primarily distributed within the northwestern, north-central, and urban areas. This figure is drawn from point-based well records rather than the gridded, kriged trend surface used in our own exploratory analysis above; individual wells in the most heavily-extracted pockets of the Barind Tract can therefore show a substantially steeper decline than the 0.15–0.20 m yr^{-1} spatially-averaged maximum we report, which is smoothed across neighbouring pixels by the interpolation process. Further studies repeatedly link the Groundwater declines in the Barind Tract to the growing expansion of dry-season boro rice irrigation [58, 56, 49]. GRACE satellite gravimetry identifies a 37.5 km^3 decline in Groundwater between 2002 and 2020 in Bangladesh, with irrigation being the dominant driver of these declines [20].

Second, of the three architectures we evaluated—CNN-LSTM, ST-GNN, and ST-GTN—our Spatiotemporal Graph Transformer Network (ST-GTN) has delivered the strongest and most consistent performance across the validation (2017—2018), test (2019), and forecast (2020) periods, attaining validation RMSE of 0.288-0.435 m and Nash-Sutcliffe Efficiency values reaching 0.956. Its advantage over the graph-convolutional and the convolutional-recurrent baseline lies precisely where hydrology is most complex, being the Barind Tract during the dry season as well as the monsoon transition zones. Of the three models tested, the CNN-LSTM baseline performed the worst, showing significant systematic under-prediction bias and negative NSE on the held-out test year.

Third, all three models, regardless of overall performance, share and exhibit two failure modes: the systematic underestimation of peak groundwater depths in the Barind Tract and a pronounced decline on monsoon forecast skill in 2020. Although quantile-mapping post-processing was effective in removing distributional bias within the pre-monsoon and winter forecasts, it could not repair the underlying spatial structure of monsoon-season errors.

We note one methodological effect alongside these physical explanations: permutation importance's tendency to understate a feature's true contribution when correlated alongside other retained features. Shuffling one channel can leave correlated channels intact for the model to fall back on. From the plausible correlation between temperature, radiation, and precipitation in a monsoon climate, certain aspects of precipitation's apparent unimportance may reflect this collinearity rather than being purely a hydrological lag effect. The same caveat applies to Land Use and NDVI: although the two channels are conceptually related, since agricultural, urban, and forested land cover are expected to carry systematically different vegetation signatures, Land Use's dominance and NDVI's comparatively low rank are also consistent with the model falling back on Land Use once NDVI is permuted, rather than with NDVI genuinely carrying no independent signal (Chapter 5, Section 5.6.4). Under the marginal, one-channel-at-a-time criterion used here, these explanations cannot be fully disambiguated from a genuine physical effect. In future work, imposing a conditional or Shapley-based importance measure would be necessary to disentangle these caveats.

A fourth finding, developed through the post-hoc feature importance analysis presented in Chapter 5, concerns the relative contribution of input covariates to forecast accuracy. Contrary to initial expectations, total precipitation and NDVI ranked among the least important predictors across all three architectures under both permutation importance and gradient·input attribution, whilst minimum and maximum temperature, land use, and surface thermal radiation ranked most consistently. This result is not interpreted as unambiguous evidence that precipitation and vegetation are hydrologically irrelevant—the methodological caveat above means the rankings are also partly shaped by feature redundancy—but, to the extent they do reflect genuine physical signal, as the diagnostic signature of three interacting mechanisms: a weeks-to-months vadose-zone lag between rainfall and water-table response; the dominance of anthropogenic extraction over climatic recharge in the irrigated northwest; and the role of temperature and radiation as indirect proxies for irrigation-demand timing. Section 6.3 develops each mechanism in detail. Sections 6.2–6.5 interpret each of these findings in finer detail, with Section 6.6 covering the limitations faced in this study and Section 6.7 documenting the implications for groundwater management derived from said findings.

Model Architecture and Predictive Skill: Why the Graph Transformer Wins

We can assess the superior performance of the ST-GTN model to the combination of two architectural properties: the explicit graph-structured representation of spatial dependence between groundwater monitoring locations and the self-attention mechanism over the temporal dimension. The latter which especially permits the model to weight informative past states without the vanishing-memory constraints of recurrence [83].

These findings align with an emerging consensus of incorporating Transformer architecture in machine-learning hydrology. In the most directly comparable study, Sun, Chang, and Chang [78] evaluated a Transformer network against CNN, LSTM, and MLP benchmarks for groundwater-level prediction within the Zhuoushui River basin of Taiwan, through incorporating a 20-year hydrometeorological dataset at a 10-day time-step. They revealed that the Transformer outperformed all their benchmarks, and attributed its advantage to the self-attention mechanism's capacity to capture complex, long-range relationships among the main drivers of groundwater levels like rainfall and river flow. From earlier benchmark comparisons, it's established that there is no single conventional architecture that is the titular best at forecasting groundwater levels. Through comparing CNN, LSTM, and NARX models across a large sample of German Wells, Wunsch, Liesch, and Broda [95] identified broadly similar accuracy among recurrent and conventional approaches and suggest that the marginal gains reported for attention-based models represent a legitimate architectural advance rather than it being a dataset idiosyncrasy.

Our findings extend this literature in two respects. Firstly, whereas Sun, Chang, and Chang [78] addressed a relatively short horizon of 10-20 days in their forecasting problem, highlighting dominance by rainfall and river-flow signals, our study demonstrates the Transformer advantage at a seasonal scale over a monsoon-driven deltaic aquifer system which bears intense anthropogenic forcing: a substantially different recharge regime of a sharper spatial heterogeneity, furthermore exemplified by the irrigation drawdown of the Barind Tract. Second, the ST-GTN outperformed not only the standalone convolution and recurrent baselines but also the CNN-LSTM hybrid, setting implications that even by combining convolutional feature extraction with recurrent temporal modeling, it still cannot close the gap opened by attention over graph-structured space-time data. Even so, these comparisons must still be made with caution. Given the differing forecast horizons, covariates, and hydrogeological settings, the agreement between studies should instead be read as converging support for transformer-style architectures to be a promising general direction for groundwater forecasting and not as explicit evidence that any specific architecture can translate across settings.

Additionally, a relevant secondary finding details ST-GTN to have also converged the fastest: 23 epochs against the 33 epochs ran for CNN-LSTM and 63 for ST-GNN, while CNN-LSTM proved to be the most computational demanding, due to dense image patches supplying far greater quantities of training points compared to the sparse graph representation. Where national agencies face hardware constraints of the kind as described in Section 6.6, graph-based models otherwise offer a favorable accuracy-per-compute profile.

Climatic Versus Anthropogenic Drivers of Water-Table Dynamics

The cross-model feature importance analysis (Chapter 5, Figure 28) confirms that land use and maximum temperature are the two most consistently influential channels under permutation importance across all three architectures, whilst total precipitation and NDVI rank in the lower half of all twelve covariates. Minimum temperature, thermal radiation, and surface latent heat flux occupy intermediate positions, forming a coherent temperature–radiation cluster. Gradient·input sensitivity scores for the graph-based models corroborate permutation rankings for the temperature variables, whilst elevating NDVI and latent heat flux slightly relative to their permutation ranks, suggesting these channels carry localised attribution signals that the transformer's attention mechanism exploits in specific spatiotemporal contexts even when their global shuffling impact is modest

[75, 9].

Contrary to initial expectations, NDVI and total precipitation contributed far less to forecast accuracy than minimum–maximum temperature and surface radiation across all models. Given the conventional understanding of precipitation as the primary driver of groundwater recharge, these findings prompt explanation, and we advance three non-exclusive mechanisms.

Delayed hydrological response. We ascertain groundwater lags precipitation input due to infiltrating water to must transit the unsaturated (vadose) zone before reaching the water table. This lag is a well-documented, first-order feature of recharge dynamics, scaling strongly with vadose-zone thickness, which ranges from under two months where water table depths are shallow to many years where they are deep [62, 27, 91]. In our study area, regions of highest management concern—the Barind and Madhupur uplands specifically—show the deepest water tables, and thus the longest expected precipitation lags. A model supplied only with short-lag precipitation features therefore under-attribute importance to precipitation despite being the underlying physical driver; the signal arrives at the water table only weeks to months after the covariate window closes.

Extraction-dominated variability. Spatial anomalies present in water-table depth within the Barind and Madhupur Tracts are governed less by climatic recharge than by abstraction for dry-season irrigation [58, 56]. Intensive pumping in this region has contributed to fundamentally alter the recharge regime itself, by inducing additional recharge through creating subsurface storage capacity [69]. At basin, scale, Shamsudduha et al. [70] quantified this coupled pumping-recharge dynamic—the "Bengal Water Machine"—to highlight dry-season abstraction by roughly 16 million smallholder farmers cumulatively captured 79–90 km^3 worth of freshwater between 1988 and 2018 [70]. Shamsudduha et al. has also documented that the potential of this new mechanism is limited, with withdrawals at their highest in the Barind area. In this sort of system, raw precipitation totals are a weak proxy in highlighting net water-table change. Instead, the dominant signal is anthropogenic demand, which our feature set covers only indirectly.

Temperature and radiation as demand proxies. We assess two pathways that could plausibly explain the high importance of minimum-maximum temperature and surface radiation in our groundwater forecasts. Not only do they modulate evapotranspiration and therefore do not recharge, but that they also correlate with seasonal timing of irrigation demand. It may be the case that the model is therefore reading extraction pressure through its climatic proxies, consistent with the finding that these variables outperform precipitation primarily in irrigation-intensive regions. We treat this as the physically most plausible reading of the rankings rather than a conclusion the feature importance results confirm outright: as noted above, permutation importance cannot fully separate a channel's genuine causal contribution from redundancy with correlated channels retained elsewhere in the feature set, so the same ranking pattern would also arise, at least in part, from temperature and land use simply absorbing signal that a more direct extraction covariate would otherwise carry.

Disentangling these mechanisms—in particular, testing if lagged precipitation features constructed and leads consistent with vadose-zone transit times can recover predictive importance—is a concrete and high-priority direction for future work. A broader implication for model design from these findings is that in heavily-irrigated aquifer systems, covariate sets built primarily from climate and vegetation indices are structurally insufficient. We should instead incorporate and treat direct or proxy measures of extraction—such as tubewell density, electrification of pumps, irrigated-area maps, etc.—as first-class predictors.

The Barind Tract: Systematic Underestimation of Extraction-Driven Drawdown

All three models have systematically underestimated peak groundwater depths in the Barind tract, which we interpret as the interaction of two compounding factors.

Indirect withdrawal representation. As argued in Section 6.3, water-table depth in this region is driven substantially by anthropogenic extraction for dry-season boro irrigation [58, 70]. With the lack of any direct representation of withdrawal volumes in our feature set, the models must instead infer extraction indirectly, through factors such as land-use, temperature, and seasonal-timing proxies, bounding the magnitude of drawdown that they can express. This results in a censoring effect, where predictions saturate below the true extremes of groundwater depth distribution.

Spatial resolution constraint. The Barind depth anomalies are spatially concentrated, their characteristic scale potentially falling beneath the effective resolution imposed by our patching system (patch size 64, stride 32—a constraint dictated by available GPU memory; see Section 6.6.). The Barind Tract itself is hydrogeologically distinctive within Bangladesh, bearing elevated terrain, relatively low rainfall of approximately 1,100–1,300 mm yr^{-1} compared to the national mean, and a thick clay-rich superficial layer that exacerbates runoff while hindering infiltration; an estimated 85% of the tract is classified to have low recharge potential [2]. Sharp lateral gradients between this upland and surrounding recharge-responsive lowlands are precisely the type of features that spatial smoothing at a coarse patch resolution will average away.

We propose two remedies: the implementation of direct extraction covariates like tubewell density or abstraction-volume estimates—where obtainable from the Barind Multipurpose Development Authority (BWDB) and finer spatial patching permitted by greater computational capacity or memory-efficient architectures.

The 2020 Monsoon Skill Decline and the Limits of Post Hoc Bias Correction

All models showcased a pronounced decline in forecasting skill during the 2020 monsoon period. We assess the plausible combination of three factors to produce this. From a data standpoint, 2020 is both the terminal and most temporally distant year of the forecast set. No model observes any dynamics within or beyond it, so anomalous hydrological behavior for that year arises as pure forecast error instead of it being partially absorbed in training. Hydrogeologically, monsoon recharge involves sharp, spatially discontinuous lateral gradients within groundwater response, specifically across the transition zones between the Barind and Madhupur uplands and surrounding lowlands—gradients that were difficult to resolve at our chosen patch resolution.

The behavior of the quantile-mapping (QM) post-processor is diagnostic here. Although QM has successfully reduced distributional bias in the pre-monsoon and winter forecasts, it failed to repair the inherent spatial structure of monsoon errors, particularly the sharp lateral gradients across central Bangladesh. This is exactly what the bias-correction literature predicts: QM is a univariate transformation applied independently at each location, thus correcting only the marginal distribution of predictions, leaving spatial (and inter-variable) dependence structures untouched or misrepresented [42, 25]. QM may be able to remove systematic over- or under-prediction, but it cannot introduce spatial information that the underlying model has failed to learn. Residual monsoon errors are therefore a model-capacity problem rather than a calibration problem. Appropriate responses are but structural: use a finer spatial resolution, introduce additional irrigation-related covariates, or implement physics-informed architectural constraints on lateral flows.

Limitations

Spatial generalization versus transductive evaluation. A question raised during the defense of this thesis was whether ST-GTN's markedly strong performance reflects genuine overfitting rather than genuine skill, and it is worth addressing directly rather than assuming either answer. All three models were evaluated under a strictly temporal train/validation/test split: every well location—and, for the graph-based models, every node in the fixed, country-wide adjacency—is present during training, with only future timesteps held out for validation, test, and forecast. This is a transductive evaluation of temporal extrapolation at already-known locations, not an inductive test of spatial generalization to unseen geography. Two pieces of evidence weigh against the more alarming form of overfitting, namely memorization: ST-GTN's training and validation losses converge to near-identical values with no widening gap across epochs (Section 5.2), unlike CNN-LSTM's persistent train-validation gap, which we do interpret as mild overfitting; and ST-GTN's degradation from validation to test (Table 5) is comparatively graceful (NSE 0.853 $\rightarrow$ 0.592, R^2 0.854 $\rightarrow$ 0.687) next to ST-GNN's far sharper collapse over the same transition (R^2 0.744 $\rightarrow$

0.059), which is more consistent with what memorization-driven overfitting would look like. Neither observation, however, can rule out a subtler form of overfitting: because the model has seen every well's position and local neighbourhood throughout training, its strong metrics may partly reflect familiarity with the specific spatial graph topology of Bangladesh's well network rather than a function of the covariates alone that would transfer to unmonitored locations. Distinguishing these possibilities requires a regional holdout scheme—withholding entire subsections of the country's well network (e.g., by administrative division or hydrogeological zone) from training entirely, then evaluating on the withheld region—which was suggested as a priority direction for future work during the defense and is detailed further in the Conclusion. Until that test is run, ST-GTN's reported skill should be read as applying to forecasting future conditions at already-monitored wells specifically, not as evidence that the model would generalize to unmonitored parts of the country.

Observational coverage. While our study assembled water-table coverage from approximately 1,000 monitoring wells with usable records, drawn from the 1,300-well BWDB network, records could not be extended beyond 2020, and observations from roughly 2015 and onward proved difficult to obtain. In addition, several wells exhibit either intermittent records or are absent from the available datasets entirely. As forecast evaluation rests on a single held-out year, any anomaly in 2020 exerts disproportionate influence on reported forecast skill. Although the densest in any alluvial basin [70], Bangladesh's monitoring network still remains uneven in access and continuity, a constraint noted more broadly for AI adoption in the countries water institutions [70, 47].

Climate covariates were sourced from ECMWF's ERA-Interim [17], which has ceased production on August 31st, 2019 and has since been superseded by ERA5. This is an external data-discontinuation constraint rather than it being an access limitation, and highlights why climate covariates could not be extended into 2019-2020 without a dataset transition. Efforts were made to obtain climate records through ER5 but were met with consistency issues trying to fit them onto an established pipeline in addition to access/processing time constraints. As ERA5 does cover the same time period as this study while extending it beyond 2020, future work has it so the climate covariate stack should be fully migrated to use ERA5 instead.

Computational capacity. Model training was performed on two 8-GB GPUs, which capped the patching scheme to a patch size of 64 and a stride of 32; finer patching exhausted available VRAM for every architecture tested. This constraint bounds the effective spatial resolution of all results and is directly implicated in Barind underestimation (Section 6.4). Of our three models, the CNN-LSTM is the most computationally demanding model despite being the least accurate, mainly due to its dense patch representation multiplying training points relative to the more sparse graph-based models.

Missing abstraction data. The lack of direct-irrigation withdrawal covariates is the most consequential gap in the feature set, for the reasons specified in Sections 6.3 and 6.4. The cross-model feature importance analysis (Chapter 5) further quantifies this gap: precipitation and NDVI consistently rank last, while land use and temperature—imperfect proxies for extraction demand—rank first, a pattern that would be expected if abstraction is the dominant but unmeasured driver.

Scope of evaluation. Beyond methodological contribution, these empirical findings assess the case for treating the Barind and Madhupur Tracts as priority regions for groundwater management. This convergence of monotonic long-term depletion trends, persistent cross-seasonal extraction pressure, and the spatial concentration of irrigated architecture within these tracts testifies a trajectory of accelerating water stress neither seasonal recharge nor current management frameworks appear sufficient to reverse. This conclusion we can assess to be consistent with the identification of groundwater depletion as a principal threat to food and water security as addressed in the Bangladesh Delta Plan 2100 [56, 20, 47].

At the same time, the Bengal Water Machine literature cautions against uniform, blanket restrictions on abstraction, specifically that dry-season pumping in favorable hydrogeological settings induces monsoon recharge and underwrites the boro harvest on which national food security depends on [70, 48]. From both bodies of evidence, the need for spatially targeted management is necessary—to permit sustainable conjunctive use where induced recharge operates, while restricting abstraction in the low-recharge Barind uplands where it does not. Skillful, seasonally resolved, spatially distributed forecasts of the kind developed here are precisely the instruments such targeting demands. Through incorporating abstraction covariates, finer resolution, and post-2020 observations, an operational ST-GTN forecasting system could support the Bangladesh Water Development board in distinguishing depletion hotspots well in advance of the dry season, to not only inform irrigation scheduling and extraction licensing, but also anticipate early warnings to communities dependent on shallow tubewells—offering a concrete decision-support capability to the implementation of the Delta Plan 2100.

Conclusion

Groundwater depletion in Bangladesh exposes one of the most pressing water security challenges in South Asia, furthermore driven by the intersection of a monsoon-dominated climate, intensive groundwater-fed irrigation, and the rapid urbanization of one of the world's most densely populated nations. Even so, the national capacity to anticipate groundwater change—rather than to simply monitor it—remains limited. This thesis addressed this gap in the existing forecasting literature by developing and evaluating spatiotemporal deep learning models capable of forecasting seasonal groundwater depth across the entirety of Bangladesh at a high spatial resolution through a multi-source covariate dataset spanning land use, vegetation, climate, and population dynamics from 2000 to 2020, built upon observations recorded from approximately 1,000 wells with usable records, drawn from the roughly 1,300-well Bangladesh Water Development Board (BWDB) network. It contributes to not only a methodological advance—the application of graph transformers to national-scale groundwater forecasting—but also to an empirical characterization of Bangladesh's groundwater system that connects long-term anthropogenic trends to seasonal predictive skill.

This thesis makes two coupled contributions. Methodologically, it offers—to the best of our knowledge—the first application of a graph transformer architecture to national-scale groundwater forecasting within a monsoon-driven deltaic aquifer system, combined with a systematic benchmark ran against graph-convolutional (ST-GNN) and convolutional-recurrent (CNN-LSTM) alternatives under identical data conditions. It empirically offers a characterization of Bangladesh's groundwater system that aligns long-term anthropogenic depletion trends with seasonal predictive skill, by recognizing where, when, and why forecasting is at its most difficult and what it implies for both model design and water management.

Summary of findings by Research Question

RQ1 — How do advanced spatiotemporal AI models (CNN-LSTM, ST-GNN, ST-GTN) compare when forecasting groundwater depth across Bangladesh's aquifer systems?

Of the three models, the Spatiotemporal Graph Transformer Network (ST-GTN) proved the strongest and most consistent performer across the validation (2017-2018), test (2019), and forecast (2020) periods. Although the ST-GNN was a clear improvement over the CNN-LSTM baseline, it generalized less robustly to the held-out test year. In contrast, the CNN-LSTM model performed the worst, highlighted by systematic under-prediction bias, negative NSE on the test set, and marked degradation in the 2020 monsoon season — not to mention, being the most computationally expensive model to train. ST-GTN additionally converged in only 23 epochs, compared to the 33 epochs for CNN-LSTM and 63 for ST-GNN. Its advantage was at its greatest precisely in the regions and seasons bearing the highest hydrological complexity: the Barind Tract in the dry season and the monsoon transition zones, to which we attribute to the model's self attention mechanism's capability of modeling long-term spatial dependencies and non-stationary temporal dynamics jointly. These findings align with—and extend to the seasonal deltaic setting—the emerging consensus and promise of transformer-style architectures generalizing much more robustly compared to convolutional or recurrent benchmarks in spatially heterogeneous hydrological systems [78].

RQ2 — How much do location and environmental covariates (temperature, rainfall, land use, vegetation) contribute to forecast accuracy?

The cross-model feature importance analysis—combining permutation importance [9] and gradient·input attribution [75] across all three architectures—produced a significant counterintuitive finding: NDVI and total precipitation ranked among the least important predictors for all three architectures, whilst minimum–maximum temperature, land use, thermal radiation, and latent heat flux showed consistently higher importance. Land use and maximum temperature emerged as the top two features in the cross-model combined rank. The consensus across architecturally distinct models strengthens confidence that these rankings reflect genuine physical drivers rather than artifacts of any single model's inductive biases. We interpret these rankings not as evidence that precipitation is hydrologically unimportant, but as the diagnostic signature of three interacting mechanisms: the weeks-to-months vadose-zone lag between rainfall and water-table response; the dominance of anthropogenic extraction over climatic recharge in the irrigated northwest, where land use and temperature function as imperfect proxies for irrigation-demand timing; and the structural absence of direct abstraction data from the covariate set. In heavily irrigated aquifer systems, feature sets built primarily from climate and vegetation indices are structurally incomplete; direct or proxy measures of abstraction should be treated as high-priority predictors in future work.

RQ3 — How robust are the models to the realities of Bangladesh's monitoring data—gaps, intermittent records, and a single-year forecast horizon?

Data conditions are what materially shaped model skill, for every architecture shared two failure modes. One, the systematic underestimation of peak groundwater depths in the Barind Tract, due to extraction-driven anomalies being spatially concentrated beneath the effective resolution of our patching scheme and are thus underrepresented in the covariate set. The second being a pronounced decline in monsoon-season skill in the forecast year (2020), being the most temporally distant year from training; anomalous hydrology emerges as pure forecast error. Though Quantile-mapping post-processing removed distributional bias in the pre-monsoon and winter forecasts. As a univariate per-location correction, it fails to repair the spatial structure of monsoon errors. This confirms that these residuals are more of a model-capacity problem rather than it being a calibration problem. A full sensitivity analysis to monitoring-network density and record completeness was beyond our scope of this study and is instead delegated as future work below.

RQ4 — Can combining AI with satellite remote sensing improve the basis for early warning of depletion-prone areas?

This thesis establishes three of the necessary components for such a basis, yet it stops short of a validated, operational early-warning system. First, the cross-model feature importance analysis covered in Chapter 5 showcases material improvements in forecast accuracy from satellite-derived covariates as opposed to just being there in the input stack. There, land-use ranks first in the cross-model combined importance score, and population density is ranked as the second most influential channel for CNN-LSTM in particular, which shows evidence that by combining AI with remote-sensing products, genuine, measurable predictive gains are recorded instead of the inputs ultimately being redundant. Second, the spatially and seasonally-resolved skill maps produced for each model—covering metrics RMSE, NSE, and Bias by season and region), identify precisely where and when forecasts are reliable enough to uphold monitoring decisions and where they are not. For example, the Barind Tract and monsoon transition zones highlighted reduced skill, which is itself the diagnostic basis that an early-warning system needs before flagging a region as high-confidence versus flagging one of uncertainty. Third, long-term population, land-use, and groundwater trend analysis independently converges upon the Barind and Madhupur Tracts as priority targets, serving the spatial prioritization component present in an early-warning framework. However, what this thesis did not provide is a validated, threshold-based warning capability, with the lack of a depletion threshold, lead-time evaluation, or calibrated probabilistic risk framework constructed or tested against actual historical depletion events. Rather than a completed contribution of this thesis, to convert the skillful, spatially continuous forecasts developed here into operational early-warning tools remains as a necessary next step for future work.

RQ5 — What institutional and technical constraints condition the operational use of such tools?

Rather than by formal analysis, this thesis engaged RQ5 strictly through direct experience and identified three decisive constraints. First, the restricted and discontinuous access to monitoring data, where records were unobtainable beyond 2020, and years roughly 2015 and onward were difficult to obtain. Second, the computational limits of two 8 GB GPUs capped our patching scheme to a patch size of 64 and a stride of 32, thus bounding the effective spatial resolution we could use. And finally, the absence of published irrigation-abstraction data at a usable resolution. These constraints are not incidental inconveniences; each constraint maps directly onto a documented failure mode of the models and constitute an empirical, ground-truth account of what deploying AI forecasting within Bangladesh's water institutions will ultimately demand.

Contributions

1. A benchmarked graph-transformer forecasting framework for national-scale seasonal groundwater prediction that demonstrates state-of-the-art skill (NSE up to 0.956) in a monsoon-driven, heavily anthropogenically forced deltaic aquifer—a significantly more challenging setting to forecast than the short-horizon, climate-driven cases previously reported.

2. An integrated multi-source spatiotemporal dataset (spanning 2000-2020) that merges BWDB well observations with climate, MODIS NDVI, land-use, and population layers, passed through a reproducible Dash/Pytorch processing pipeline for patch- and graph-based model training.

3. An empirical two-regime characterization of Bangladesh's groundwater system that separates recharge-responsive lowlands from structurally depleting uplands, quantified trend rates identifying the Barind and Madhupur Tracts as priority management targets.

4. A diagnosis of where and why forecasting fails—from extraction driven extremes, monsoon transition zones, and distant forecast horizon—with the demonstration that post hoc bias correction cannot substitute for model capacity, alongside a cross-model covariate-importance analysis (permutation importance and gradient·input attribution) that exposes anthropogenic demand—proxied by land use and temperature—to dominate predictability in irrigated regions, while precipitation and NDVI rank last due to vadose-zone lag and the masking of climatic recharge by extraction pressure [9, 75].

Future Work

The limitations identified in Chapter 6 outline a concrete research agenda, listed here and ordered roughly by expected impact.

Test spatial, not just temporal, generalization. The evaluation scheme used throughout this thesis withholds future timesteps but not geography: every well location, and every node in the fixed graph used by ST-GNN and ST-GTN, is present throughout training. Raised as a priority during the defense of this thesis, a regional holdout—splitting Bangladesh into multiple subsections (e.g., by administrative division or hydrogeological zone) and withholding an entire subsection's wells from training before evaluating on it—would directly test whether ST-GTN's strong validation and test skill (Chapter 6) reflects genuine, inductive learning of the covariate–groundwater relationship, or partly reflects transductive familiarity with the specific well-network topology it was trained on. This is a necessary step before treating the current results as evidence that the model would transfer to unmonitored locations, and should be prioritized alongside, or ahead of, the other extensions listed below.

Extend and densify the observational record. By securing BWDB observations beyond 2020, as well as recovering the sparsely covered post-2015 records and incorporating currently absent monitoring wells; models would be able to learn from a wider range of hydrological conditions, and thus reduce the disproportionate influence of any single anomalous year on the reported forecast skill. Multi-year rolling-origin evaluation instead of the single held-out forecast year would be executed. In addition, the climate covariate will instead be migrated to use ERA5 instead of ERA-Interim to extend the study period beyond 2020.

Incorporate direct abstraction covariates. The single most promising model improvement is the addition of irrigation-extraction information, such as tubewell density, pump electrification, abstraction-volume estimates, and/or irrigated-area maps from the Barind Multipurpose Development Authority and the BWDB. Section 6.4 argues that the Barind estimation is, at root, a missing-covariate problem; this is a test of that hypothesis.

Engineer lag-aware climate features. By constructing precipitation features at leads that are consistent with vadose-zone transit times (months for shallow water tables, longer for where the unsaturated zone is thick), it would directly test the delayed-response explanation for precipitation's low measured importance, and it may potentially restore a physically interpretable role for rainfall in the feature set.

Increase effective spatial resolution. Finer patching—enabled by larger GPU memory, gradient checkpointing, mixed-precision training, or more memory-efficient attention variants—targets the sharp lateral gradients of the monsoon transition zones and the spatially concentrated Barind anomalies that the current patch-64/stride-32 scheme smooths away.

Explore physics-informed and hybrid architectures. Embedded constraints on lateral flow, mass balance, or recharge physics into the loss function or architecture offers a principled route to correct the spatial structural errors that quantile mapping could not and possibly improve generalization under data-scarce conditions typical of the region.

Quantify predictive uncertainty. Operational decision support demands calibrated uncertainty rather than point forecasts. Deep ensembles, Monte Carlo dropout, or conformal prediction over the ST-GTN architecture would convert seasonal depth forecasts into probabilistic depletion-risk maps suited for early-warning thresholds.

Formalize the sensitivity analysis of RQ3. Systematic ablation over monitoring-network density, record completeness, and forecast horizon would establish minimum data conditions under which forecasts remains skillful—particularly the information the BWDB board needs in order to focus on network investments.

Move toward operational deployment. The end goal as framed in Chapter 1—an AI-based decision-support tool for BWDB in support of the Bangladesh Delta Plan 2100—requires packaging the forecasting pipeline for routine seasonal execution, integrating near-real-time data feeds, a nd co-designing outputs like hotspot maps, warning thresholds, and licensing-relevant summaries for the institutions that would use them.

Closing Statement

Beyond the methodological contributions, this thesis reinforces the empirical case for treating the Barind and Madhupur tracts as priority regions for groundwater management in Bangladesh. Through the convergence of monotonic long-term depletion, persistent cross-seasonal extraction pressure, and the spatial concentration of irrigated agriculture in these tracts, a trajectory of accelerating water stress that neither seasonal recharge nor current management frameworks could sufficiently reverse is highlighted. At the same time, the evidence assembled here, argues for spatially targeted intervention rather than it being uniform: dry-season pumping sustains national food security and, in more favorable hydrogeological settings, even induces monsoon recharge [70]—the case for restriction is specific to low-land recharge uplands, not to be generalized.

Forecasting tools of the kind developed in this study, once extended with richer abstraction data, finer resolution, and calibrated uncertainty, represent a practical pathway toward exactly that kind of evidence-based, spatially targeted groundwater governance—one capable of informing irrigation scheduling, extraction licensing, and climate-adaptation planning across some of the world's most groundwater-dependent agricultural systems.

Appendix A

Table 6: Full hyperparameter fine-tuning summary.

Model	Hyperparameters	Best Values	Search Space
CNN-LSTM	Epochs	33	[10:100]
	Channels	[64,96,128]	[[32:64],[64:128],[128:256]]
	LSTM Hidden	256	[64, 128, 192, 256]
	LSTM Layers	3	[1:3]
	Dropout	0.18993043718198038	[0.0:0.3]
	Use Attention	True	[True,False]
	Attn Heads	4	[2,4,8]
	Weight Decay	1.622401095006138e-05	[1e-8,1e-3]
	Learning Rate	0.0002113416820082882	[1e-5,5e-4]
	Grad Clip	0.6262099474925733	[0.5:2.0]
	Scheduler	plateau	[cosine,plateau]
	Batch Size	4	[4,8,16]
ST-GNN	Epochs	72	[10:100]
	G Hidden	96	[32,64,96,128]
	G Layers	2	[1,3]
	RNN Hidden	192	[64,128,192,256]
	Dropout	0.15721378117753168	[0.0:0.3]
	Learning Rate	0.00012137906507392256	[1e-5,5e-4]
	Weight Decay	1.5794812986984505e-07	[1e-8,1e-3]
	Grad Clip	1.0724741173967205	[0.5,2.0]
	Scheduler	cosine	[cosine,plateau]
	Batch Size	4	[4,8,16]
	Spatial Div Weight	0.1912049607917827	[0.0:1.0]
ST-GTN	D Model	96	[64,96,128,192,256]
	Number of Heads	2	[2,4,8]
	Epochs	72	[10:1000]
	Number of Layers	3	[2:5]
	D FF	768	[128,256,512,768,1024]
	Dropout	0.035878273781490506	[0.0:0.3]
	Learning Rate	0.0001628476512948763	[1e-5,5e-4]
	Weight Decay	6.366859160799428e-05	[1e-8:1e-3]
	Grad Clip	1.3419157963542443	[0.5:2.0]
	Scheduler	cosine	[cosine, plateau]
	Batch Size	4	[4,8,16]
	Spatial Div Weight	0.02157828539866089	[0.0:0.2]

Appendix B

Table 7: Time taken to load the dataset into data loaders across all three models (in seconds).

	Data Loading Time (in seconds)		
	CNN-LSTM	ST-GNN	ST-GTN
Build Data Stack	10.7 sec	11.2 sec	10.9 sec
Data Normalization	237.4 sec	253.6 sec	267.4 sec
Data Stack to Numpy	246.5 sec	247.2 sec	252.2 sec
Create Dataset	1.7 sec	2.5 sec	2.5 sec
Total Time	496.3 sec	514.4 sec	533 sec

References

[1] B. A. Aderemi et al. "Groundwater levels forecasting using machine learning models: A case study of the groundwater region 10 at Karst Belt, South Africa". In: *Systems and Soft Computing* 5 (2023), p. 200049.

[2] M. I. Adham et al. "Study on groundwater recharge potentiality of Barind Tract, Rajshahi District, Bangladesh using GIS and remote sensing technique". In: *Journal of the Geological Society of India* 75.2 (2010), pp. 432–438. DOI: 10.1007/s12594-010-0039-3.

[3] A. A. Ahmed et al. "Applications of machine learning to water resources management: A review of present status and future opportunities". In: *Journal of Cleaner Production* 441 (2024), p. 140715.

[4] T. Bai and P. Tahmasebi. "Graph neural network for groundwater level forecasting". In: *Journal of Hydrology* 616 (2023), p. 128792.

[5] M. I. Bhuyan et al. "Effect of soil and water salinity on dry season boro rice production in the south-central coastal area of Bangladesh". In: *Heliyon* 9 (2023), e19180.

[6] V. Boerger et al. *The State of the World's Land and Water Resources for Food and Agriculture – Systems at breaking point. Synthesis report 2021*. Tech. rep. Food and Agriculture Organization of the United Nations, 2021.

[7] H. Brammer. "Physical Geography and Environment". In: *Bangladesh: National Cultures and Heritage: An Introductory Reader*. Dhaka: Asiatic Society of Bangladesh, 2014, pp. 23–48.

[8] H. Brammer. "Floods, cyclones, drought and climate change in Bangladesh: a reality check". In: *International Journal of Environmental Studies* 73 (2016), pp. 865–886.

[9] L. Breiman. "Random Forests". In: *Machine Learning* 45.1 (2001), pp. 5–32. DOI: 10.1023/A:1010933404324.

[10] C. Chen et al. "A comparative study among machine learning and numerical models for simulating groundwater dynamics in the Heihe River Basin, northwestern China". In: *Scientific Reports* 10 (2020), p. 3904.

[11] K. Cho et al. "Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 1724–1734.

[12] N. T. Chowdhury. "Water management in Bangladesh: an analytical review". In: *Water Policy* 12 (2010), pp. 32–51.

[13] J. Chung et al. "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling". In: *arXiv preprint arXiv:1412.3555* (2014).

[14] R. Connor and M. Miletto. *The United Nations World Water Development Report 2023: partnerships and cooperation for water; executive summary*. Tech. rep. UNESCO, 2023.

[15] F. Corradini et al. "A Systematic Literature Review of Spatio-Temporal Graph Neural Network Models for Time Series Forecasting and Classification". In: *arXiv preprint arXiv:2410.22377* (2024).

[16] C. Dalin et al. "Groundwater depletion embedded in international food trade". In: *Nature* 543 (2017), pp. 700–704.

[17] D. P. Dee et al. "The ERA-Interim reanalysis: configuration and performance of the data assimilation system". In: *Quarterly Journal of the Royal Meteorological Society* 137 (2011), pp. 553–597.

[18] S. Emamgholizadeh, K. Moslemi, and G. Karami. "Prediction the Groundwater Level of Bastam Plain (Iran) by Artificial Neural Network (ANN) and Adaptive Neuro-Fuzzy Inference System (ANFIS)". In: *Water Resources Management* 28 (2014), pp. 5433–5446.

[19] J. S. Famiglietti. "The global groundwater crisis". In: *Nature Climate Change* 4 (2014), pp. 945–948.

[20] A. Getirana et al. "Avert Bangladesh's looming water crisis through open science and better data". In: *Nature* 610 (2022), pp. 626–629.

[21] T. Gleeson et al. "HESS Opinions: Improving the evaluation of groundwater representation in continental to global scale models". In: *Hydrology and Earth System Sciences Discussions* 2020 (2020), pp. 1–39.

[22] P. Goovaerts. "Geostatistics for Natural Resources Evaluation". In: *Applied Geostatistics Series* (1997).

[23] P. Goovaerts et al. "Geostatistical modeling of the spatial variability of arsenic in groundwater of southeast Michigan". In: *Water Resources Research* 41 (2005).

[24] A. Graves and J. Schmidhuber. “Framewise phoneme classification with bidirectional LSTM and other neural network architectures”. In: *Neural Networks* 18.5–6 (2005), pp. 602–610.

[25] P. Hess and N. Boers. “A generative machine learning approach for improving precipitation from Earth system models”. In: *arXiv preprint arXiv:2406.15026* (2024). Documents that quantile mapping, as a univariate per-grid-cell correction, cannot improve spatial patterns or temporal dynamics.

[26] S. Hochreiter and J. Schmidhuber. “Long short-term memory”. In: *Neural Computation* 9.8 (1997), pp. 1735–1780.

[27] M. Hocking and B. F. J. Kelly. “Groundwater recharge and time lag measurement through Vertosols using impulse response functions”. In: *Journal of Hydrology* 535 (2016), pp. 22–35. DOI: 10.1016/j.jhydrol.2016.01.042.

[28] J. Jeong, J. Jeong, and S. Moon. “Integrating physics-based constraints with gated recurrent unit for improved groundwater level fluctuation prediction”. In: *Journal of Hydrology* 662 (2025), p. 134018.

[29] E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open Source Scientific Tools for Python*. https://www.scipy.org/. Accessed: 2026. 2015.

[30] A. Kapoor et al. *Examining COVID-19 Forecasting using Spatio-Temporal Graph Neural Networks*. arXiv preprint arXiv:2007.03113. arXiv:2007.03113 [cs.LG]. 2020.

[31] T. N. Kipf and M. Welling. “Semi-Supervised Classification with Graph Convolutional Networks”. In: *International Conference on Learning Representations (ICLR)*. 2017.

[32] K. Klemmer, N. S. Safir, and D. B. Neill. “Positional encoder graph neural networks for geographic data”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2023, pp. 1379–1389.

[33] J. Koch et al. “High Resolution Water Table Modeling of the Shallow Groundwater Using a Knowledge-Guided Gradient Boosting Decision Tree Model”. In: *Frontiers in Water* 3 (2021), p. 701726. DOI: 10.3389/frwa.2021.701726.

[34] F. Kratzert et al. “Rainfall–runoff modelling using Long Short-Term Memory (LSTM) networks”. In: *Hydrology and Earth System Sciences* 22.11 (2018), pp. 6005–6022. DOI: 10.5194/hess-22-6005-2018.

[35] Y. LeCun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE*. Vol. 86. 11. 1998, pp. 2278–2324.

[36] B. Lim et al. “Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting”. In: *International Journal of Forecasting*. Vol. 37. 4. 2021, pp. 1748–1764.

[37] M. T. Liu et al. “Short-Term Prediction of Groundwater Level Based on Spatiotemporal Correlation”. In: *Water Resources* 51 (2024), pp. 207–220.

[38] D. Long et al. “Deriving scaling factors using a global hydrological model to restore GRACE total water storage changes for China’s Yangtze River Basin”. In: *Remote Sensing of Environment* 168 (2015), pp. 177–193.

[39] J. Long, E. Shelhamer, and T. Darrell. “Fully Convolutional Networks for Semantic Segmentation”. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 3431–3440.

[40] M. Luling et al. “GlobPOP: a gridded annual global population dataset from 1990 to 2020”. In: *Scientific Data* 10 (2023), p. 682.

[41] M. Mainuddin and M. Kirby. “National food security in Bangladesh to 2050”. In: *Food Security* 7 (2015), pp. 633–646.

[42] D. Maraun. “Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue”. In: *Journal of Climate* 26.6 (2013), pp. 2137–2143. DOI: 10.1175/JCLI-D-12-00821.1.

[43] J. Margat and J. Van der Gun. *Groundwater around the world: a geographic synopsis*. CRC Press, 2013.

[44] S. L. Markstrom et al. *GSFLOW-Coupled Ground-water and Surface-water FLOW model based on the integration of the Precipitation-Runoff Modeling System (PRMS) and the Modular Ground-Water Flow Model (MODFLOW-2005)*. Techniques and Methods 6. US Geological Survey, 2008, p. 240.

[45] G. May-Lagunes et al. “Forecasting groundwater levels using machine learning methods: The case of California’s Central Valley”. In: *Journal of Hydrology X* 21 (2023), p. 100161.

[46] M. A. Mirdad. “Investigation of Ground Water Table in the South-East (Chittagong) Part of Bangladesh”. In: *American Journal of Civil Engineering* 1 (2014), p. 53.

[47] M. Moshfika, S. Biswas, and M. S. Mondal. “Assessing Groundwater Level Declination in Dhaka City and Identifying Adaptation Options for Sustainable Water Supply”. In: *Sustainability* 14 (2022), p. 1518.

[48] A. Mukherji. “The “water machine” of Bengal”. In: *Science* 377.6612 (2022), pp. 1292–1293. DOI: 10.1126/science.ade0393.

[49] S. M. T. Mustafa et al. "Identification of the influencing factors on groundwater drought and depletion in north-western Bangladesh". In: *Hydrogeology Journal* 25.5 (2017), pp. 1357–1375. DOI: 10.1007/s10040-017-1547-7.

[50] V. Nourani, R. G. Ejlali, and M. T. Alami. "Spatiotemporal groundwater level forecasting in coastal aquifers by hybrid artificial neural network-geostatistics model: a case study". In: *Environmental Engineering Science* 28 (2011), pp. 217–228.

[51] V. Nourani, A. A. Mogaddam, and A. O. Nadiri. "An ANN-based model for spatiotemporal groundwater level forecasting". In: *Hydrological Processes* 22 (2008), pp. 5054–5066.

[52] V. Nourani et al. "Applications of hybrid wavelet–Artificial Intelligence models in hydrology: A review". In: *Journal of Hydrology* 514 (2014), pp. 358–377.

[53] A. H. Nury et al. "Assessment of groundwater variability using ARIMA, random forest, and LSTM-RNN in the northeastern region of Bangladesh". In: *H2Open Journal* 8 (2025), pp. 336–360.

[54] E. J. Pebesma. "Multivariable geostatistics in S: the gstat package". In: *Computers & Geosciences* 30 (2004), pp. 683–691.

[55] A. S. Qureshi, Z. U. Ahmad, and T. J. Krupnik. "Moving from Resource Development to Resource Management: Problems, Prospects and Policy Recommendations for Sustainable Groundwater Management in Bangladesh". In: *Water Resources Management* 29 (2015), pp. 4269–4283.

[56] A. S. Qureshi, Z. Ahmed, and T. J. Krupnik. *Groundwater management in Bangladesh: An analysis of problems and opportunities*. Tech. rep. Dhaka, Bangladesh: International Water Management Institute (IWMI) / CGIAR Research Program on Water, Land and Ecosystems, 2014.

[57] R Core Team. *R: A Language and Environment for Statistical Computing*. https://www.R-project.org/. Vienna, Austria: R Foundation for Statistical Computing, 2019.

[58] M. M. Rahman and A. Q. M. Mahbub. "Groundwater depletion with expansion of irrigation in Barind Tract: A case study of Tanore Upazila". In: *Journal of Water Resource and Protection* 4.8 (2012), pp. 567–575. DOI: 10.4236/jwarp.2012.48066.

[59] M. Rodell et al. "Emerging trends in global freshwater availability". In: *Nature* 557 (2018), pp. 651–659.

[60] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Springer. 2015, pp. 234–241.

[61] L. Rosa et al. “Global agricultural economic water scarcity”. In: *Science Advances* 6 (2020), eaaz6031.

[62] N. R. Rossman et al. “Vadose zone lag time and potential 21st century climate change effects on spatially distributed groundwater recharge in the semi-arid Nebraska Sand Hills”. In: *Journal of Hydrology* 519 (2014), pp. 656–669. DOI: 10.1016/j.jhydrol.2014.07.057.

[63] M. Sadat-Noori, W. Glamore, and D. Khojasteh. “Groundwater level prediction using genetic programming: the importance of precipitation data and weather station location on model accuracy”. In: *Environmental Earth Sciences* 79 (2019), p. 37.

[64] Zahraa Al Sahili and Mariette Awad. “Spatio-temporal graph neural networks: A survey”. In: *arXiv preprint arXiv:2301.10569* (2023).

[65] M. Sahoo et al. “Space–time forecasting of groundwater level using a hybrid soft computing model”. In: *Hydrological Sciences Journal* 62 (2017), pp. 561–574.

[66] B. R. Scanlon et al. “Groundwater depletion and sustainability of irrigation in the US High Plains and Central Valley”. In: *Proceedings of the National Academy of Sciences* 109 (2012), pp. 9320–9325.

[67] T. Shah. “The groundwater economy of South Asia: an assessment of size, significance and socio-ecological impacts”. In: *The agricultural groundwater revolution: Opportunities and threats to development*. 2007, pp. 7–36.

[68] M. Shamsudduha et al. “Recent trends in groundwater levels in a highly seasonal hydrological system: the Ganges–Brahmaputra–Meghna Delta”. In: *Hydrology and Earth System Sciences* 13.12 (2009), pp. 2373–2385. DOI: 10.5194/hess-13-2373-2009.

[69] M. Shamsudduha et al. “The impact of intensive groundwater abstraction on recharge to a shallow regional aquifer system: Evidence from Bangladesh”. In: *Hydrogeology Journal* 19.4 (2011), pp. 901–916. DOI: 10.1007/s10040-011-0723-4.

[70] M. Shamsudduha et al. “The Bengal Water Machine: Quantified freshwater capture in Bangladesh”. In: *Science* 377 (2022), pp. 1315–1319.

[71] Z. Sheikh Khozani et al. “Combining autoregressive integrated moving average with Long Short-Term Memory neural network and optimisation algorithms for predicting ground water level”. In: *Journal of Cleaner Production* 348 (2022), p. 131224. DOI: 10.1016/j.jclepro.2022.131224.

[72] X. Shi et al. “Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 28. 2015, pp. 802–810.

[73] A. H. M. Siddique-E-Akbor et al. “Satellite Precipitation Data–Driven Hydrological Modeling for Water Resources Management in the Ganges, Brahmaputra, and Meghna Basins”. In: *Earth Interactions* 18 (2014), pp. 1–25.

[74] S. Siebert et al. “Groundwater use for irrigation – a global inventory”. In: *Hydrology and Earth System Sciences* 14 (2010), pp. 1863–1880.

[75] K. Simonyan, A. Vedaldi, and A. Zisserman. *Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps*. arXiv preprint arXiv:1312.6034. arXiv:1312.6034 [cs.CV]. 2014.

[76] A. Singh. “Groundwater resources management through the applications of simulation modeling: A review”. In: *Science of the Total Environment* 499 (2014), pp. 414–423.

[77] A. Y. Sun et al. “Combining physically based modeling and deep learning for fusing GRACE satellite data: can we learn from mismatch?” In: *Water Resources Research* 55 (2019), pp. 1179–1195.

[78] W. Sun, L.-C. Chang, and F.-J. Chang. “Deep dive into predictive excellence: Transformer’s impact on groundwater level prediction”. In: *Journal of Hydrology* 636 (2024), p. 131250.

[79] M. L. Taccari et al. “Spatial-temporal graph neural networks for groundwater data”. In: *Scientific Reports* 14 (2024), p. 24564.

[80] A. Talib, A. R. Desai, and J. Huang. “Spatial and temporal forecasting of groundwater anomalies in complex aquifer undergoing climate and land use change”. In: *Journal of Hydrology* (2024), p. 131525.

[81] U. Udokpoh. “Analysis of Trend in Meteorological and Hydrological Time-series using Mann-Kendall and Sen’s Slope Estimator Statistical Test in Akwa Ibom State, Nigeria”. In: *International Journal of Environment and Climate Change* 13 (2023). DOI: 10.9734/ijecc/2023/v13i102748.

[82] UNESCO. *The United Nations World Water Development Report 2022: groundwater: making the invisible visible*. Tech. rep. UNESCO, 2022.

[83] A. Vaswani et al. “Attention Is All You Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[84] P. Veličković et al. “Graph Attention Networks”. In: *International Conference on Learning Representations (ICLR)*. 2018.

[85] M. T. Vu et al. “Reconstruction of missing groundwater level data by using Long Short-Term Memory (LSTM) deep neural network”. In: *Journal of Hydrology* 597 (2021), p. 125776.

[86] Y. Wada, D. Wisser, and M. F. Bierkens. “Global modeling of withdrawal, allocation and consumptive use of surface water and groundwater resources”. In: *Earth System Dynamics* 5 (2014), pp. 15–40.

[87] Y. Wada et al. “Global depletion of groundwater resources”. In: *Geophysical Research Letters* 37 (2010).

[88] X. Wang et al. “Short-term prediction of groundwater level using improved random forest regression with a combination of random features”. In: *Applied Water Science* 8 (2018), p. 125.

[89] K. Winkler et al. “Global land use changes are four times greater than previously estimated”. In: *Nature Communications* 12 (2021), p. 2501.

[90] R. F. Woodman et al. “Global land-use and land-cover change as a driver of climate impacts”. In: *Nature Climate Change* (2026). In press.

[91] B. K. Wossenyeleh et al. “Vadose zone lag time effect on groundwater drought in a temperate climate”. In: *Water* 12.8 (2020), p. 2123. DOI: 10.3390/w12082123.

[92] J. Wu, W. A. Norvell, and R. M. Welch. “Kriging on highly skewed data for DTPA-extractable soil Zn with auxiliary information for pH and organic carbon”. In: *Geoderma* 132 (2006), pp. 471–479.

[93] S. Wu et al. “Geographically and temporally neural network weighted regression for modeling spatiotemporal non-stationary relationships”. In: *International Journal of Geographical Information Science* 35 (2021), pp. 582–608.

[94] Z. Wu et al. “Graph WaveNet for Deep Spatial-Temporal Graph Modeling”. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI)*. 2019, pp. 1907–1913. DOI: 10.24963/ijcai.2019/264.

[95] A. Wunsch, T. Liesch, and S. Broda. “Groundwater level forecasting with artificial neural networks: A comparison of long short-term memory (LSTM), convolutional neural networks (CNNs) and non-linear autoregressive networks with exogenous input (NARX)”. In: *Hydrology and Earth System Sciences* 25.3 (2021), pp. 1671–1687. DOI: 10.5194/hess-25-1671-2021.

[96] M. Xu et al. “Spatial-Temporal Transformer Networks for Traffic Flow Forecasting”. In: *arXiv preprint arXiv:2001.02908*. 2020.

[97] S. Yan, Y. Xiong, and D. Lin. “Spatial Temporal Graph Convolutional Networks for Skeleton-Based Action Recognition”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 2018, pp. 7444–7452.

[98] X. Yang and Z. Zhang. “A CNN-LSTM model based on a meta-learning algorithm to predict groundwater level in the middle and lower reaches of the Heihe River, China”. In: *Water* 14 (2022), p. 2377.

[99] Z. Yin et al. “GNNWR: an open-source package of spatiotemporal intelligent regression methods for modeling spatial and temporal nonstationarity”. In: *Geoscientific Model Development* 17 (2024), pp. 8455–8468. DOI: 10.5194/gmd-17-8455-2024.

[100] H. Yoon et al. “A comparative study of artificial neural networks and support vector machines for predicting groundwater levels in a coastal aquifer”. In: *Journal of Hydrology* 396 (2011), pp. 128–138.

[101] B. Yu, H. Yin, and Z. Zhu. “Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting”. In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*. 2018, pp. 3634–3640. DOI: 10.24963/ijcai.2018/505.

[102] Cunjun Yu et al. “Spatio-temporal graph transformer networks for pedestrian trajectory prediction”. In: *European Conference on Computer Vision*. Springer. Cham, 2020, pp. 507–523.

[103] S. Yun et al. “Graph Transformer Networks”. In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019, pp. 11960–11970.

[104] X. Zhang et al. “Advance prediction of coastal groundwater levels with temporal convolutional and long short-term memory networks”. In: *Hydrology and Earth System Sciences* 27 (2023), pp. 83–96.