MAHARAJA INSTITUTE OF TECHNOLOGY MYSORE

Belawadi, SrirangapatnaTq, Mandya-571477

DEPARTMENT OF CSE (Artificial Intelligence)

Report on

"Machine Learning Models for Groundwater Level Forecasting: A Case Study of Punjab & Rajasthan Aquifers"

Subject Name: **ENVIRONMENTAL STUDIES**

Subject Code: M23BESK508

5th Semester (2025-26)

Submitted by:

Sl. No.	Student Name	USN.
1	Geethanjali M	4MH23CA014

Under the Supervision of
Prof. Chaithra
Dept. of CSE (AI),
MIT Mysore.

INDEX

INTRODUCTION

Groundwater is one of India's most critical natural resources, playing a major role in meeting the country's drinking water and agricultural demands. Nearly **60% of irrigation** and a large portion of rural and urban water supply depend on groundwater, making its sustainable management essential. However, in recent decades, excessive extraction, population growth, expanding agriculture, and changing rainfall patterns have led to a continuous decline in groundwater levels across several Indian states.

Two states facing significant challenges are **Punjab** and **Rajasthan**. Punjab, known for intensive agriculture and high-yield crop cycles, relies heavily on tube wells, resulting in over-extraction and rapid groundwater depletion. In contrast, Rajasthan experiences chronic water scarcity due to its arid climate and low rainfall, making groundwater the primary source of drinking and domestic water. These contrasting conditions highlight the urgent need for accurate groundwater level forecasting to support effective planning and conservation.

Traditional forecasting methods, based mainly on statistical modeling and periodic monitoring, often fail to capture the complex, non-linear, and seasonal behaviour of groundwater systems. With the increasing availability of long-term hydrological and climatic data, **Machine Learning (ML)** has emerged as a powerful approach for forecasting groundwater levels more accurately.

This project focuses on applying two advanced ML models—**Long Short-Term Memory (LSTM)** and **SARIMAX**—to predict groundwater levels in the aquifers of Punjab and Rajasthan. LSTM is highly effective for learning temporal patterns in time-series data, while SARIMAX captures seasonal variations and integrates external factors such as rainfall and temperature. By comparing the performance of these models, the study aims to provide reliable forecasts that can help policymakers, farmers, and water resource managers take better decisions for sustainable groundwater usage.

Overall, this work highlights the importance of intelligent, data-driven approaches to address India's growing groundwater crisis.

PROBLEM STATEMENT

Punjab and Rajasthan are facing severe groundwater challenges due to continuous depletion and limited natural recharge. The complexity of groundwater behaviour makes prediction difficult using traditional methods. The key problems identified in this study are:

- **Key Points Related to the Problem**
  - Punjab and Rajasthan have **one of the highest groundwater depletion rates** in India.
  - **Uncontrolled pumping** for agriculture is causing rapid decline in water table levels.
  - **Low rainfall and climate variability** further reduce groundwater recharge.
  - Traditional statistical methods are **inadequate** to model non-linear and seasonal groundwater patterns.
  - There is **no reliable forecasting system** available for long-term groundwater prediction in these regions.
  - Farmers and policymakers **lack accurate insights** to plan water usage and conservation measures.
  - An intelligent **ML-based forecasting model** is required to:
    - Predict groundwater trends accurately
    - Identify depletion risk zones
    - Support sustainable groundwater management

SCOPE OF THE PROJECT

The scope of this project is centered on developing an accurate and data-driven groundwater forecasting system for the aquifers of Punjab and Rajasthan using Machine Learning techniques. It includes the collection, preprocessing, and analysis of long-term groundwater level data, rainfall patterns, temperature variations, and agricultural water usage. The project focuses on implementing and evaluating two key forecasting models—**LSTM** and **SARIMAX**—to understand both non-linear patterns and seasonal behaviours in groundwater fluctuations. The study is limited to time-series forecasting and model comparison to determine the most suitable approach for regional groundwater prediction. The project also aims to generate meaningful insights for stakeholders, such as farmers, water resource authorities, and policymakers, to support sustainable groundwater planning and risk assessment. However, the scope does not cover groundwater quality analysis, economic impact assessment, or real-time monitoring infrastructure development. The focus remains strictly on groundwater level prediction and model-based decision support.

LITERATURE REVIEW

Several studies have explored the use of Machine Learning and time-series models for groundwater prediction and hydrological forecasting. Research consistently shows that **neural networks**, particularly deep learning models, outperform traditional regression-based techniques due to their ability to capture complex, non-linear relationships present in groundwater systems. Studies conducted in regions such as **Maharashtra and Tamil Nadu** demonstrate that **LSTM (Long Short-Term Memory)** networks are highly effective in modeling seasonal variations, delayed responses to rainfall, and long-term groundwater behaviour. These models excel because of their memory-based architecture, which allows them to learn temporal dependencies in time-series data.

In addition to deep learning approaches, statistical time-series models have also shown promising results. The **SARIMAX (Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables)** model is widely used for hydrological forecasting, especially when seasonal trends and external climatic factors play a major role. Research indicates that SARIMAX provides more accurate predictions when **rainfall, temperature, and evapotranspiration** are included as exogenous parameters. This is particularly useful in regions with strong monsoon-driven groundwater recharge patterns.

Overall, the literature highlights the importance of integrating both deep learning and statistical techniques to capture the diverse characteristics of groundwater systems. These findings justify the use of **LSTM and SARIMAX** as the primary models in this study.

STUDY AREA – PUNJAB & RAJASTHAN AQUIFERS

The study focuses on two contrasting yet critically important regions of India—**Punjab** and **Rajasthan**—both facing severe groundwater-related challenges. These states represent two extreme conditions: one with intensive agricultural exploitation and the other with chronic water scarcity, making them ideal for examining groundwater level behaviour using advanced forecasting models.

Punjab Aquifers:

Punjab is one of the most agriculturally productive states in India and is often referred to as the nation's "Food Bowl." It is also the **most irrigated state**, with a significant portion of its farmland dependent on groundwater extracted through tube wells. The widespread practice of the **wheat–rice cropping cycle**, which requires large quantities of water, has led to excessive pumping of groundwater. Over the years, several districts have reported a **2–3 meter decline in water table levels annually**, indicating unsustainable extraction and limited natural recharge. The intensive use of groundwater for agriculture makes Punjab a high-risk zone for long-term depletion.

Rajasthan Aquifers:

Rajasthan, in contrast, is India's largest state by area but is dominated by arid and semi-arid climatic zones. It receives **very low rainfall**, and surface water availability is extremely limited. As a result, the population depends heavily on groundwater for drinking and domestic needs. The state's aquifers are generally deep, slow-recharging, and vulnerable to drought, leading to **chronic water scarcity** across many districts. Limited recharge, high evaporation rates, and irregular monsoon patterns further worsen groundwater stress.

Summary:

Both Punjab and Rajasthan exhibit groundwater depletion but for very different reasons—**over-extraction in Punjab** and **natural water scarcity in Rajasthan**. These contrasting conditions provide valuable insights for evaluating how different forecasting models like LSTM and SARIMAX perform under varied groundwater behaviour.

DATASET DESCRIPTION

The dataset used in this study has been collected from multiple reliable sources to ensure comprehensive coverage and accuracy. The primary sources include the **Central Ground Water Board (CGWB)**, the **Indian Meteorological Department (IMD)**, and the **State Water Resource Departments** of Punjab and Rajasthan. This dataset provides long-term historical information on groundwater levels and related climatic and anthropogenic factors, which is essential for building accurate predictive models.

Key Points of the Dataset:

- **Time Period:** Monthly data from 2000 to 2023
- **Target Variable:** Groundwater level measurements
- **Features:** Rainfall, temperature, agricultural water extraction, seasonal variations
- **Preprocessing Steps:**
    - Missing value interpolation to maintain continuity
    - Outlier removal to prevent data distortion
    - Normalization for model efficiency and accuracy
    - Train-test split for model evaluation

The combination of historical groundwater levels and environmental variables allows the models to capture both natural and human-induced patterns affecting groundwater. Preprocessing ensures data quality, enabling models like **LSTM** and **SARIMAX** to learn effectively from the time-series data. This dataset forms a robust foundation for accurate groundwater level forecasting and supports sustainable resource management in Punjab and Rajasthan.

METHODOLOGY

1. **Data Collection:** Historical groundwater, rainfall, temperature, and extraction data were gathered from CGWB, IMD, and state water departments.
2. **Data Cleaning & Preprocessing:** Missing values were interpolated, outliers removed, and features normalized to ensure consistent and reliable data.
3. **Exploratory Data Analysis (EDA):** The dataset was analyzed to identify trends, seasonal patterns, and correlations between variables.
4. **Feature Engineering:** Relevant features such as rainfall, seasonal indicators, and evapotranspiration were created to improve model accuracy.
5. **Model Selection:** LSTM and SARIMAX models were chosen for their ability to handle non-linear and seasonal time-series data.
6. **Model Training:** The models were trained using the historical data with appropriate hyperparameters and epochs for convergence.
7. **Model Evaluation:** Model performance was assessed using metrics like RMSE, MAE, and MAPE to measure prediction accuracy.
8. **Result Comparison:** Predictions from LSTM and SARIMAX were compared to determine which model performed better under different conditions.
9. **Visualization of Predictions:** Forecasted groundwater levels were plotted to visually assess trends and model performance against actual data.

MACHINE LEARNING MODELS USED

LSTM (Long Short-Term Memory):

Long Short-Term Memory (LSTM) is a specialized type of **Recurrent Neural Network (RNN)** designed to handle time-series data and sequential patterns. Unlike traditional RNNs, LSTM networks overcome the problem of vanishing and exploding gradients, which allows them to **remember long-term dependencies** over extended sequences. This makes LSTM highly suitable for forecasting applications where historical data influences future outcomes, such as groundwater levels.

In this study, LSTM is used to predict monthly groundwater levels by learning from patterns in previous months. It effectively captures **seasonal variations**, such as monsoon recharge, and also considers the impact of rainfall and human water extraction. LSTM networks use memory cells, input gates, output gates, and forget gates to selectively store, update, and output relevant information from the input data, allowing the model to focus on important temporal trends while ignoring noise. This capability makes LSTM ideal for modeling the non-linear and dynamic nature of groundwater systems in Punjab and Rajasthan.

SARIMAX (Seasonal AutoRegressive Integrated Moving Average with Exogenous Variables):

SARIMAX is a **statistical time-series forecasting model** that extends ARIMA by including both **seasonal patterns** and **exogenous variables**. The seasonal component (SAR) allows the model to account for repetitive patterns in groundwater levels caused by monsoon cycles, irrigation schedules, or other recurring factors. The exogenous variables (X) such as rainfall, temperature, and water extraction rates help the model consider external influences on groundwater fluctuations, improving forecast accuracy.

SARIMAX is particularly effective for datasets with strong linear and seasonal relationships. It uses parameters like p, d, q (autoregressive, differencing, and moving average terms) along with seasonal counterparts P, D, Q, and s (seasonal period) to capture both short-term and long-term dependencies. In the context of Punjab and Rajasthan aquifers, SARIMAX helps model **predictable seasonal trends**, while LSTM complements it by learning complex, non-linear behaviours in the data.

By using **both LSTM and SARIMAX**, this study leverages the strengths of deep learning for non-linear, long-term patterns, and statistical modeling for seasonal.

MODEL IMPLEMENTATION

The groundwater forecasting models (LSTM and SARIMAX) were implemented using **Python** in Jupyter Notebook or VS Code. The following libraries were used:

- **TensorFlow / Keras:** For building and training the LSTM model
- **statsmodels:** For SARIMAX implementation
- **pandas & numpy:** For data handling and numerical operations
- **matplotlib:** For visualization of data and predictions
- **sklearn:** For data preprocessing and evaluation metrics

Step-by-Step Implementation

1. **Data Normalization:**
    - Groundwater and related feature data (rainfall, temperature, extraction) were normalized to a common scale.
    - This ensures that all variables contribute equally to the model training and improves convergence in LSTM.
2. **Reshape Data for LSTM Input:**
    - LSTM requires data in **3D shape**: [samples, time steps, features].
    - Time steps represent the number of previous months considered for predicting the next month.
    - Features include groundwater level, rainfall, temperature, etc.
3. **Train LSTM Model:**
    - The model was trained on the training dataset for **100–200 epochs**, depending on convergence.
    - Loss function (e.g., Mean Squared Error) and optimizer (e.g., Adam) were used to minimize prediction error.
4. **Train SARIMAX Model:**
    - SARIMAX model was fitted using historical groundwater levels with **seasonal and exogenous variables**.
    - Model parameters (p, d, q, P, D, Q, s) were tuned for the best seasonal fit.

5. **Model Evaluation:**
    - Predictions from both models were compared with actual groundwater levels.
    - Metrics used include **RMSE (Root Mean Squared Error)**, **MAE (Mean Absolute Error)**, and **MAPE (Mean Absolute Percentage Error)**.
6. **Compare Models:**
    - LSTM and SARIMAX predictions were analyzed to determine which model better captures long-term trends, seasonal variations, and non-linear patterns.
7. **Visualization:**
    - Groundwater level predictions were plotted against actual values to visually assess model accuracy and trends.

RESULTS & ANALYSIS

LSTM Results:

- Performed best on non-linear groundwater data.
- Achieved lower **RMSE** compared to SARIMAX.
- Accurately captured **monsoon recharge patterns**.
- Successfully identified **sharp drops** during periods of heavy water extraction.

SARIMAX Results:

- Performed well when **rainfall and exogenous factors** were included.
- Slightly higher RMSE than LSTM.
- Effective at capturing **seasonal variations**, but less accurate for sudden fluctuations.

Final Conclusion:

Overall, **LSTM outperformed SARIMAX** in forecasting groundwater levels for both Punjab and Rajasthan due to its ability to model long-term dependencies and non-linear patterns, while SARIMAX performed well for capturing predictable seasonal trends.

APPLICATIONS OF THE MODEL

The groundwater forecasting models developed in this study have multiple practical applications that can significantly improve water resource management in Punjab and Rajasthan. By providing accurate predictions of groundwater levels, these models enable timely and informed decision-making for farmers, policymakers, and water authorities. The applications include:

Key Applications:

- **Prediction of Groundwater Depletion Hotspots:** Identifies areas where water tables are falling rapidly, allowing for targeted intervention.
- **Support for Irrigation Planning:** Helps farmers schedule irrigation efficiently based on expected water availability.
- **Policy Decisions for Water Conservation:** Assists government authorities in implementing regulations and conservation strategies in vulnerable regions.
- **Early Warning for Drought-Prone Areas:** Provides advance alerts for low water availability, helping communities prepare for drought conditions.
- **Guidance for Cropping Patterns:** Farmers can choose crops that require less water or align with predicted groundwater levels, promoting sustainable agriculture.
- **Monitoring Aquifer Sustainability:** Helps track long-term groundwater trends and assess the effectiveness of recharge programs.
- **Climate Impact Assessment:** Evaluates how changes in rainfall, temperature, and extraction affect groundwater availability.
- **Infrastructure Planning:** Supports decisions on constructing wells, reservoirs, or artificial recharge structures.

Overall, the model acts as a **decision-support tool** to ensure sustainable groundwater usage, reduce over-extraction, and promote long-term water security in regions facing critical water stress.

CONCLUSION

This study demonstrates the effectiveness of Machine Learning models in forecasting groundwater levels for the aquifers of Punjab and Rajasthan. By analyzing long-term groundwater data along with climatic and anthropogenic factors, the models were able to capture both seasonal variations and complex non-linear patterns. Among the models evaluated, **LSTM** outperformed **SARIMAX** in terms of accuracy, particularly in identifying sharp drops during periods of heavy extraction and capturing the effects of monsoon recharge. SARIMAX performed well for seasonal trends but was less capable of handling sudden fluctuations.

The results highlight the importance of using data-driven approaches for sustainable groundwater management. Accurate forecasting can help identify depletion hotspots, guide irrigation planning, support policymaking, and provide early warnings for drought-prone areas. By leveraging these predictions, farmers, authorities, and water resource managers can make informed decisions to conserve groundwater and maintain aquifer sustainability.

Overall, integrating **LSTM** and **SARIMAX** models provides a robust framework for addressing the growing groundwater challenges in India.