# Crop Yield Forecasting

Anmol Gupta,

Bsc. (Hons) Statistics, 3rd Year,

Ramanujan College, University of Delhi

Project Guide / Mentor Name: Mr. Ankit Lodh

Period of Internship: 19th May 2025 - 15th July 2025

Report submitted to: IDEAS – Institute of Data Engineering, Analytics and Science Foundation, ISI Kolkata

# 1. Abstract

This project focuses on forecasting agricultural yield at a daily resolution using hydrological indicators such as reservoir Level and Current Live Storage. The aim is to build a state–crop-specific pipeline that leverages time series forecasting models (Prophet) to predict future reservoir behavior, followed by machine learning (Random Forest) for yield estimation. Historical data from 2010–2022 was collected and cleaned for multiple states and crops. For each pair, Prophet models were trained to forecast reservoir metrics 180 days into the future. These forecasts were then used to engineer lagged and rolling features that capture temporal water stress. Using these features, Random Forest models were trained to predict yield. The pipeline was executed in batch mode across all state–crop CSVs.

# 2. Introduction

Agriculture in India remains heavily dependent on timely and adequate water supply, which in turn is governed by reservoir storage and river regulation. Traditional yield forecasting relies on meteorological variables (rainfall, temperature) and satellite indices, but these can be noisy or unavailable at high frequency. By contrast, reservoir Level and Current Live Storage are directly measured, continuously recorded, and implicitly integrate the cumulative effect of upstream rainfall, evaporation, irrigation withdrawals, and catchment inflows. This project builds a daily-resolution forecasting pipeline that first uses **Prophet** to predict future reservoir behaviour and then leverages **Random Forest** regression to translate those hydrological forecasts into crop yield estimates for specific state–crop combinations.

The pipeline addresses several practical needs:

- **Relevance**: Enables planners and farmers to anticipate yield fluctuations weeks or months in advance, informing planting, irrigation scheduling, and market decisions.
- **Technology**: Combines time-series modelling (Facebook Prophet) with ensemble machine learning (scikit-learn's Random Forest) in a reproducible Python workflow.
- **Background Survey**: We reviewed literature on hydrological forecasting, crop modelling, and machine-learning-based yield prediction. Key references include reservoir inflow modelling, yield proxy selection, and feature-engineering best practices.
- **Procedure**:
    1. **Data Preparation & Shortlisting**: Collated daily reservoir metrics (FRL, Level, Current Live Storage), state level temperature and rainfall proxies, and shortlisted only those state–crop pairs that had fully consistent daily data (no large gaps) and known yield figures throughout this period—ensuring reliable model training and evaluation.
    2. **Forecasting**: Train Prophet on each series of reservoir Level and Storage to generate 180-day forecasts.
    3. **Feature Engineering**: Compute 7-day rolling means and 7-day lags of the forecasts to capture short-term hydrological memory.

4. **Yield Modeling**: Train Random Forest regressors on the historical overlap; evaluate with $R^2$, RMSE, and MAE.
5. **Production**: Batch-mode prediction of future yields for all state–crop combinations; export CSV and plot outputs.

- **Purpose**: Demonstrate that reservoir-based hydrological forecasts can serve as a robust, measured proxy for crop yield prediction at a daily scale—offering a practical tool for water resource managers and agricultural stakeholders.

### Topics Covered in the First Two Weeks

During the initial training phase of the internship, I received instruction on the following topics:

- **Power BI**
- **Research Project Introduction**
- **Career Design**
- **Prompt Engineering & Generative AI Introduction**
- **Questionnaire Design**
- **Survey Methodology**
- **Python Fundamentals**
    1. Basic syntax and data structures
    2. Functions and loops
    3. Object-Oriented Programming
- **Text Analytics**

# 3. Project Objective

• Build a reproducible, daily-level pipeline to forecast agricultural yield using reservoir data.

• Leverage Prophet to generate future Level & Storage forecasts per state–crop.

• Engineer lagged and rolling features to capture hydrological memory.

• Train and evaluate a Random Forest model on historical overlaps.

• Produce 180-day ahead yield forecasts for each state–crop combination.

# 4. Methodology

## 4.1 Data Collection and Shortlisting

- **Sources**: We began with six CSV files—one for each crop (gram, massor, mustard, potato, rabi rice, wheat)—containing daily state-level (Andhra Pradesh, Chhattisgarh, Gujarat, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Odisha, Rajasthan, Tamil Nadu, Telangana, Uttar Pradesh, Uttarakhand, West Bengal) reservoir metrics (FRL, Level, Current Live Storage), temperature (max/min), rainfall, and annual yield values for 2000–2022.
- **Shortlisting**: To ensure data consistency, we retained only those state–crop pairs that (a) had no large gaps in daily records over the entire period, and (b) possessed non-missing annual yield values for each year. Picking up all those state-crop pairs that had consistent values from 2010–2022 resulted in 10–12 valid state–crop CSVs per crop, each with approximately 4,700 daily rows.

## 4.2 Data Cleaning and Feature Engineering

- **Daily Features**: For each shortlisted CSV, we engineered the following daily features in the folder engineered_state_crop_csvs/:
    1. **avg_temp** = (temperature_max + temperature_min) / 2
    2. **water_stress** = 100 − Live Cap FRL
    3. **7-day rolling means** for rainfall, Level, and Storage
    4. **7-day lags** for rainfall, Level, and Storage
- **Implementation**: All data cleaning and feature creation were implemented in Python (pandas). The resulting files contain raw metrics and these engineered columns.

## 4.3 Reservoir Forecasting with Prophet

- **Objective**: Generate 180-day forecasts of reservoir Level and Current Live Storage for each state–crop.
- **Procedure**:
    1. **Time Series Extraction**: From each engineered CSV, select the historical series of “Level” (or “Current Live Storage”), drop missing values, and rename columns to ds (date) and y (value).
    2. **Model Training**: Fit a Prophet model with daily seasonality enabled.
    3. **Future Projection**: Use make_future_dataframe(periods=180) to extend the series and predict yhat for each future date.

- **Output**: Two 180-day forecast tables per state–crop—one for Level and one for Storage.

### 4.4 Forecast-Based Feature Reconstruction

- **Merging**: Join the Level and Storage forecasts on date.
- **Re-engineering**: Compute the same 7-day rolling means and 7-day lags on the forecasted values (e.g., level_7d_avg, level_7d_lag, etc.).
- **Cleaning**: Drop the first seven days (where lags are undefined) to produce a clean forecast-feature table (fc_df) for model input.

### 4.5 Yield Model Development

- **Data Alignment**: Merge fc_df with historical yield from the original engineered CSV on matching dates; drop any rows with missing yield.
- **Feature Matrix and Labels**:
    - **X** = [level_7d_avg, level_7d_lag, storage_7d_avg, storage_7d_lag]
    - **y** = yield
- **Train/Test Split**: Randomly shuffle and split 80% training / 20% testing (no fixed seed) to assess model robustness across varying seasons.
- **Model Selection**: A RandomForestRegressor (200 trees) was chosen for its ability to capture non-linear interactions with minimal hyperparameter tuning.
- **Validation**: Evaluate on the test set using $R^2$, RMSE, and MAE.

### 4.6 Production Forecasting

- **Future Prediction**: Apply the trained RF model to the 180-day forecast-feature table (fc_df) beyond the historical date range to obtain predicted daily yields.
- **Output Storage**: Save per state–crop future yield forecasts to predicted_yield/ as CSV and corresponding line-plot PNG.

### 4.7 Code and Reproducibility

- **GitHub Repository**: [https://github.com/MRANMERA/Yield_Analysis](https://github.com/MRANMERA/Yield_Analysis)
- **EDA**: yield_analysis_EDA.ipynb contains the exploratory data analysis and initial visualizations.
    - [https://colab.research.google.com/drive/1KEbuE1gvumeAPXGOoqB95bSSDmpxkCXH?usp=sharing](https://colab.research.google.com/drive/1KEbuE1gvumeAPXGOoqB95bSSDmpxkCXH?usp=sharing)

- **Forecast Pipeline**: yield_analysis_forecast.ipynb implements the full forecasting and yield prediction pipeline.
    - https://colab.research.google.com/drive/1SH5RzczZVeEMcHxzh7emm7q_wrSjAaQi?usp=sharing

# 5. Data Analysis and Results

## 5.1 Data Summary

| state_name | crop_name | apy_item_interval_st | temperature_recorde | state_temperature_m | state_temperature_m | state_rainfall_val | yield | FRL | Live Cap FRL | Level | Current Live Storage |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Andhra Pradesh | gram | 2000 | 2000-01-01 | 30.38 | 14.47 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 266.3 | 6.39 |
| Andhra Pradesh | gram | 2000 | 2000-01-02 | 30.04 | 13.96 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 266.18 | 6.33 |
| Andhra Pradesh | gram | 2000 | 2000-01-03 | 29.92 | 12.98 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 266.09 | 6.286 |
| Andhra Pradesh | gram | 2000 | 2000-01-04 | 29.98 | 12.23 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 266.03 | 6.257 |
| Andhra Pradesh | gram | 2000 | 2000-01-05 | 29.77 | 13.24 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 265.97 | 6.228 |
| Andhra Pradesh | gram | 2000 | 2000-01-06 | 30.42 | 12.31 | 0 | 1.22615 | 152.2966667 | 2.838333333 | 266.3 | 6.39 |

**Sample of the dataset from merged_gram_reservoir.csv**

| **Column Name** | **Meaning** |
|---|---|
| **state_name** | Indian state where data is recorded |
| **crop_name** | Type of crop (e.g., gram, wheat, etc.) |
| **apy_item_interval_start** | Year of sowing season (e.g., 2000) |
| **temperature_recorded_date** | Date of weather data (daily) |
| **state_temperature_max_val** | Max temperature recorded on the day |
| **state_temperature_min_val** | Min temperature recorded on the day |
| **state_rainfall_val** | Rainfall in mm on the day |
| **yield** | Yield |
| **FRL** | Full Reservoir Level (Water storage limit) |
| **Live Cap FRL** | Percentage of storage relative to FRL |
| **Level** | Water level in the reservoir |
| **Current Live Storage** | Actual water stored on the date |

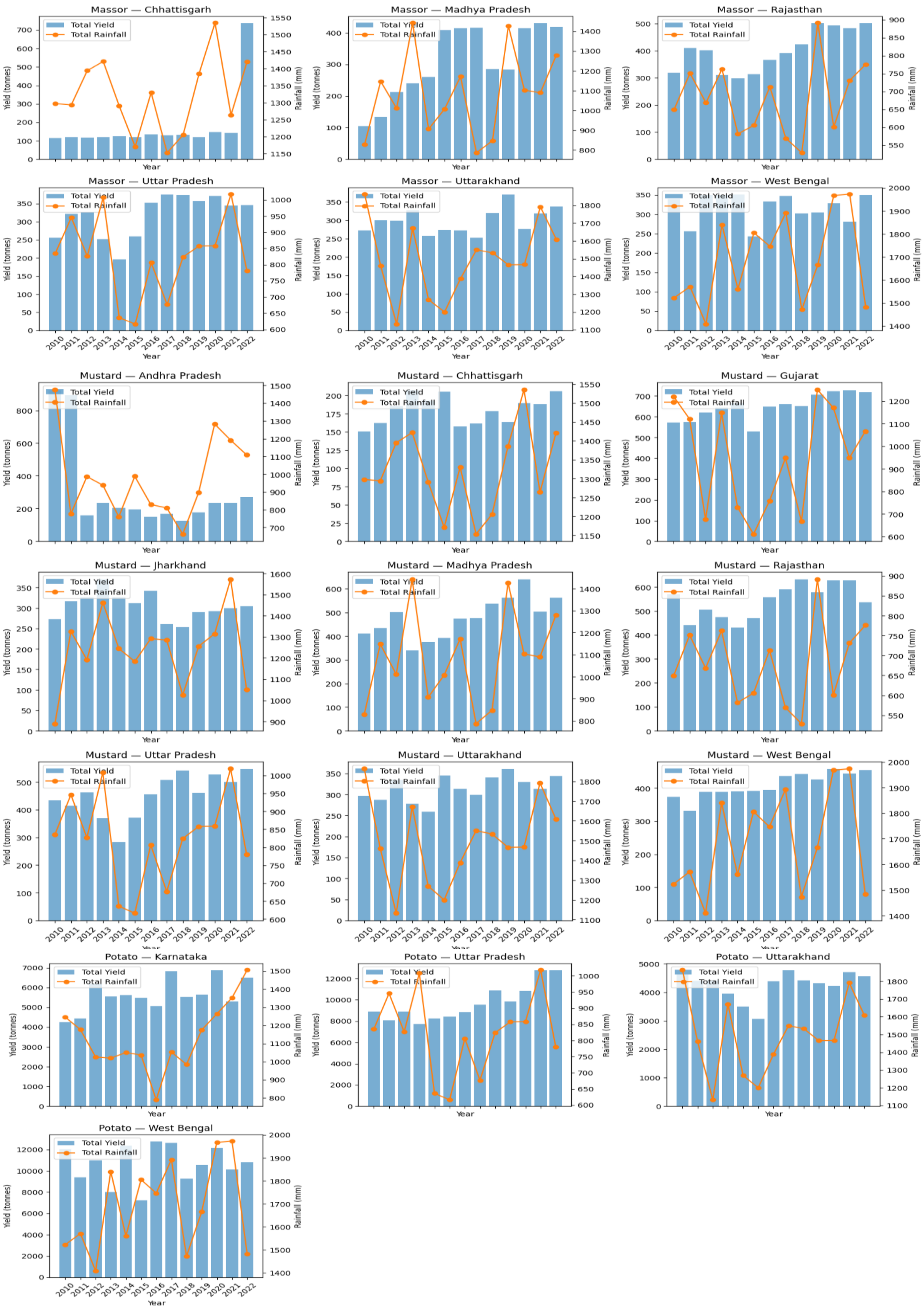
| **Item** | **Details** |
|---|---|
| **Time Frame** | 2000–2022 (raw), filtered to 2010–2022 for modelling |
| **Number of States** | 14 (AP, CH, GJ, JH, KA, MP, MH, OD, RJ, TN, TL, UP, UK, WB) |
| **Columns with missing entries** | Level, Current Live Storage |
| **Missing Patterns** | Inconsistent gaps by year/state (manual entry issues) |
| **Daily Rows per CSV** | ~4,700 |

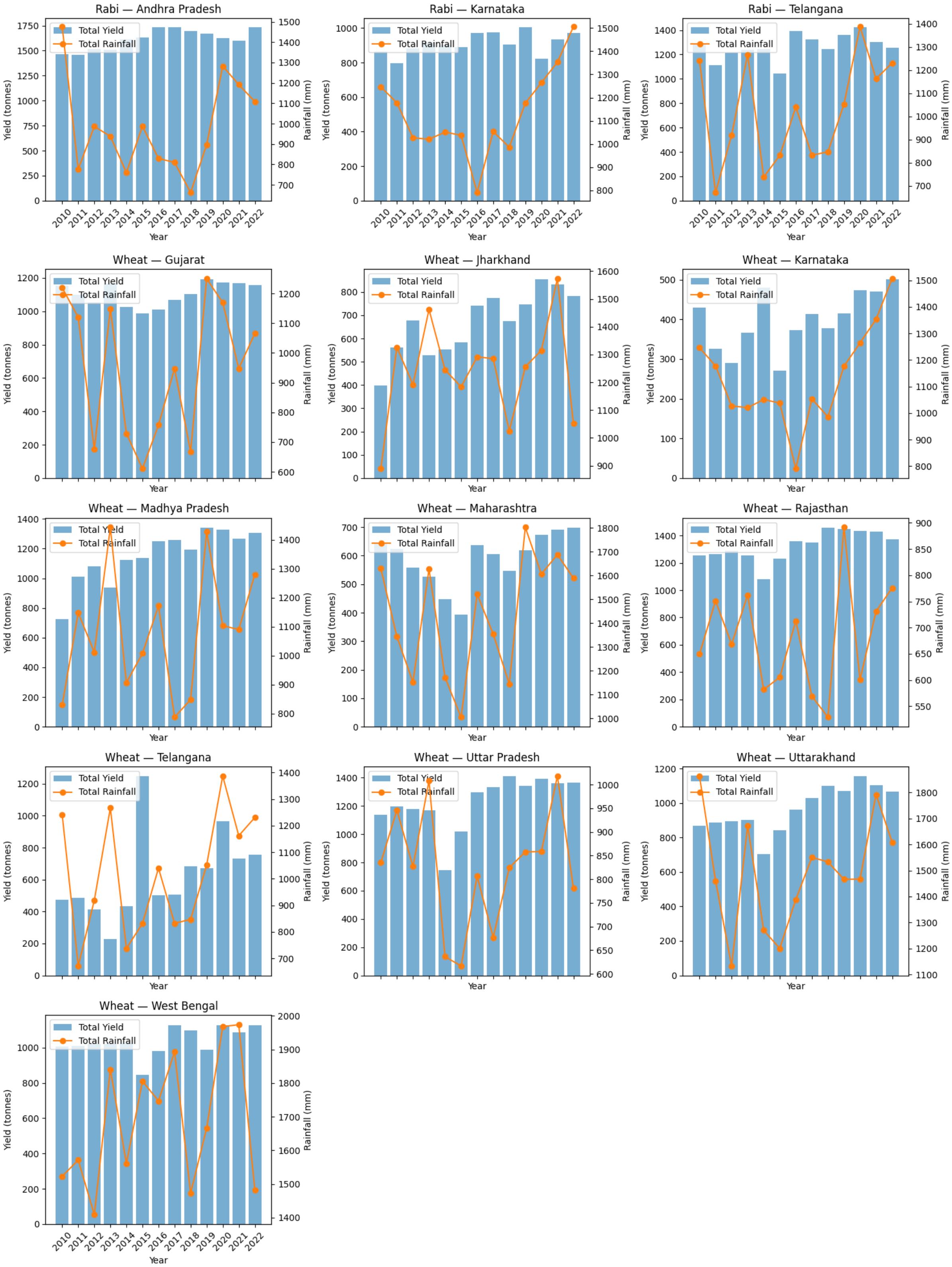
**Note**: To ensure reliable training, we **shortlisted only those state–crop pairs** with complete daily data and known annual yield for every year from 2010 through 2022.

## 5.2 Shortlisted State–Crop Pairs

| **Crop** | **States** |
|---|---|
| **Gram** | Andhra Pradesh, Chhattisgarh, Gujarat, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Rajasthan, Telangana, Uttar Pradesh, Uttarakhand, West Bengal |
| **Massor** | Chhattisgarh, Madhya Pradesh, Rajasthan, Uttar Pradesh, Uttarakhand, West Bengal |
| **Mustard** | Andhra Pradesh, Chhattisgarh, Gujarat, Jharkhand, Madhya Pradesh, Rajasthan, Uttar Pradesh, Uttarakhand, West Bengal |
| **Potato** | Karnataka, Uttar Pradesh, Uttarakhand, West Bengal |
| **Rabi Rice** | Andhra Pradesh, Karnataka, Telangana |
| **Wheat** | Gujarat, Jharkhand, Karnataka, Madhya Pradesh, Maharashtra, Rajasthan, Telangana, Uttar Pradesh, Uttarakhand, West Bengal |

## 5.3 Rainfall and yield patterns.


Massor — Chhattisgarh
Massor — Madhya Pradesh
Massor — Rajasthan
Massor — Uttar Pradesh
Massor — Uttarakhand
Massor — West Bengal
Mustard — Andhra Pradesh
Mustard — Chhattisgarh
Mustard — Gujarat
Mustard — Jharkhand
Mustard — Madhya Pradesh
Mustard — Rajasthan
Mustard — Uttar Pradesh
Mustard — Uttarakhand
Mustard — West Bengal
Potato — Karnataka
Potato — Uttar Pradesh
Potato — Uttarakhand
Potato — West Bengal
Total Yield
Total Rainfall
Yield (tonnes)
Rainfall (mm)
Year




To investigate the climatic dependencies and temporal dynamics of agricultural yield, we analyzed **rainfall and yield patterns** for all shortlisted state–crop combinations across the study period (2010–2022). This step was essential to determine whether **generalized modeling** would suffice or if **individual models** were necessary for different regions and crops.

- **Visualizing Annual Trends**

We generated year-wise time series plots for rainfall and yield to observe how these two variables behaved over time. The visualizations revealed several distinct rainfall patterns:

- In some regions, rainfall followed a stable and predictable pattern every year.
- Other regions displayed significant year-to-year fluctuations, including sharp declines or spikes in monsoon activity, directly affecting reservoir storage and agricultural output.

These differences made it evident that rainfall's influence on yield varies substantially depending on the geographical and agricultural context.

- **Need for Localized Modeling**

Based on the rainfall–yield visualizations:

- We observed that a single model cannot capture the heterogeneity across combinations.
- Each combination presented unique patterns and dependencies, influenced by microclimates, irrigation infrastructure, and crop calendar.

Hence, we adopted a strategy of individualized modeling for each state–crop pair, allowing us to train models that are sensitive to local agro-climatic characteristics.

- **Multi-Metric Temporal Profiling**

To support our understanding further, we developed multi-metric monthly trend plots, which visualized:

- Monthly rainfall
- Monthly yield
- Reservoir Level
- Current Live Storage

These were plotted across multiple years and segregated by state, enabling comparisons across different growing seasons. This helped us identify:

- Whether rainfall and reservoir levels were synchronized.
- If storage metrics served as a more reliable proxy for water availability than direct rainfall.
- How yield lagged or aligned with water availability indicators.

The plots used faceted visualizations, where each metric had its own subplot, and years were color-coded. This allowed clear interpretation of inter-metric relationships on a month-by-month basis.

- **Key Insights**

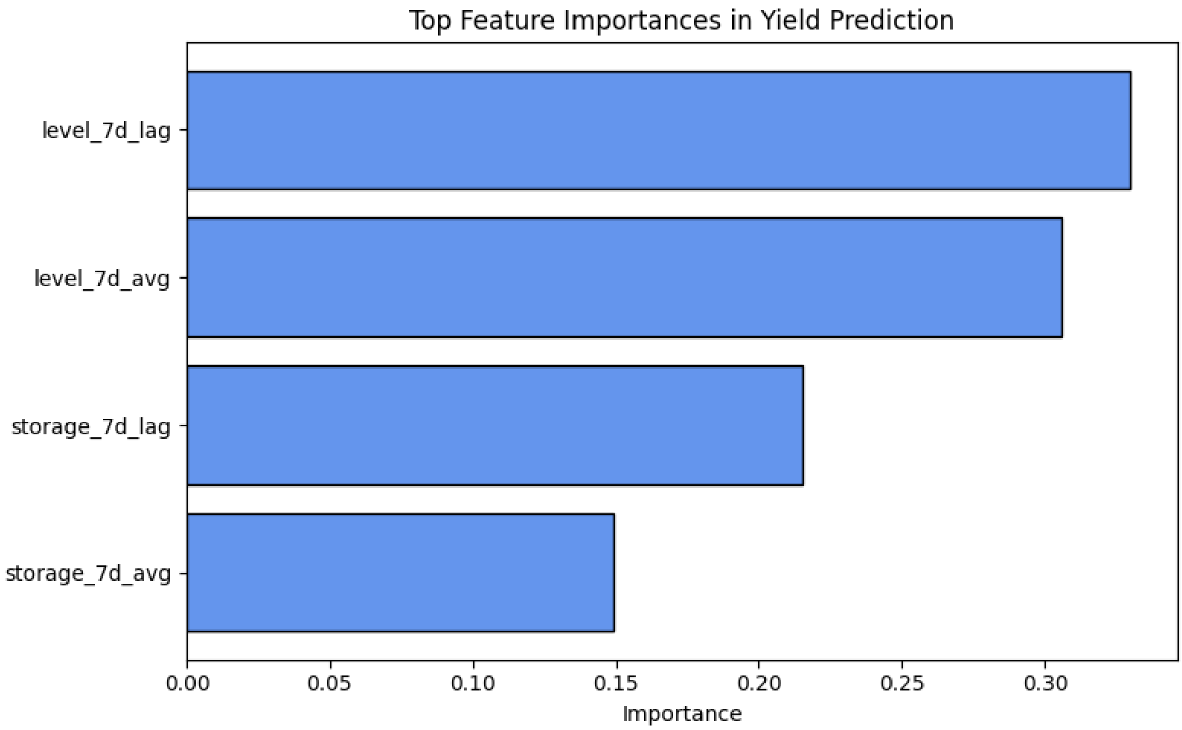
- In some states, rainfall timing and magnitude were consistent year over year, simplifying the forecasting challenge.
- In others, variability in water availability—either from rainfall or storage—made it necessary to incorporate lag features and rolling averages.
- In certain cases, yield seemed to track storage metrics more closely than direct rainfall, highlighting the relevance of incorporating engineered water-based features into the model.

## 5.4 Feature Engineering

For each state–crop CSV, we computed:

| Feature | Description |
|---|---|
| `avg_temp` | (max_temp + min_temp) / 2 |
| `water_stress` | 100 − Live Cap FRL |
| `rainfall_7d_avg` | 7-day rolling mean of rainfall |
| `level_7d_avg` | 7-day rolling mean of forecasted reservoir level |
| `storage_7d_avg` | 7-day rolling mean of forecasted live storage |
| `rainfall_7d_lag` | 7-day lag of rainfall |
| `level_7d_lag` | 7-day lag of forecasted reservoir level |
| `storage_7d_lag` | 7-day lag of forecasted live storage |

These features capture both the recent trend and short-term memory of water availability, which are critical drivers of crop yield.




### 5.4 Model Performance Summary

To evaluate the performance of our yield forecasting pipeline, we trained and tested a **Random Forest regression model** on each shortlisted state–crop pair, using consistent data from 2010–2022. Each model was trained using reservoir-derived features (rolling averages and lags of Level and Current Live Storage), and performance was evaluated using standard regression metrics:

- **R² Score (Coefficient of Determination)**
- **RMSE (Root Mean Squared Error)**
- **Absolute Error)**

| | state_crop | n_samples | R2 | RMSE | MAE |
|---|---|---|---|---|---|
| 12 | massor_chhattisgarh | 4647 | 0.974179 | 0.071017 | 0.014027 |
| 31 | rabi_rice_andhrapradesh | 4734 | 0.881495 | 0.088082 | 0.045247 |
| 1 | gram_chhattisgarh | 4647 | 0.869774 | 0.082258 | 0.046204 |
| 38 | wheat_maharashtra | 4741 | 0.845738 | 0.099009 | 0.056889 |
| 18 | mustard_andhrapradesh | 4734 | 0.799160 | 0.326810 | 0.128769 |
| 6 | gram_maharashtra | 4741 | 0.787229 | 0.075138 | 0.043224 |
| 11 | gram_westbengal | 4639 | 0.773797 | 0.071240 | 0.040313 |
| 26 | mustard_westbengal | 4639 | 0.772123 | 0.047632 | 0.029492 |
| 2 | gram_gujarat | 4733 | 0.766662 | 0.148098 | 0.089154 |
| 5 | gram_madhyapradesh | 4741 | 0.746525 | 0.153687 | 0.100415 |
| 19 | mustard_chhattisgarh | 4647 | 0.741579 | 0.026410 | 0.015082 |
| 40 | wheat_telangana | 4741 | 0.738262 | 0.351579 | 0.191881 |
| 43 | wheat_westbengal | 4639 | 0.731588 | 0.103015 | 0.059088 |
| 33 | rabi_rice_telangana | 4741 | 0.729454 | 0.141870 | 0.082959 |
| 22 | mustard_madhyapradesh | 4741 | 0.727695 | 0.116149 | 0.083003 |
| 37 | wheat_madhyapradesh | 4741 | 0.723926 | 0.239402 | 0.137930 |
| 0 | gram_andhrapradesh | 4734 | 0.721157 | 0.121607 | 0.066922 |
| 9 | gram_uttarakhand | 4719 | 0.718267 | 0.015801 | 0.010311 |
| 8 | gram_telangana | 4741 | 0.702645 | 0.148576 | 0.080216 |
| 15 | massor_uttarakhand | 4719 | 0.696096 | 0.051361 | 0.032958 |
| 29 | potato_uttarpradesh | 4715 | 0.694450 | 2.459801 | 1.590087 |
| 41 | wheat_uttarakhand | 4719 | 0.689473 | 0.191491 | 0.121665 |
| 17 | massor_westbengal | 4639 | 0.673466 | 0.057104 | 0.036161 |
| 28 | potato_uttarakhand | 4719 | 0.670658 | 0.759603 | 0.446217 |
| 20 | mustard_gujarat | 4733 | 0.669725 | 0.094050 | 0.057851 |
| 30 | potato_westbengal | 4639 | 0.650202 | 2.728195 | 1.701881 |
| 34 | wheat_gujarat | 4733 | 0.640349 | 0.112819 | 0.073870 |
| 13 | massor_madhyapradesh | 4741 | 0.618105 | 0.187542 | 0.119047 |
| 7 | gram_rajasthan | 4741 | 0.612743 | 0.087215 | 0.053102 |
| 14 | massor_rajasthan | 4741 | 0.610336 | 0.128251 | 0.076346 |
| 35 | wheat_jharkhand | 4633 | 0.605993 | 0.222391 | 0.159114 |
| 3 | gram_jharkhand | 4633 | 0.595132 | 0.117337 | 0.079339 |
| 24 | mustard_uttarakhand | 4719 | 0.570559 | 0.052303 | 0.034488 |
| 23 | mustard_rajasthan | 4741 | 0.569359 | 0.119980 | 0.079769 |
| 39 | wheat_rajasthan | 4741 | 0.558938 | 0.193640 | 0.124214 |
| 21 | mustard_jharkhand | 4633 | 0.553819 | 0.056505 | 0.040055 |
| 10 | gram_uttarpradesh | 4715 | 0.529450 | 0.238939 | 0.153998 |
| 42 | wheat_uttarpradesh | 4715 | 0.487578 | 0.352834 | 0.216177 |
| 4 | gram_karnataka | 4741 | 0.471764 | 0.085904 | 0.060172 |
| 16 | massor_uttarpradesh | 4715 | 0.471338 | 0.110586 | 0.070448 |
| 25 | mustard_uttarpradesh | 4715 | 0.467534 | 0.148292 | 0.098801 |
| 32 | rabi_rice_karnataka | 4741 | 0.455307 | 0.115528 | 0.079846 |
| 27 | potato_karnataka | 4741 | 0.440701 | 1.602868 | 1.153899 |
| 36 | wheat_karnataka | 4741 | 0.406259 | 0.148866 | 0.105079 |

### Overall Performance Range

- The best performing models achieved $R^2$ scores above 0.85, with minimal RMSE and MAE, indicating a strong ability to capture the relationship between reservoir metrics and yield.

- A majority of combinations yielded $R^2$ scores between 0.6 and 0.85, suggesting moderate to good predictability, depending on data consistency and crop–region dynamics.

- Some models recorded $R^2$ values below 0.5, often accompanied by high RMSE and MAE. These cases reflected challenges in predictability, likely due to:

- Highly erratic rainfall and storage behavior,
- Low correlation between yield and water availability metrics,
- Potential influence of unobserved confounders (e.g., pests, fertilizers, local management).

**Feature Importance Interpretation**

Across almost all models, the 7-day lagged and averaged reservoir metrics had varying importance. This confirms that the time-delayed effect of water stress is critical in determining yield outcomes—a biologically grounded insight, as plant stress responses are rarely instantaneous.

Interestingly:

- In better-performing models, average storage and level consistently held higher importance than their lagged counterparts.
- In less consistent models, no single feature dominated, implying more complex or chaotic systems influencing yield.

**Summary of Results**

A total of 44 models were trained and evaluated. Highlights:

- $R^2$ scores ranged from 0.97 (near-perfect fit) to 0.40 (weak predictive power).
- MAE ranged from ~0.01 to ~2.0, with lower MAE indicating better precision.
- In terms of RMSE, well-performing models stayed under 0.1, while weaker ones exceeded 1.5.

This performance distribution is a strong indicator that context-aware modeling is essential, and future work can build upon this framework by integrating additional agricultural and climatic parameters.

# 6. Conclusion

Pursuing this reservoir-based yield prediction project has been a enlightening experience of hydrology and agriculture's interplay. Our per state-crop model method, confined to combinations with full daily data for 2010-2022, produced a wide range of predictive abilities:

- Some models achieved exceptionally high $R^2$ scores (up to **0.97**), which—while impressive—may hint at slight overfitting to historical idiosyncrasies.

- The bulk of models fell within an **$R^2$ range of 0.80 down to 0.40**, demonstrating solid performance for many regions and highlighting areas where water availability is only one piece of the yield puzzle.

A particularly **striking finding** was the dominant influence of **7-day rolling averages** of reservoir Level and Storage. In high-performing cases, these two features often comprised **80–90% of the model's importance**, confirming that **sustained water availability**, rather than day-to-day fluctuations, drives crop productivity. Conversely, in lower-performing models, the more evenly spread feature importances suggested that **additional factors**—such as soil fertility, extreme temperatures, or agronomic practices—need to be incorporated.

On a personal level, I'm struck by how **localized behaviors** demand **localized solutions**. Regions with consistent monsoon-driven reservoir cycles benefited greatly from our streamlined pipeline; areas with erratic rainfall required a deeper dive into supplementary data sources. This reinforced the principle that effective agricultural analytics must be **context-aware**.

**Recommendations for future work**:

1. **Broaden Feature Set**: Introduce temperature extremes, soil moisture indices, or fertilizer usage to capture missing variance in underperforming pairs.
2. **Multi-Scale Temporal Features**: Experiment with longer lags (14- or 30-day rolling averages) and seasonal lag variables to capture extended hydrological memory.
3. **Advanced Modeling Techniques**: Evaluate gradient-boosting frameworks (XGBoost, LightGBM) or quantile regression forests to both enhance accuracy and provide predictive intervals for risk assessment.
4. **Operational Deployment**: Integrate this pipeline into a real-time dashboard that automatically ingests new reservoir readings, updates forecasts, and delivers actionable yield projections to stakeholders.