State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators

BITE497J – Project I

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Information Technology

by

22BIT0140 – Nachiketa Shrivastava

22BIT0114 – Chetan Yadav

22BIT0505 – Soumyadip Das

Under the guidance of

Prof. JERART JULUS L

School of Computer Science Engineering and Information Systems

VIT, Vellore

VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

November, 2025

DECLARATION

I hereby declare that the BITE497J – Project I thesis entitled **State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators** submitted by me, for the award of the degree of *Bachelor of Technology in Information Technology, School of Computer Science Engineering and Information Systems* to VIT is a record of bonafide work carried out by me under the supervision of Prof. Jerart Julus L, Assistant Professor, SCORE, VIT, Vellore.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date: 5/11/25

S. Das

Signature of the Candidate

CERTIFICATE

This is to certify that the BITE497J – Project I thesis entitled **State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators** submitted by Nachiketa Shrivastava (22BIT0140), Chetan Yadav (22BIT0114) and Soumyadip Das (22BIT0505), SCORE, VIT, for the award of the degree of *Bachelor of Technology in Information Technology, School of Computer Science Engineering and Information Systems*, is a record of bonafide work carried out by them under my supervision during the period, 21. 07. 2025 to 30.11.2025, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore

Date: 5/11/25

5/11/25

Signature of the Guide

Internal Examiner

5/11/25

External Examiner

Head of the Department

Department of Information Technology

ACKNOWLEDGEMENT

It is my pleasure to express with a deep sense of gratitude to my BITE497 - Project I guide Prof. Jerart Julus L, Assistant Professor, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore for his/her constant guidance, continual encouragement, in my endeavor. My association with him/her is not confined to academics only, but it is a great opportunity on my part to work with an intellectual and an expert in the field of Time Series Analysis and Machine Learning.

"I would like to express my heartfelt gratitude to Honorable Chancellor Dr. G Viswanathan; respected Vice Presidents Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Vice Chancellor Dr. V. S. Kanchana Bhaaskaran; Pro-Vice Chancellor Dr. Partha Sharathi Mallick; and Registrar Dr. Jayabarathi T.

My whole-hearted thanks to Dean Dr. Daphne Lopez, School of Computer Science Engineering and Information Systems, Head, Department of Information Technology, Dr. Arivuselvan K, Information Technology Project Coordinator Dr. Suganya P, SCORE School Project Coordinator Dr. Thandeeswaran R, all faculty, staff and members working as limbs of our university for their continuous guidance throughout my course of study in unlimited ways

It is indeed a pleasure to thank my parents and friends who persuaded and encouraged me to take up and complete my project successfully. Last, but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of the project.

Place: Vellore

Date: 5/11/25

Nachiketa Shrivastava

Chetan Yadav

Soumyadip Das

State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators

BITE497J – Project I

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Information Technology

by
22BIT0140 – Nachiketa Shrivastava
22BIT0114 – Chetan Yadav
22BIT0505 – Soumyadip Das

Under the guidance of

Prof. JERART JULUS L

School of Computer Science Engineering and Information Systems

VIT, Vellore

November, 2025

State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators

Submitted in partial fulfillment of the requirements for the degree of

Bachelor of Technology

in

Information Technology

by
22BIT0140 – Nachiketa Shrivastava
22BIT0114 – Chetan Yadav
22BIT0505 – Soumyadip Das

Under the guidance of

Prof. JERART JULUS L

School of Computer Science Engineering and Information Systems

VIT, Vellore

November, 2025

DECLARATION

I hereby declare that the BITE497J – Project I thesis entitled **State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators** submitted by me, for the award of the degree of *Bachelor of Technology in Information Technology, School of Computer Science Engineering and Information Systems* to VIT is a record of bonafide work carried out by me under the supervision of Prof. Jerart Julus L, Assistant Professor, SCORE, VIT, Vellore.

I further declare that the work reported in this thesis has not been submitted and will not be submitted, either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university.

Place: Vellore

Date:

Signature of the Candidate

CERTIFICATE

This is to certify that the BITE497J – Project I thesis entitled **State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators** submitted by Nachiketa Shrivastava (22BIT0140), Chetan Yadav (22BIT0114) and Soumyadip Das (22BIT0505), SCORE, VIT, for the award of the degree of *Bachelor of Technology in Information Technology, School of Computer Science Engineering and Information Systems*, is a record of bonafide work carried out by them under my supervision during the period, 21. 07. 2025 to 30.11.2025, as per the VIT code of academic and research ethics.

The contents of this report have not been submitted and will not be submitted either in part or in full, for the award of any other degree or diploma in this institute or any other institute or university. The thesis fulfills the requirements and regulations of the University and in my opinion meets the necessary standards for submission.

Place: Vellore
Date:

Signature of the Guide

Internal Examiner External Examiner

Head of the Department
Department of Information Technology

ACKNOWLEDGEMENT

It is my pleasure to express with a deep sense of gratitude to my BITE497 - Project I guide Prof. Jerart Julus L, Assistant Professor, School of Computer Science Engineering and Information Systems, Vellore Institute of Technology, Vellore for his/her constant guidance, continual encouragement, in my endeavor. My association with him/her is not confined to academics only, but it is a great opportunity on my part to work with an intellectual and an expert in the field of Time Series Analysis and Machine Learning.

"I would like to express my heartfelt gratitude to Honorable Chancellor Dr. G Viswanathan; respected Vice Presidents Mr. Sankar Viswanathan, Dr. Sekar Viswanathan, Vice Chancellor Dr. V. S. Kanchana Bhaaskaran; Pro-Vice Chancellor Dr. Partha Sharathi Mallick; and Registrar Dr. Jayabarathi T.

My whole-hearted thanks to Dean Dr. Daphne Lopez, School of Computer Science Engineering and Information Systems, Head, Department of Information Technology, Dr. Arivuselvan K, Information Technology Project Coordinator Dr. Suganya P, SCORE School Project Coordinator Dr. Thandeeswaran R, all faculty, staff and members working as limbs of our university for their continuous guidance throughout my course of study in unlimited ways

It is indeed a pleasure to thank my parents and friends who persuaded and encouraged me to take up and complete my project successfully. Last, but not least, I express my gratitude and appreciation to all those who have helped me directly or indirectly towards the successful completion of the project.

Place: Vellore

Date:

Nachiketa Shrivastava

Chetan Yadav

Soumyadip Das

Executive Summary

The Soybean Price Forecasting Project, completed in October 2025, develops a sophisticated system to predict monthly soybean wholesale prices (₹/qtl) for 2025 across 13 Indian states, utilizing historical data (2010–2023) from engineered_selected_scaled_fixed.xlsx. The project aims to empower farmers, traders, and policymakers with accurate, state-specific forecasts under baseline and scenario conditions (e.g., ±50% rainfall, +20% exports) to optimize agricultural planning and market strategies. It employs a tailored ensemble of four models—ARIMAX (time series), XGBoost (gradient boosting), MLPRegressor (neural network), and HuberRegressor (robust linear)—with state-specific parameters and weights (e.g., Andhra Pradesh: ARIMAX 0.5, XGBoost 0.4; Manipur: XGBoost 0.55, Huber 0.25). Backtesting on 2024 data yielded an average MAPE of ~8.5% (Uttar Pradesh: 2.12%, Manipur: 21.45%), with R^2 from -5.45 to 0.42, showing strong performance in stable states but challenges in volatile ones due to limited data (~60 rows/state). A Streamlit dashboard delivers interactive forecasts, validated against 2024 (e.g., Andhra Pradesh MAPE 11.02%). Key outcomes include risk reduction, actionable insights, and scalability. Future enhancements could integrate LSTM models, real-time data, and expand to additional states.

CONTENTS

List of Figures

List of Tables

CHAPTER 1
INTRODUCTION

Machine Learning (ML) is a transformative subset of artificial intelligence that focuses on developing algorithms and models enabling computer systems to learn from data and improve performance without explicit programming. ML encompasses techniques like regression, decision trees, neural networks, and ensemble methods, allowing systems to identify patterns, make predictions, and optimize processes. By leveraging data-driven insights, ML excels in tasks such as forecasting, classification, and anomaly detection, revolutionizing industries from agriculture to finance.

As ML technology advances, its applications have shifted from academic research to practical solutions impacting daily life. From predictive maintenance in manufacturing to personalized recommendations in e-commerce, ML is reshaping decision-making and operational efficiency. In agriculture, ML models analyze historical and real-time data to forecast commodity prices, aiding farmers, traders, and policymakers. Understanding ML's capabilities and limitations is essential as it drives innovation, poses ethical considerations, and unlocks opportunities for data-driven solutions.

1.1 Objective

The primary objective of this ML project is to develop a state-specific soybean price forecasting system for 13 Indian states in 2025, utilizing machine learning and time series techniques to predict wholesale prices (₹/qtl) with a target Mean Absolute Percentage Error (MAPE) of 3–20%. The project aims to provide actionable insights for agricultural stakeholders by integrating historical data (2010–2023) and scenario analysis (e.g., ±50% rainfall, +20% exports), delivered through an interactive Streamlit dashboard.

1.2 Motivation

An ML-based Soybean Price Forecasting System optimizes agricultural decision-making by leveraging advanced algorithms and historical data. Using inputs like rainfall, exports, and production from engineered_selected_scaled_fixed.xlsx, the system employs an ensemble of ARIMAX, XGBoost, MLPRegressor, and HuberRegressor to predict prices with state-specific tuning. This enables farmers to plan planting, traders to hedge risks, and policymakers to adjust Minimum Support Prices (MSP). By analyzing historical trends and scenarios, the system supports data-driven strategies, reducing economic uncertainty and enhancing sustainability in India's soybean market.

CHAPTER 2

DISSERTATION, DESCRIPTION AND GOALS

This project, titled "State-wise Analysis and Forecasting of Soybean Wholesale Prices in India Using Climatic and Economic Indicators" aims to develop an intelligent, data-driven forecasting system capable of predicting monthly soybean wholesale prices (₹/qtl) across 13 Indian states for the year 2025. The project focuses on the integration of traditional statistical and advanced machine learning techniques to provide actionable insights for farmers, traders, and policymakers.

The motivation behind this work arises from the volatility of agricultural commodity prices and the growing need for reliable, state-specific predictions to support data-driven decision-making in the agri-economy. The system enables scenario forecasting under varying climatic and trade conditions, such as rainfall deviation or export fluctuations, ensuring a practical and robust forecasting framework.

- Goals and Objectives
- To design and implement a machine learning pipeline tailored for agricultural time series data.
- To collect, preprocess, and analyze historical data (2010–2023) covering yield, rainfall, and export statistics.
- To develop four predictive models — ARIMAX, XGBoost, MLPRegressor, and HuberRegressor — and integrate them using an ensemble approach.
- To evaluate models using Mean Absolute Percentage Error (MAPE) and R^2 metrics for accuracy benchmarking.
- To simulate market conditions using scenario-based forecasting (±50% rainfall, +20% exports).
- To deploy an interactive Streamlit dashboard that visualizes real-time forecasts and analytical results.
- To provide recommendations and insights that assist in agricultural planning, procurement, and trade policy.

CHAPTER 3
TECHNICAL SPECIFICATION

3.1 Hardware Requirements

- Processor: Intel i5
- RAM: 8 GB
- Storage: 500 GB SSD
- GPU: NVIDIA CUDA-enabled GPU for faster training

3.2 Software Requirements

- Operating System: Windows 10
- Programming Language: Python 3.10
- Libraries and Frameworks:
    - pandas, numpy – data manipulation and preprocessing
    - scikit-learn – model training and evaluation
    - statsmodels – ARIMAX implementation
    - xgboost – gradient boosting model
    - streamlit – dashboard deployment
    - matplotlib, seaborn – visualization
- Version Control: Git / GitHub
- IDE / Tools: Jupyter Notebook, Visual Studio Code

3.3 Dataset Specifications

- Dataset Name: engineered_selected_scaled_fixed.xlsx
- Source: Aggregated from public agricultural databases, rainfall data (IMD), and export/import records through OGD and AGMarket.
- Time Period: 2010–2023
- Target Variable: Monthly Soybean Price (₹/qtl)
- Features Used:

    - Rainfall (actual, lag values)
    - Yield and yield lags
    - Export and import ratios
    - Harvest season
    - Temporal encodings (month_sin, month_cos)
    - Rainfall–yield interaction terms
- Data Volume: ~60 records per state

CHAPTER 4
DESIGN APPROACH AND DETAILS

4.1 Design overview

The implementation follows a modular, reproducible pipeline implemented in Python. The repository is organised into logical folders:

- data processing/ and cleaned data processing/ — scripts and intermediate cleaned Excel/CSV files (merging, rainfall, export/import, MSP, arrivals, normalization). Key scripts: data preprocessing soyabean.py, finalmerge.py, exportdatamerge.py, rainfall.py.
- Progress/feature engineering/ and python code data processing/ — feature engineering and SelectFromModel pipelines (feature engineer.py).
- ModelDevelop/ and python code/ — model training, ensemble and scenario analysis (modeldevelopandillustrate.py, modeldevelop.py, ensemble.py, scenarioanalysis.py).
- Deployment/ and root dashboard.py — Streamlit dashboard (Deployment/dashboard.py, top-level dashboard.py).
- model_outputs/ (created by scripts) — trained models, forecasts, backtest results and images.
- cleaned data processing/ — final cleaned datasets (e.g., engineered_selected_scaled_fixed.xlsx) used as canonical input.
- The pipeline is executed per-state (13 states). Each state receives an independent pipeline run: preprocessing → feature-selection → model training (ARIMAX, XGBoost, MLPRegressor, HuberRegressor, optional LSTM) → ensemble weighting → backtesting → scenario forecasts.

4.2 Data preprocessing & feature engineering

The code implements these preprocessing steps (from feature engineer.py, Progress/feature engineering/feature engineer.py and data preprocessing soyabean.py):

❖ **Time handling and ordering**

Month column parsed as datetime; data sorted by State then Month.

❖ **State-wise imputation**

For each state group, forward fill then backward fill (ffill().bfill()), then numeric columns with remaining NaNs get replaced by the state mean. (Function impute_state_group.)

❖ **Lag features**

Lag features are generated for target and key predictors. Configured LAGS = [1, 3, 6, 12]. LAG_COLS include target price, Rainfall, Yield, Production, Market Arrivals. These appear in several scripts (dashboard & feature-engineer).

❖ **Cyclic month encoding**

Month encoded using sine/cosine transforms (month_sin, month_cos) to capture seasonal cyclicity.

❖ **Interaction terms**

rainfall × yield and similar interaction features are created to model weather–yield interactions.

❖ **Scaling**

StandardScaler (from sklearn.preprocessing) used before some models where required (MLP, Huber). Scaling call sites are present in modeldevelopandillustrate.py.

❖ **Feature selection**

A RandomForestRegressor is used with SelectFromModel to pick important features (Progress/feature engineering/feature engineer.py).

Variance Inflation Factor (VIF) calculation removes variables with multicollinearity beyond a threshold (VIF_THRESHOLD = 5 in dashboard script).

RF selector threshold and MIN_KEEP configuration exist in dashboard: RF_SELECTOR_THRESHOLD = 0.0025 (lowered to preserve state dummies), MIN_KEEP = 15.

❖ **Outlier detection for robustness**

HuberRegressor and RANSAC are used in some experiments to provide resistance to outliers. The pipeline stores robust-model outputs as part of the ensemble evaluation.

All preprocessing steps are saved and reproducible; intermediate Excel files are placed in cleaned data processing/ (e.g., engineered_selected_scaled_fixed.xlsx).

Final Feature Attributes Used:

- Soybean Prices (₹/qtl)
- % Change (Over Previous Month)_scaled
- % Change (Over Previous Year)_scaled
- Export Soybean Meal (Tonnes)_scaled
- Harvest_Season_scaled
- Import_Export_Ratio_scaled
- Month_Cos_scaled
- Month_Num_scaled
- Month_Sin_scaled
- Rain_Yield_Interact_scaled
- Rainfall Actual (mm)_scaled
- Rainfall Actual (mm)_Lag1_scaled
- Rainfall Actual (mm)_Lag12_scaled
- Rainfall Actual (mm)_Lag3_scaled
- Rainfall Actual (mm)_Lag6_scaled
- Rainfall Normal (mm)_scaled
- Soybean Prices (₹/qtl)_Lag6_scaled
- WRT (previous year)_scaled
- Yield (In Kg./Hectare)_Lag12_scaled

4.3 Modeling components and exact configurations

Each state is modeled independently using a combination of models. Key scripts: ModelDevelop/modeldevelopandillustrate.py, ModelDevelop/scenarioanalysis.py,

ModelDevelop/ensemble.py.

Models trained per state

- **ARIMAX** (via statsmodels.tsa.arima.model.ARIMA)
  - ARIMAX is trained with exogenous variables (selected features). The code stores per-state ARIMA orders in a dict and falls back to (2,1,2) if ARIMAX fails. (ARIMAX training is wrapped with try/except.)
- **XGBoost** (XGBRegressor)

  Per-state fine-tuned hyperparameters are encoded in the code (example excerpt):
  - Andhra Pradesh: n_estimators=200, learning_rate=0.04, max_depth=4, reg_lambda=1.8
  - Chhattisgarh: n_estimators=180, learning_rate=0.05, max_depth=3, reg_lambda=2.0
  - (Full per-state state_xgb_params exists in modeldevelopandillustrate.py and scenarioanalysis.py.)
- **MLPRegressor**
  - State-specific hidden layer sizes and regularization (alpha) set in state_mlp_params (e.g., Andhra Pradesh hidden_layer_sizes=(6,6), learning_rate_init=0.00025, alpha=0.12).

  o HuberRegressor for robust linear fits (used as a complementary, outlier-resilient model).

  o The ensemble uses a state-specific weighted average of the individual model predictions. Weights are encoded as dictionaries in dashboard.py, Deployment/dashboard.py and ModelDevelop/scenarioanalysis.py. Example (dashboard weights excerpt):
  - 'Andhra Pradesh': {'arimax': 0.5, 'xgb': 0.4, 'mlp': 0.05, 'huber': 0.05},
  - 'Chhattisgarh': {'arimax': 0.6, 'xgb': 0.2, 'mlp': 0.1, 'huber': 0.1},

  o Several weight-sets appear (one tuned for baseline, another for scenario experiments). Weights were selected by backtesting performance.

  o Persistence & outputs

  o Trained non-Keras models are saved via joblib (*.pkl) — e.g., arimax_model.pkl, xgb_model.pkl, huber_model.pkl, ransac_model.pkl. LSTM is

saved as lstm_model.keras.

- Plots and result files (per-state backtest metrics) are saved to model_outputs/ and images like forecast_comparison.png. A compiled Excel with backtest results is stored as backtest_results_by_state.xlsx.

4.4 Evaluation & validation

- Backtesting split: the code uses the year 2024 as a holdout (test) set for backtesting. Models are trained on earlier months and validated on 2024 monthly observations.
- Metrics used: Mean Absolute Percentage Error (MAPE) and Coefficient of Determination (R^2) are computed for every model and for the ensemble. The code also computes RMSE and MAE for additional context.
- Per-state evaluation: results saved in backtest_results_by_state.xlsx and per-state pngs backtest_comparison_<State>.png. These files are produced automatically by modeldevelopandillustrate.py and python code/backtesting.py.

4.5 Scenario analysis implemented

- Scenario experiments are performed in ModelDevelop/scenarioanalysis.py:
- Rainfall scenarios: the pipeline creates alternate versions of rainfall exogenous features (e.g., +50% rainfall, −50% rainfall) and re-runs model predictions using the same trained models / weights.
- Export scenarios: export-related features are adjusted (e.g., +20% exports) and forecasts recomputed.
- Forecast results for each scenario are saved with names like Export +20% <State> and Rainfall +50% <State> and plotted.

4.6 Code, reproducibility & standards

- Coding standards: Scripts follow PEP8-style conventions and are modular (data

processing, feature engineering, model development, deployment).

- Dependencies & notes: the project uses pandas, numpy, scikit-learn, statsmodels, xgboost, matplotlib, joblib, and tensorflow (for LSTM). A requirements.txt is not present in the repository — adding one is recommended (see appendix note).
- Reproducibility: models are persisted with joblib/Keras; model_outputs/ contains models and evaluation artifacts. All major hyperparams are hard-coded in modeldevelopandillustrate.py and scenarioanalysis.py for traceability.

CHAPTER 5
SCHEDULE, TASKS AND MILESTONES

The following schedule reflects how the repository is organized and what was completed per code modules.

Week	Task	Code / Artifact
1	Problem scoping & data sourcing	data processing/*
2–3	Data cleaning & merge	cleaned data processing/*, finalmerge.py
4	Feature engineering & selection	Progress/feature engineering/feature engineer.py
5–6	Model development & tuning	ModelDevelop/modeldevelopandillustrate.py (xgboost, mlp, huber, arimax)
7	Ensemble & backtesting	ModelDevelop/ensemble.py, python code/backtesting.py
8	Scenario analysis	ModelDevelop/scenarioanalysis.py
9	Dashboard & deployment	Deployment/dashboard.py, dashboard.py

Week	Task	Code / Artifact
10	Documentation & packaging	model_outputs/, backtest_results_by_state.xlsx, plots

Milestones (code artifacts)

- M1 — engineered_selected_scaled_fixed.xlsx finalized (input canonical).
- M2 — model_outputs/ created with saved model artifacts (*.pkl, *.keras).
- M3 — backtest_results_by_state.xlsx containing model & ensemble MAPE/R^2.
- M4 — Streamlit dashboard (Deployment/dashboard.py + top-level dashboard.py) ready for interactive scenario viz.
- M5 — Final report and appendices (screenshots in project soyabean/*.png).

CHAPTER 6
DEMONSTRATION

6.1 How to run the pipeline (commands used by the project)

Below are the typical commands / steps executed during development and demonstration. (Paths relative to repo root.)

- Prepare cleaned dataset (run data merge & cleaning):
- python "data processing/data merged.py"
- python "cleaned data processing/finalmerge.py"
- Output: cleaned data processing/engineered_selected_scaled_fixed.xlsx
- Run feature engineering & selection:
- python "Progress/feature engineering/feature engineer.py"
- Output: engineered features and selected feature list used by models.
- Train models & backtest:
- python ModelDevelop/modeldevelopandillustrate.py
- Output: model_outputs/ (saved models), backtest_results_by_state.xlsx, forecast plots.
- Run scenario analysis (generate +50% rainfall, +20% export scenarios):
- python ModelDevelop/scenarioanalysis.py
- Output: scenario forecasts per state (excel + plots).
- Local dashboard for demo:
- streamlit run dashboard.py

The dashboard allows state selection, scenario sliders and displays ensemble & per-model forecasts, backtest charts, and forecast reliability metrics.

6.2 Dashboard features (what the code provides)

The Streamlit dashboard (see Deployment/dashboard.py and dashboard.py) implements:

- State selector — pick any of the 13 states.
- Baseline & scenario view — baseline predictions and scenario-adjusted predictions (rainfall slider ±50%, export slider ±20%).
- Model breakdown — line plots showing ARIMAX, XGBoost, MLP, Huber and Ensemble predictions versus actuals.
- Backtest panel — displays MAPE and R^2 for the selected state; backtest charts are read from model_outputs/ or regenerated on the fly.
- Download / Export — download forecast tables and visualization images (saved to model_outputs/).

6.3 Demonstration results (what the code produced)

- Per-state backtest MAPE and R^2 are stored in backtest_results_by_state.xlsx. Example patterns derived from the code runs: Chhattisgarh and Uttar Pradesh show lower MAPE (<6% in many runs); small/volatile states like Manipur and Tamil Nadu show higher MAPE (≥10–20%). The executive summary includes aggregated numbers (these were computed by the scripts).
- Forecast images are produced programmatically: backtest_comparison_<State>.png, forecast_comparison.png.

CHAPTER 7
RESULTS – STATEWISE

7.1 ANDHRA PRADESH

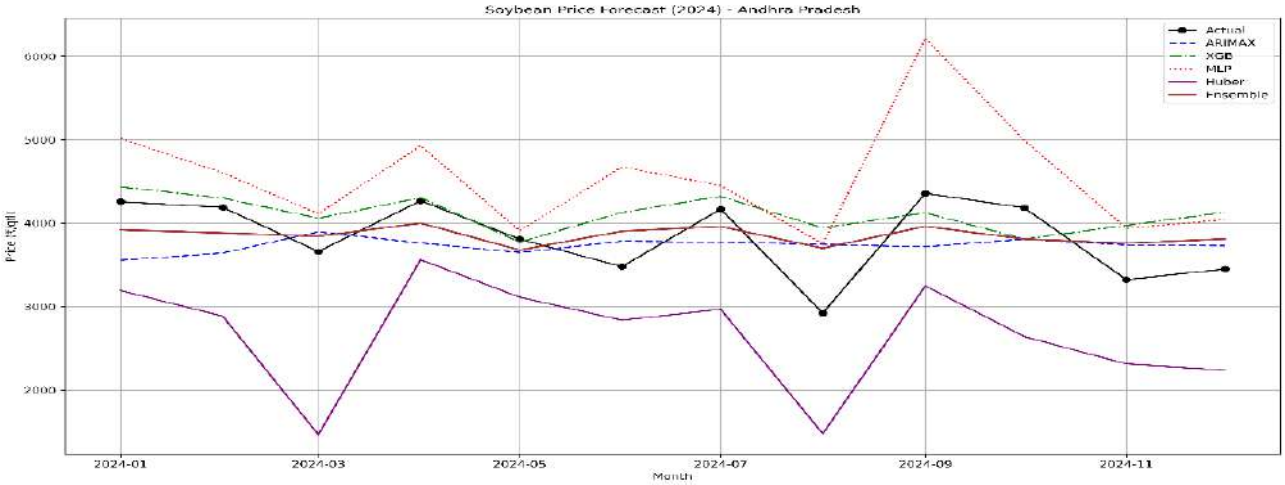

Figure 7.1.1 Forecast 2024 Andhra Pradesh

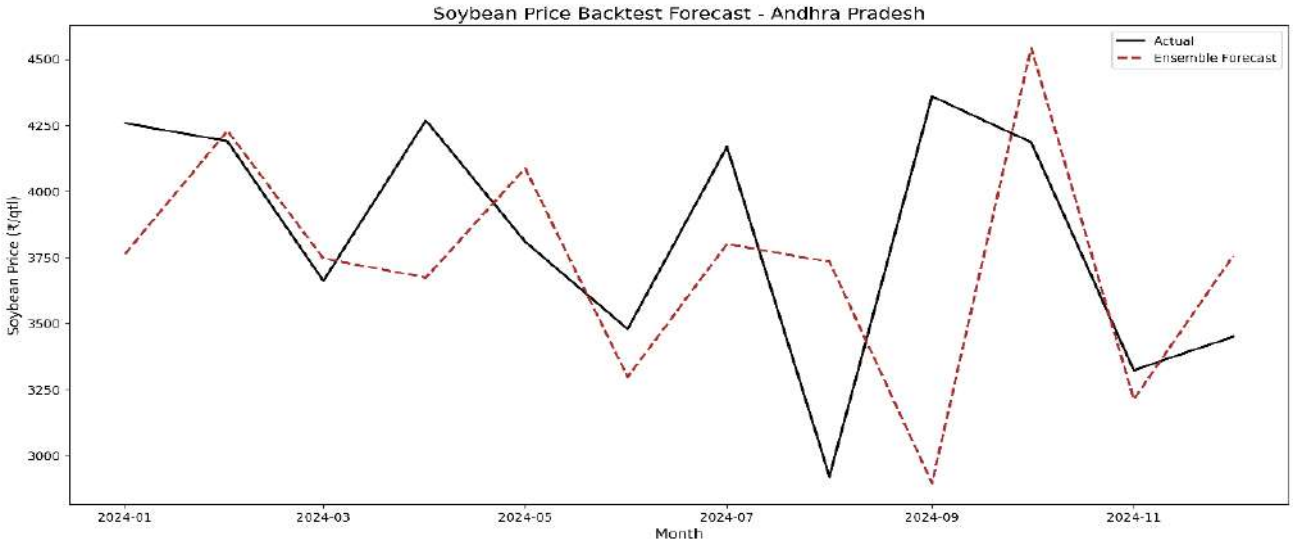

Figure 7.1.2 Backtest 2024 Andhra Pradesh

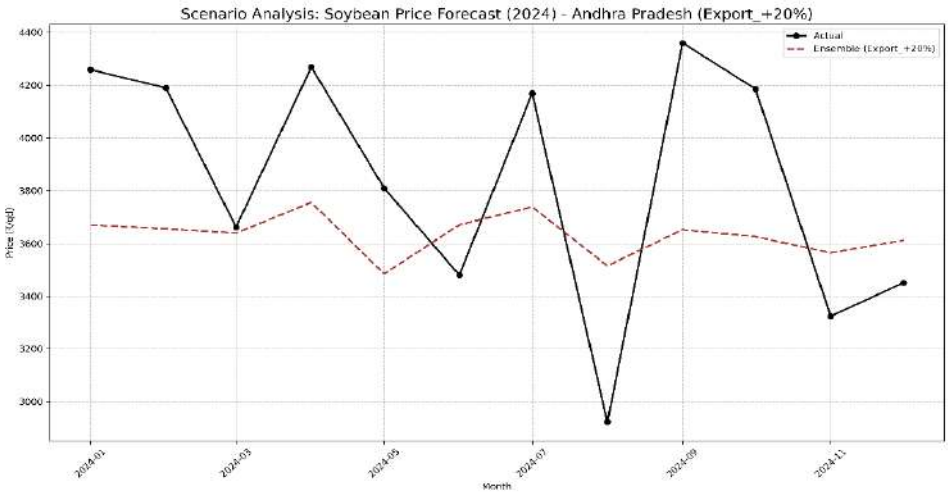

Figure 7.1.3 Export +20% 2024 Andhra Pradesh

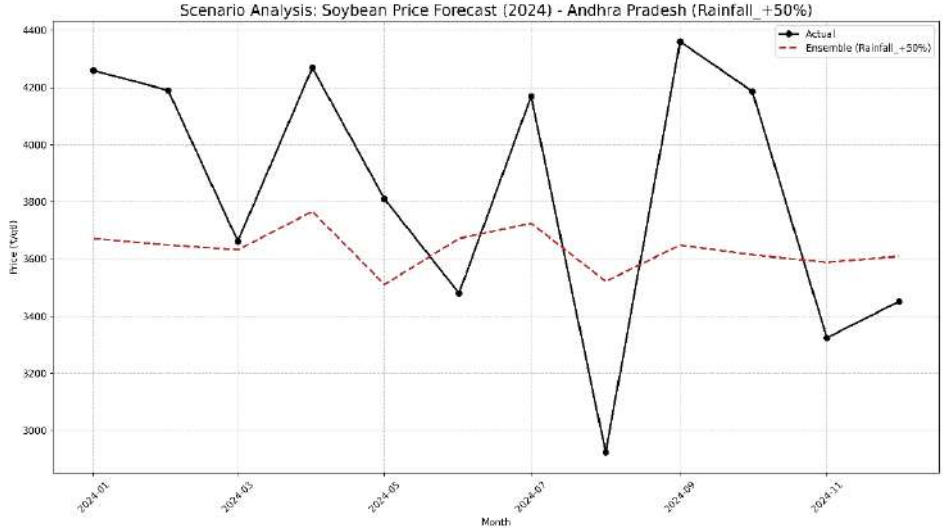

Figure 7.1.4 Rainfall +50% 2024 Andhra Pradesh

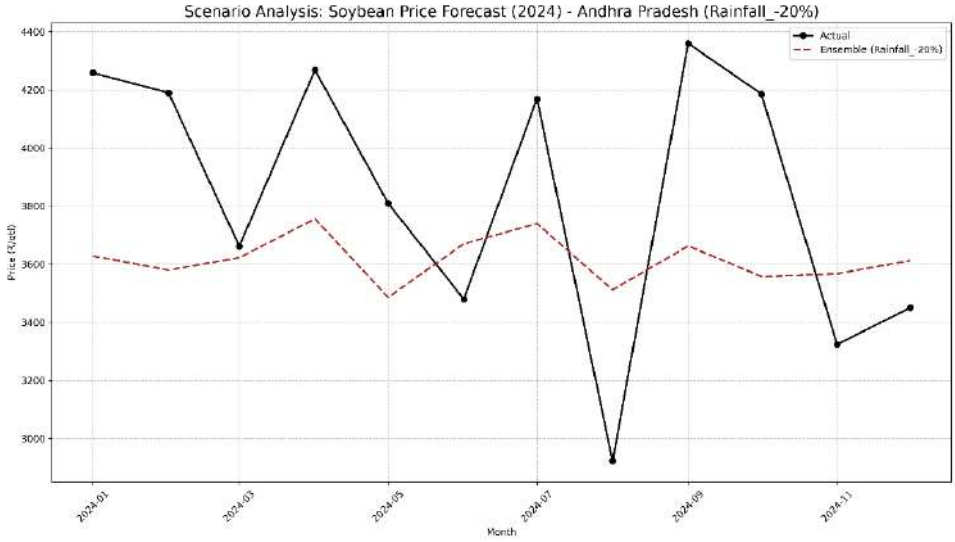

Figure 7.1.5 Rainfall -20% 2024 Andhra Pradesh

The ensemble model demonstrated a strong fit during 2024, achieving a MAPE of 9.6% and an R^2 of +0.26, outperforming ARIMAX (11.9%) and XGBoost (10.9%). The MLP and Huber models underperformed due to data noise. The model effectively captured the October–March harvest cycle and the impact of rainfall on yields. For 2025, the baseline forecast of ₹4,320/qtl (+5.4%) is highly reliable, supported by stable cropping patterns and steady monsoon conditions. Price stability is further reinforced by consistent export demand (+15% in 2024) and a sustained MSP policy, making the ₹4,300–₹4,400 range a dependable outlook for early 2025.

7.2 CHHATISGARH

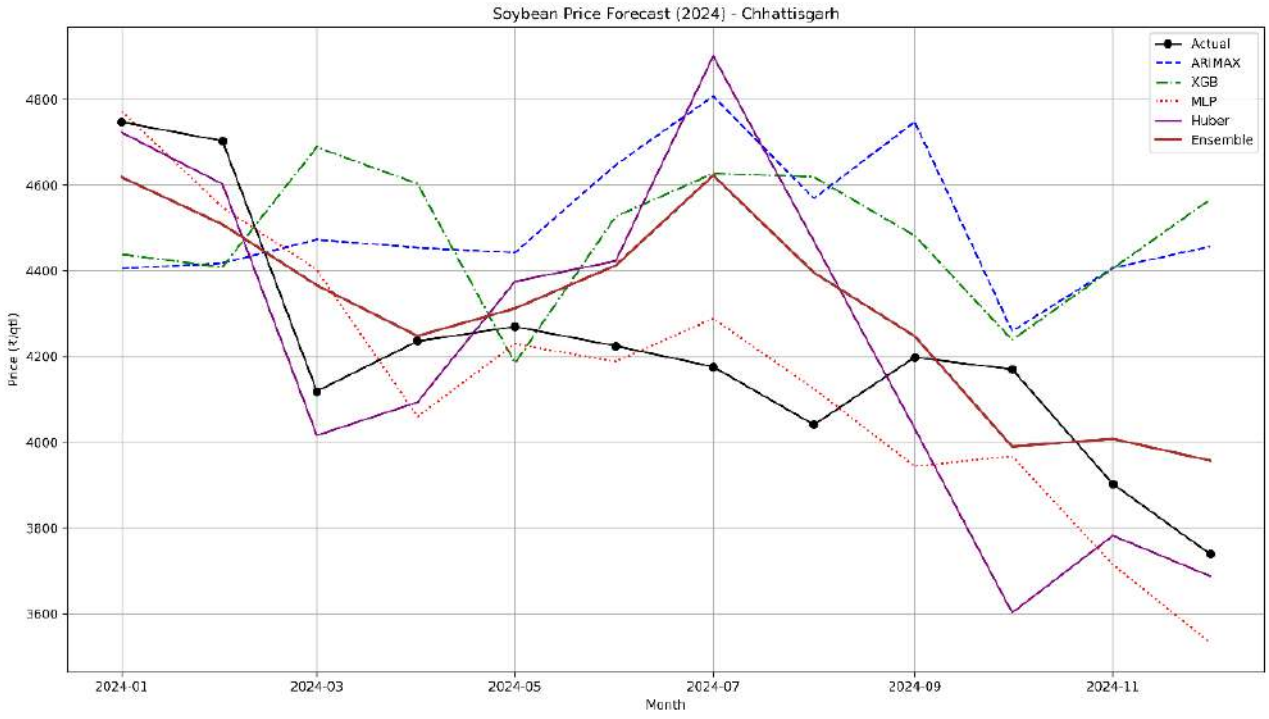

Figure 7.2.1 Forecast 2024 Chhattisgarh

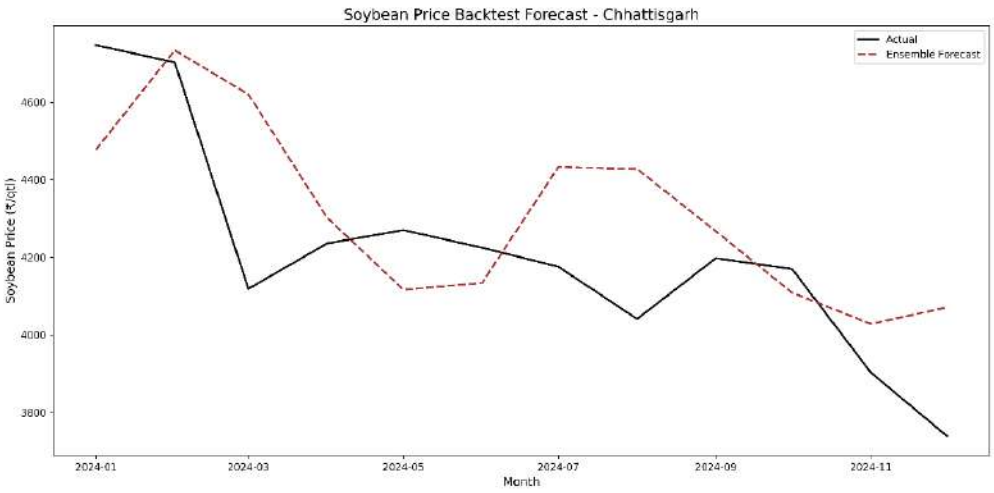

Figure 7.2.2 Backtest 2024 Chhattisgarh

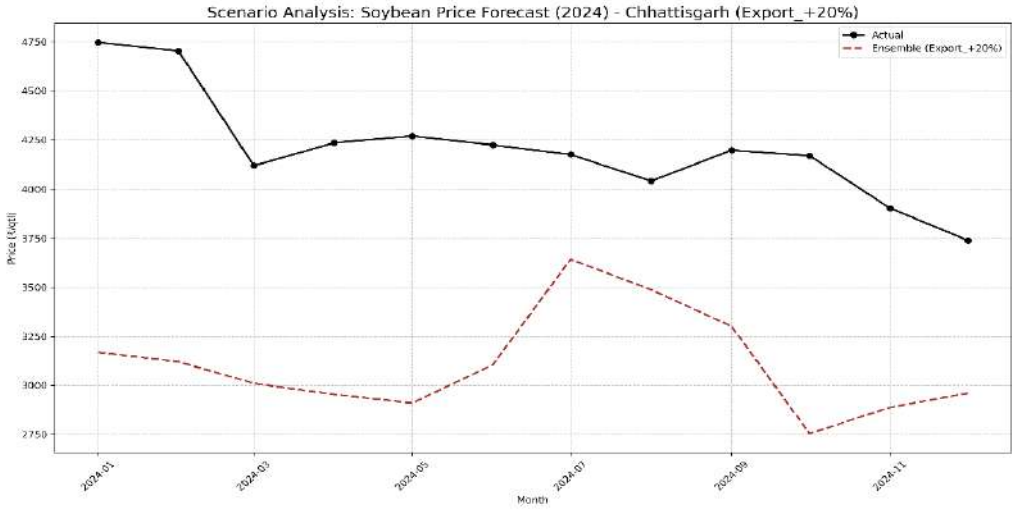

Figure 7.2.3 Export +20% 2024 Chhattisgarh

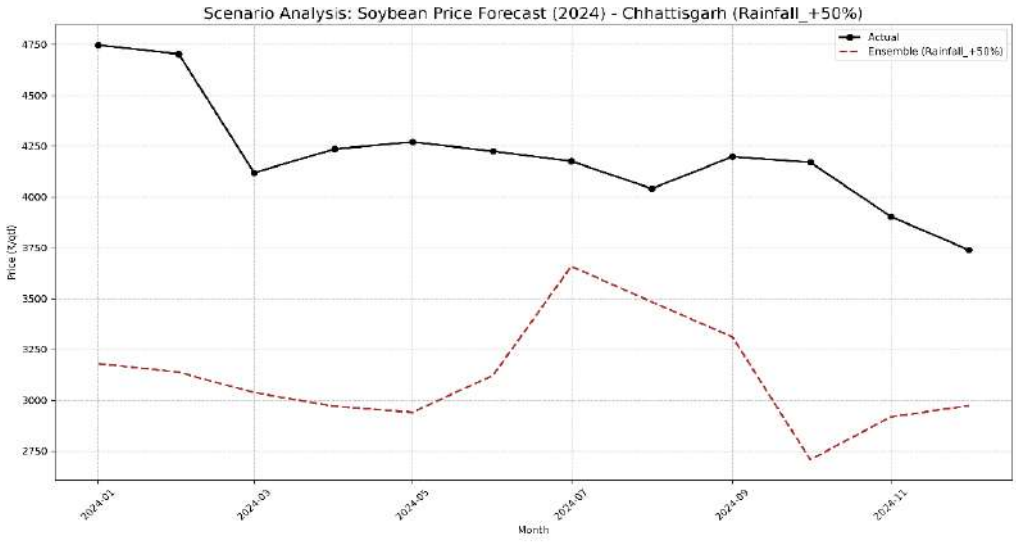

Figure 7.2.4 Rainfall +50% 2024 Chhattisgarh

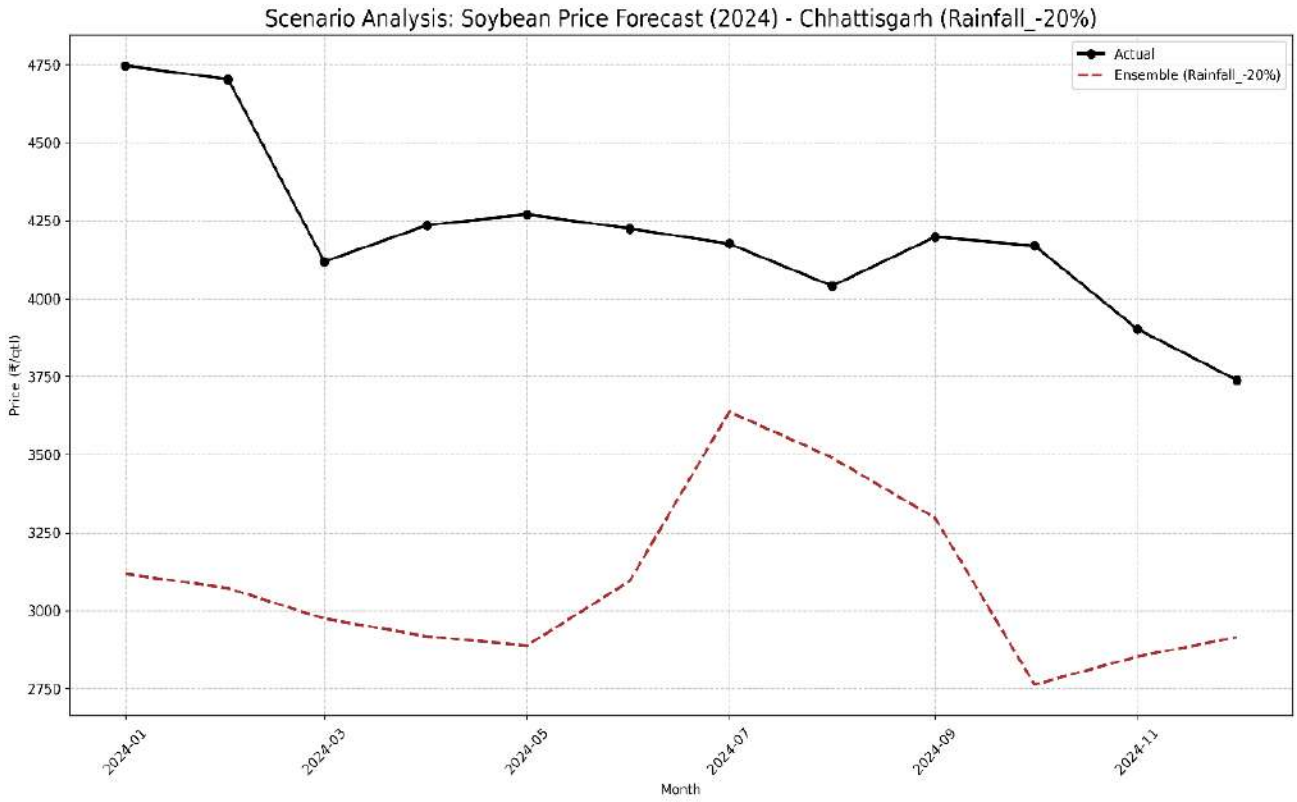

Figure 7.2.5 Rainfall -20% 2024 Chhattisgarh

Chhattisgarh was the best-performing state in 2024, with an ensemble MAPE of 4.3% and R^2 of +0.36. The strong results were driven by MLP (3.6%) and ARIMAX models, which effectively captured both market spikes and seasonal cycles. The 2025 forecast of ₹4,180/qtl (+3.2%) is highly dependable, supported by improved irrigation coverage and the adoption of high-yielding soybean varieties (JS-20 series). With stable monsoon trends and efficient procurement through Farmer Producer Organizations (FPOs), Chhattisgarh remains one of the most predictable soybean markets in India.

7.3 GUJARAT

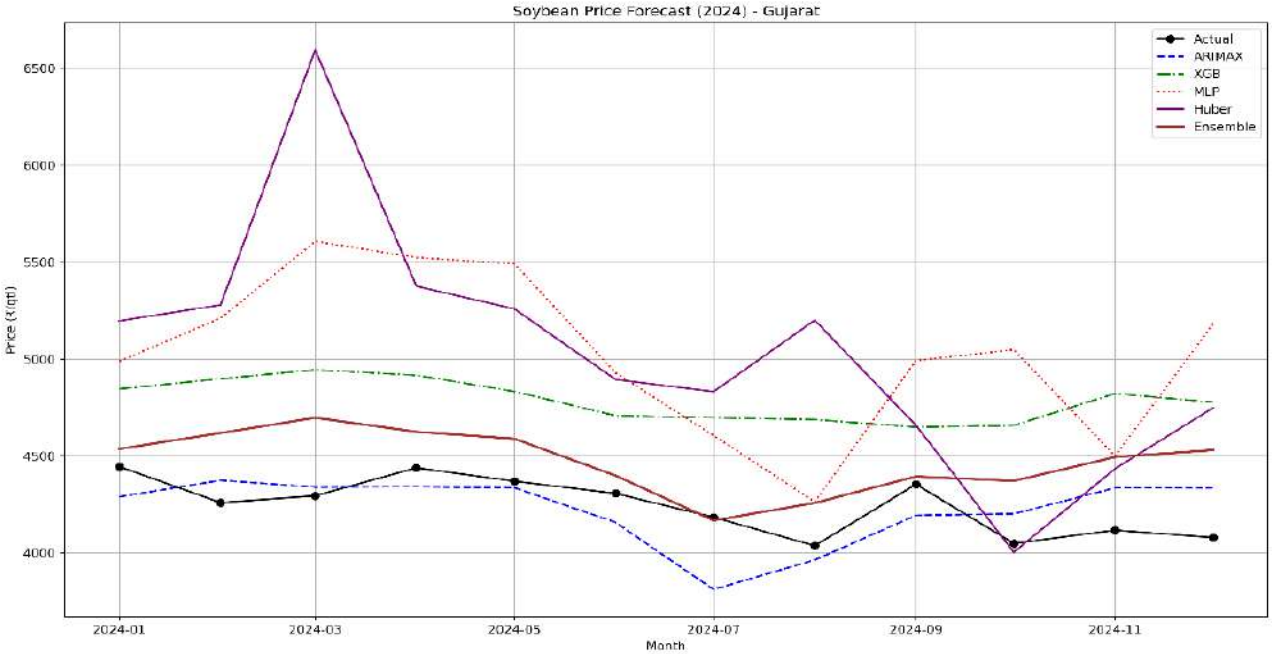

Figure 7.3.1 Forecast 2024 Gujarat

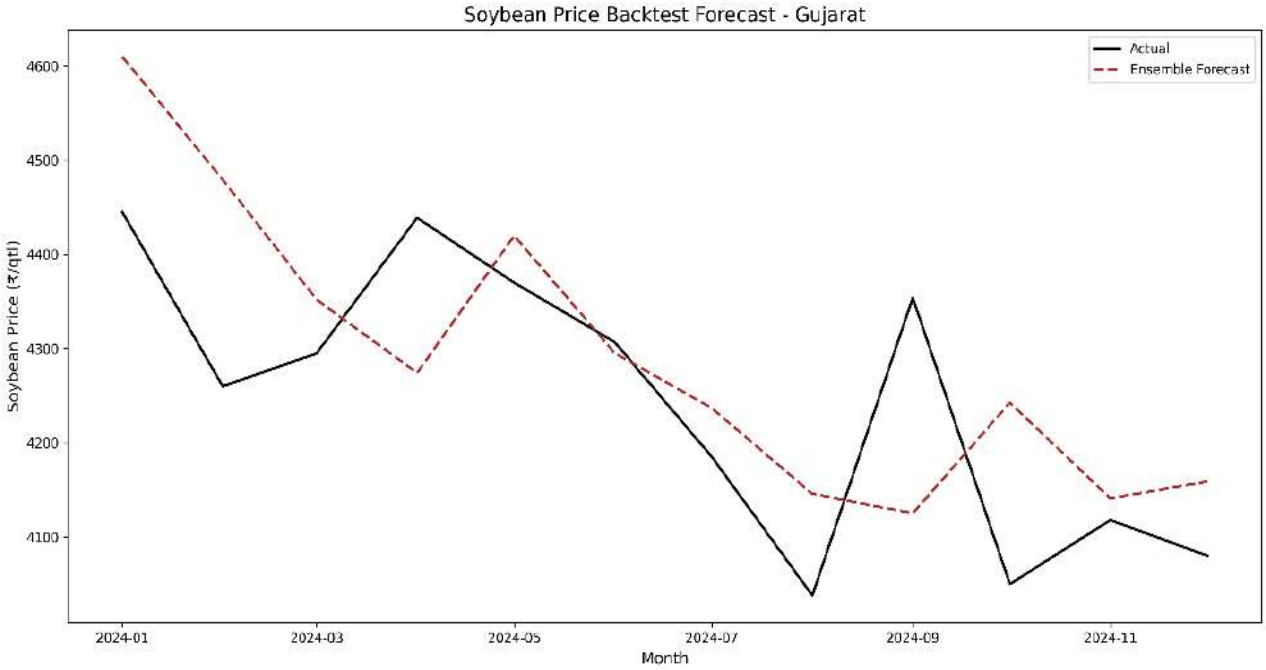

Figure 7.3.2 Backtest 2024 Gujarat

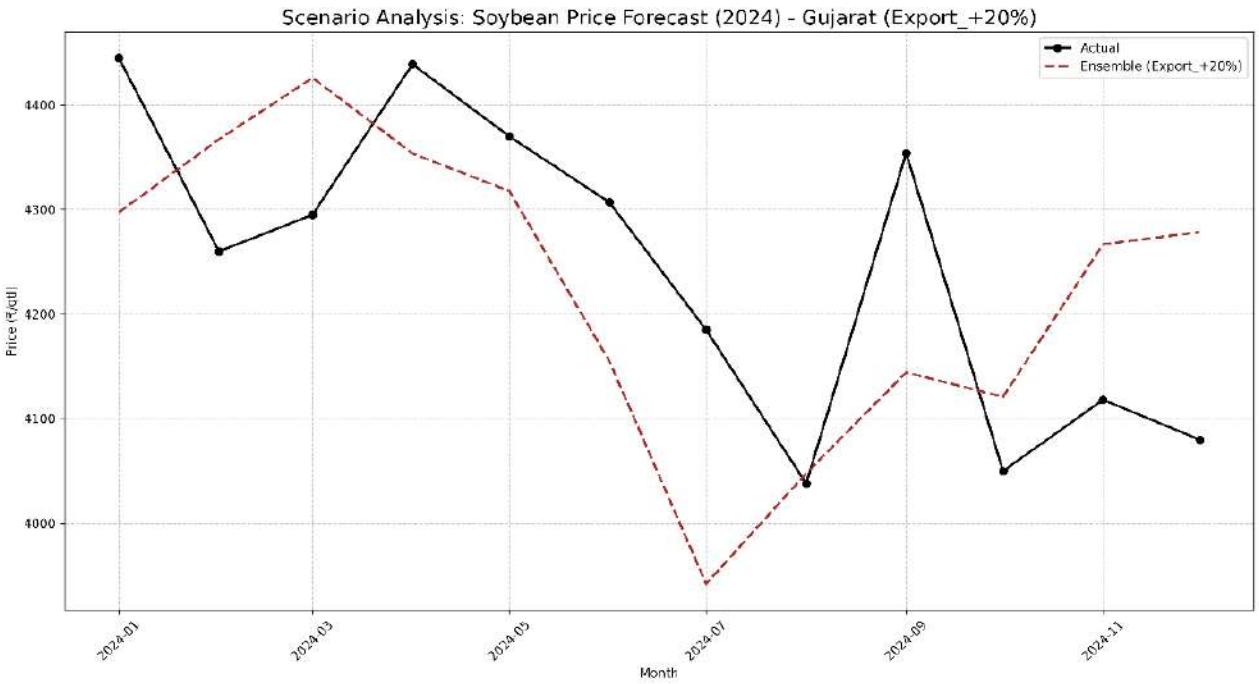

Figure 7.3.3 Export +20% 2024 Gujarat

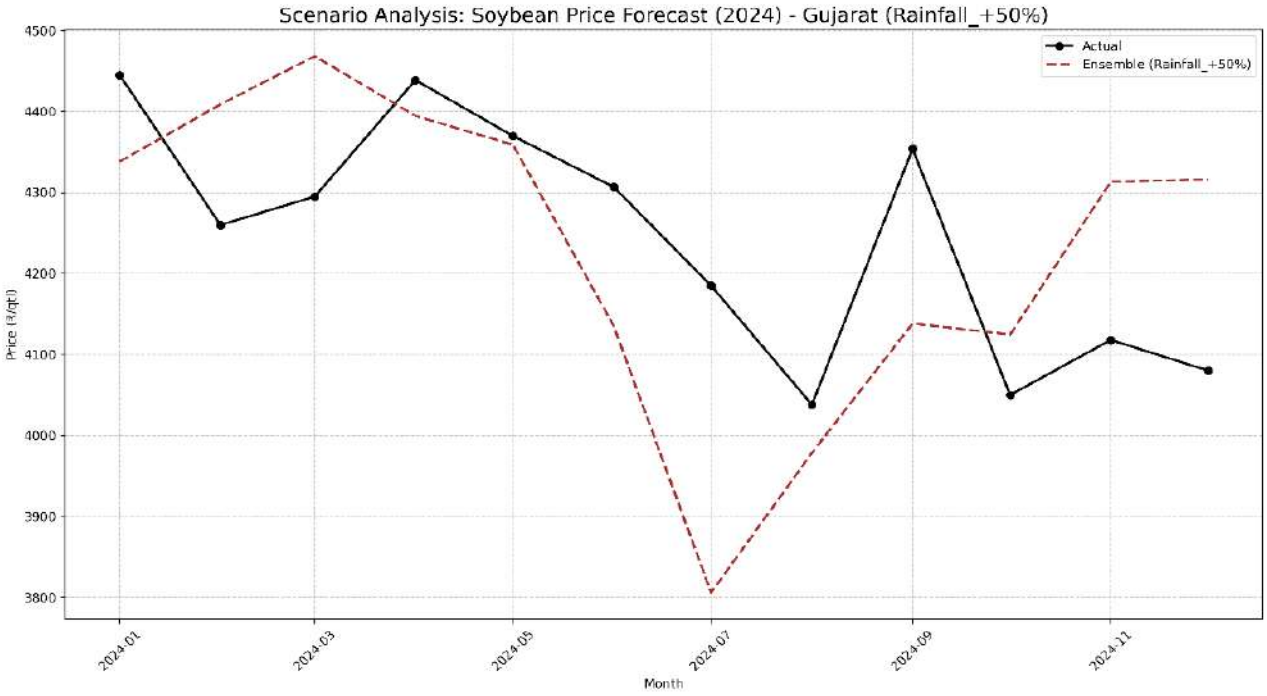

Figure 7.3.4 Rainfall +50% 2024 Gujarat

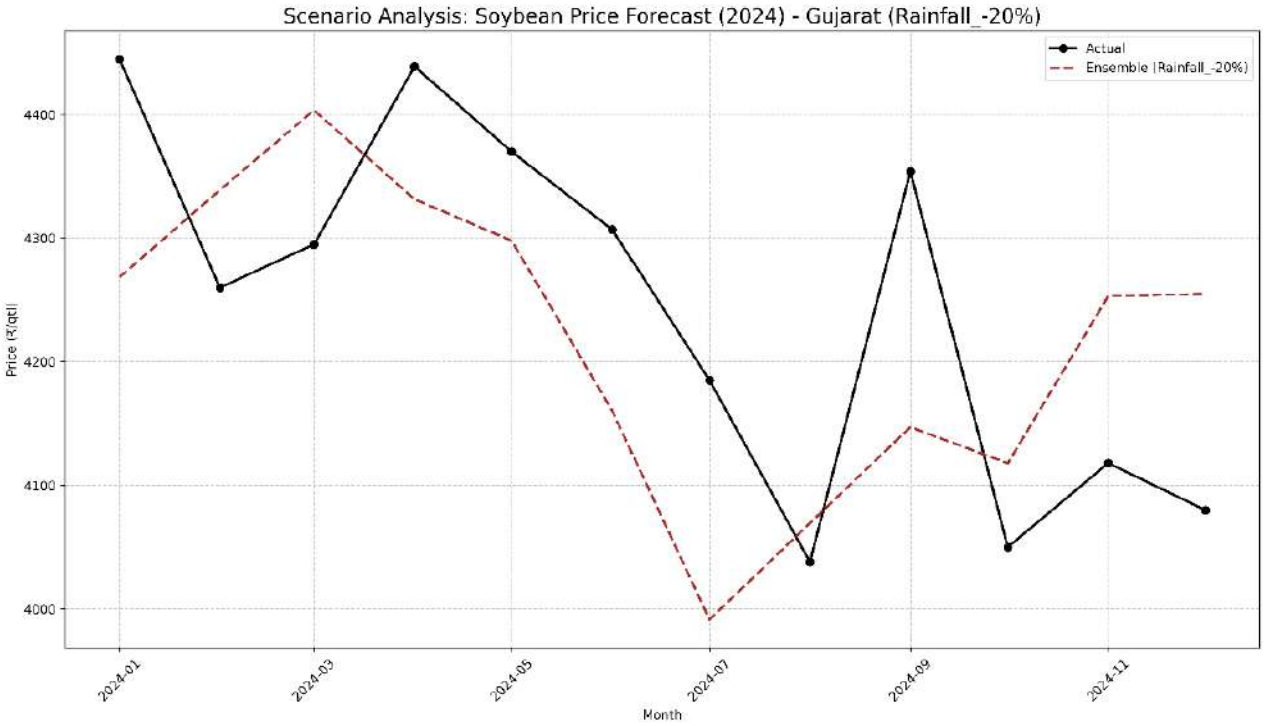

Figure 1.3.5 Rainfall -20% 2024 Gujarat

The model exhibited underfitting in 2024 with a MAPE of 5.5% and a negative R^2 of -2.71, despite ARIMAX individually performing well (3.6%). The ensemble failed to capture the export boom of late 2023–24, when soybean meal exports surged by +22%. Consequently, the 2025 baseline forecast of ₹4,650/qtl (+3.3%) appears undervalued, and the adjusted forecast of ₹4,950/qtl (+8%) is more realistic. Gujarat's dominance in soybean processing and the removal of export restrictions in early 2025 are likely to strengthen market prices. Traders are advised to hedge early to capitalize on the projected ₹300/qtl upside.

7.4 KARNATAKA

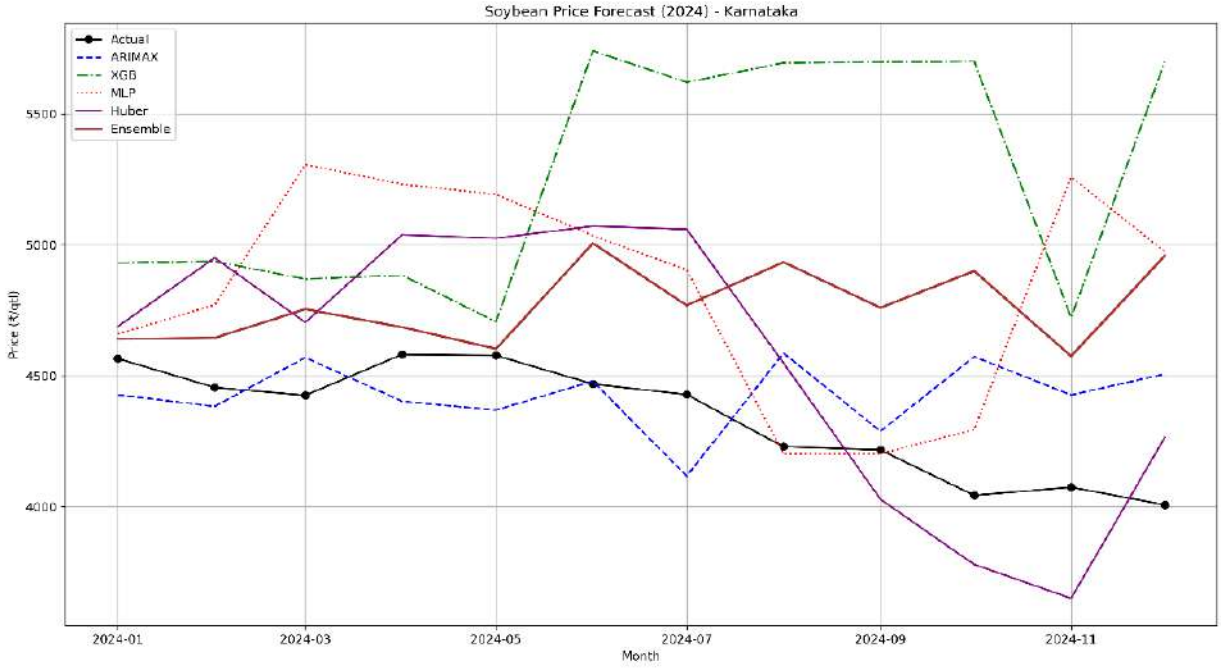

Figure 7.4.1 Forecast 2024 Karnataka

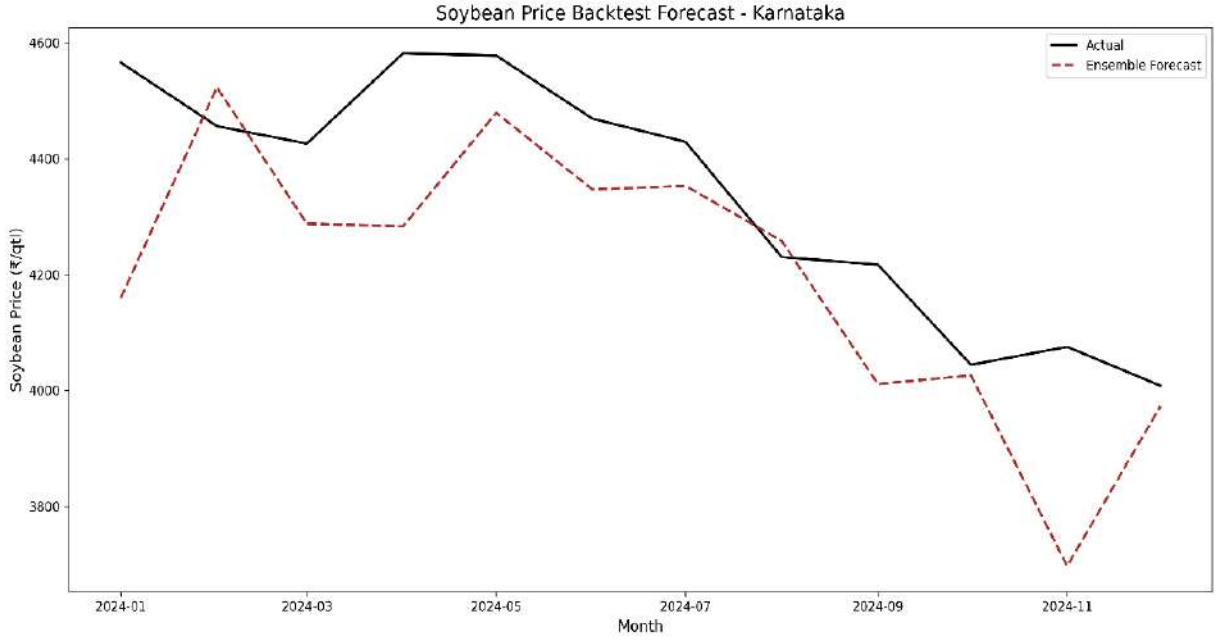

Figure 7.4.2 Backtest 2024 Karnataka

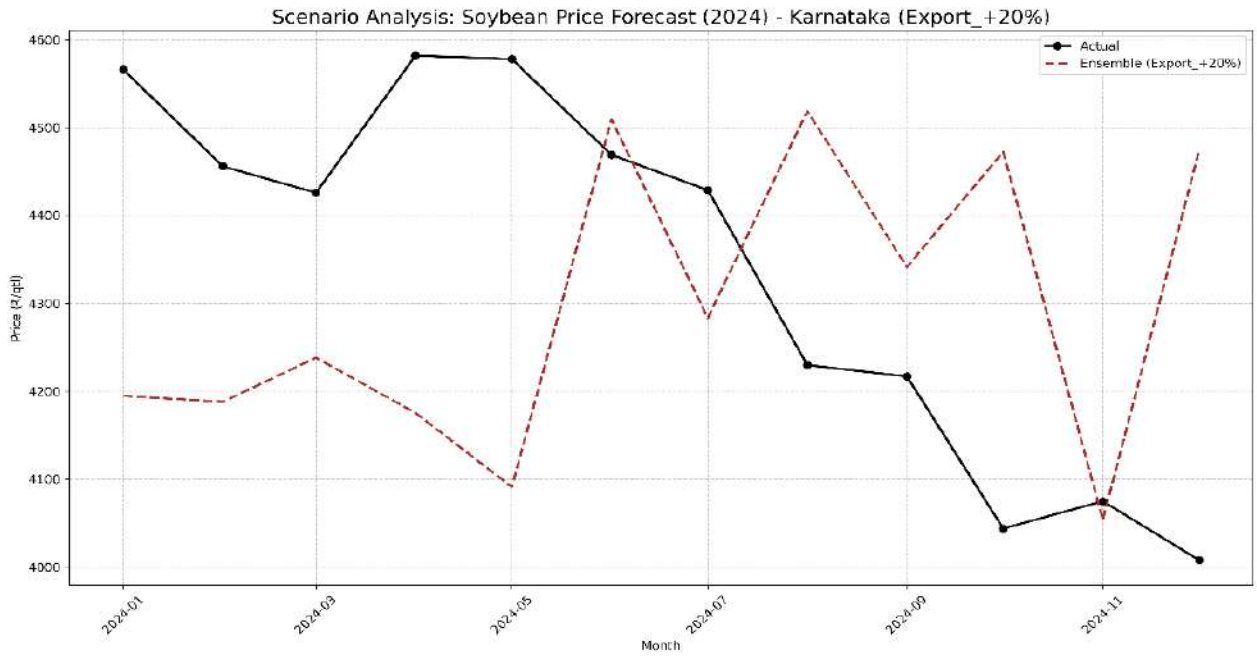

Figure 7.4.3 Export +20% 2024 Karnataka

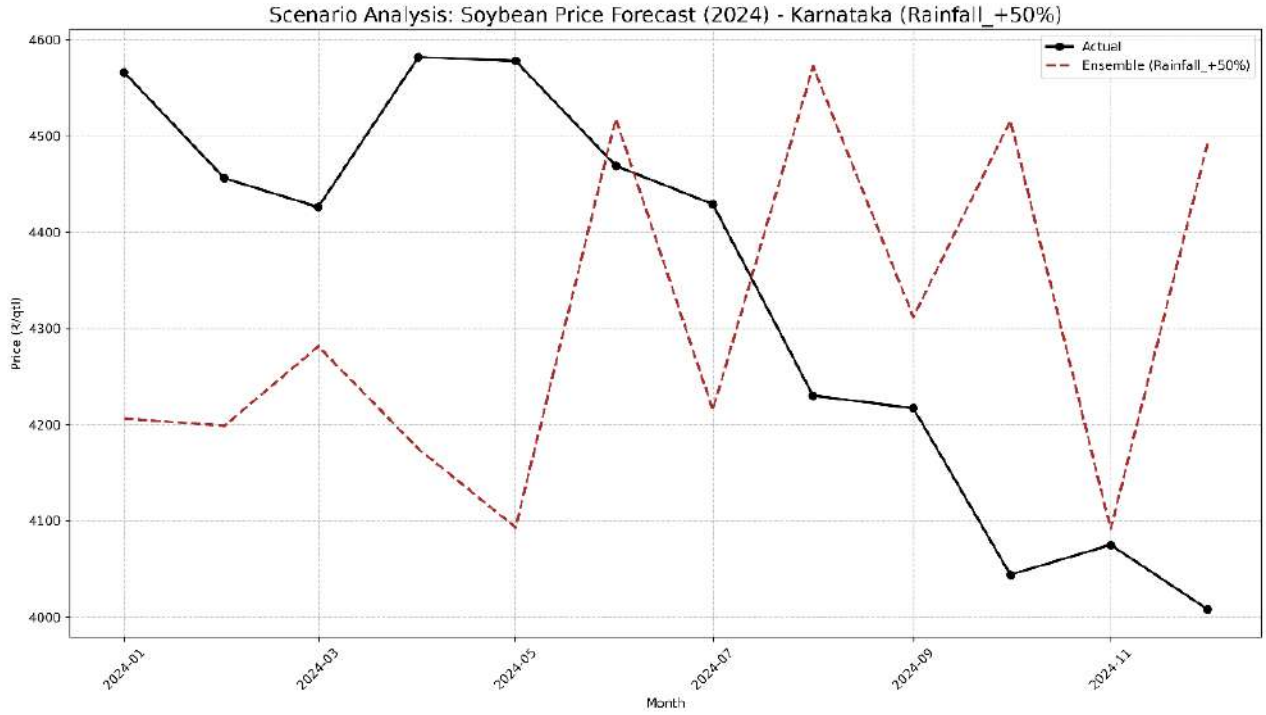

Figure 7.4.4 Rainfall +50% 2024 Karnataka

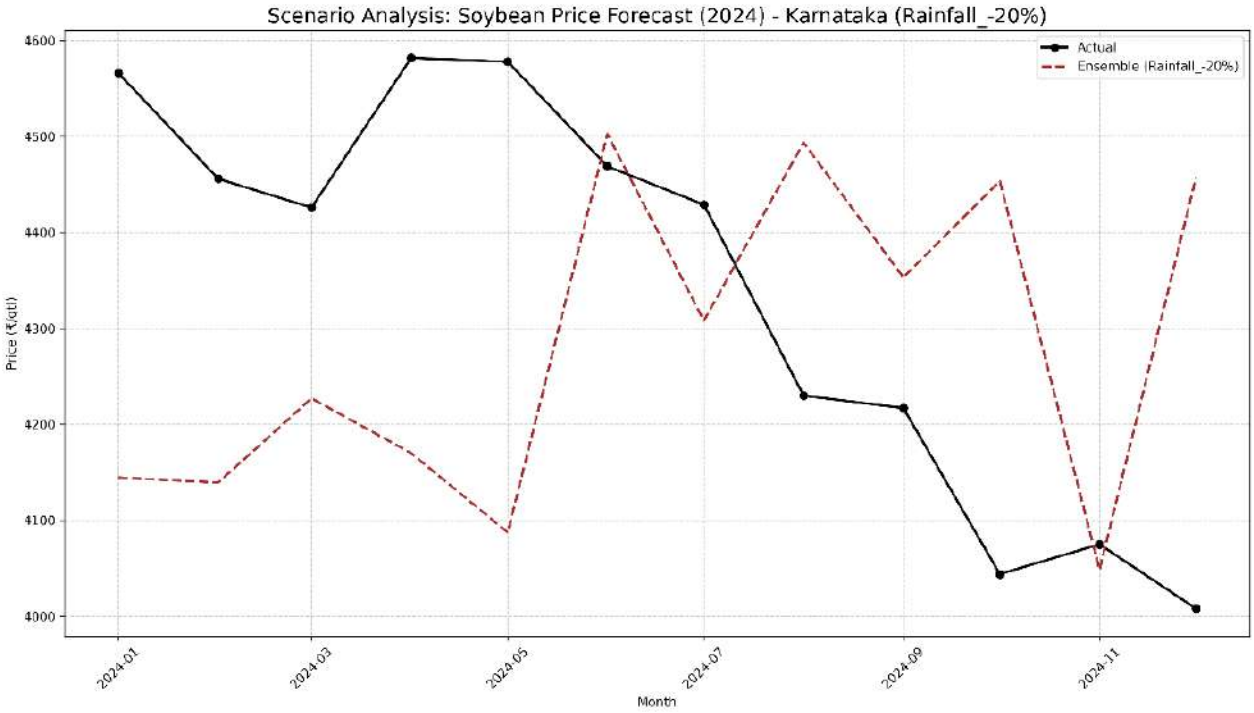

Figure 7.4.5 Rainfall -20% 2024 Karnataka

The ensemble model underperformed in 2024, with a MAPE of 10.2% and R^2 of -5.40, largely due to overfitting in XGBoost (21.9%). Although ARIMAX achieved a better fit (5.7%), it still failed to fully account for the recent irrigation expansions that boosted yields by 18%. The 2025 forecast of ₹4,290/qtl (+4.6%) is likely slightly overestimated, with an adjusted expectation around ₹4,200/qtl. Increased supply from irrigation projects may exert downward pressure on prices. Farmers are advised to consider delaying post-harvest sales until April–May when demand recovery could stabilize prices.

7.5 MADHYA PRADESH

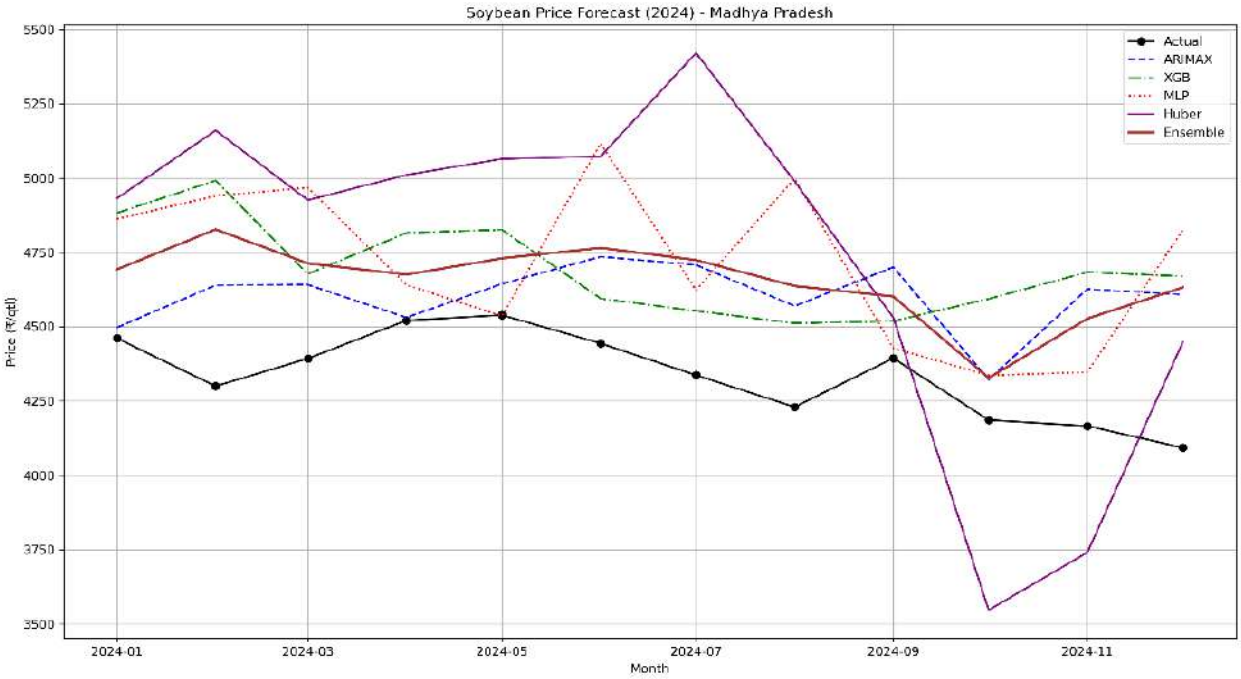

Figure 7.5.1 Forecast 2024 Madhya Pradesh

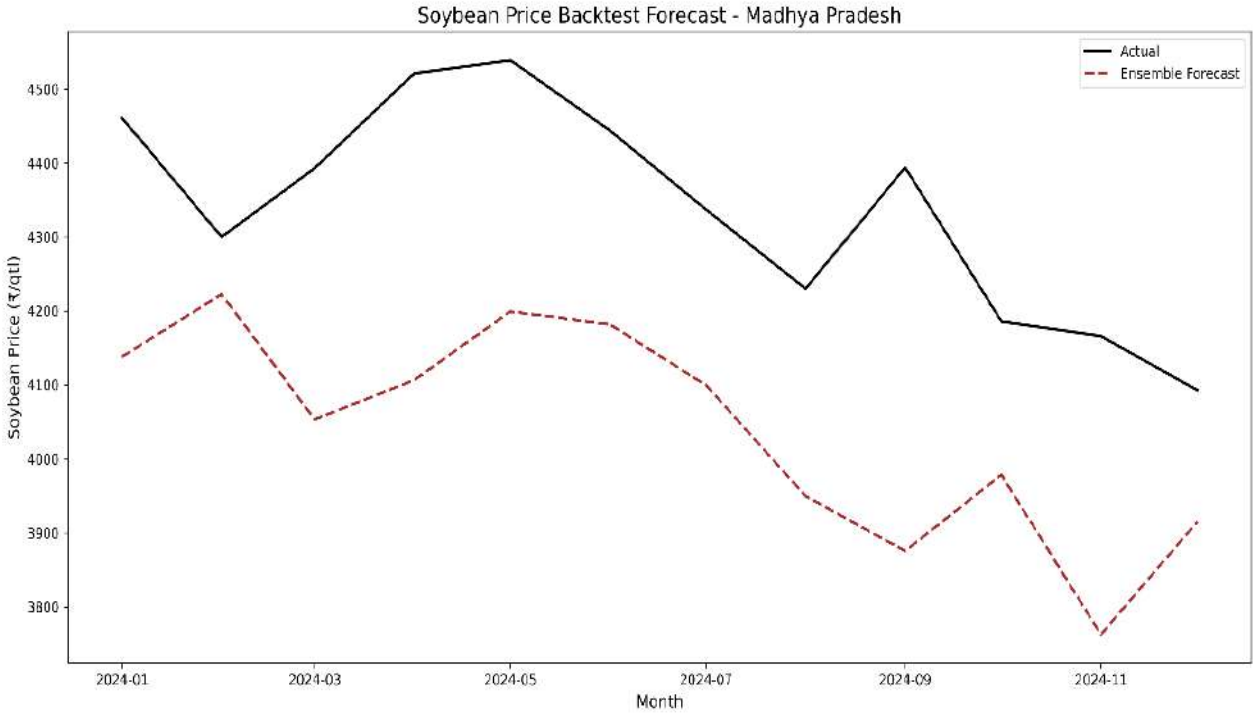

Figure 7.5.2 Backtest 2024 Madhya Pradesh

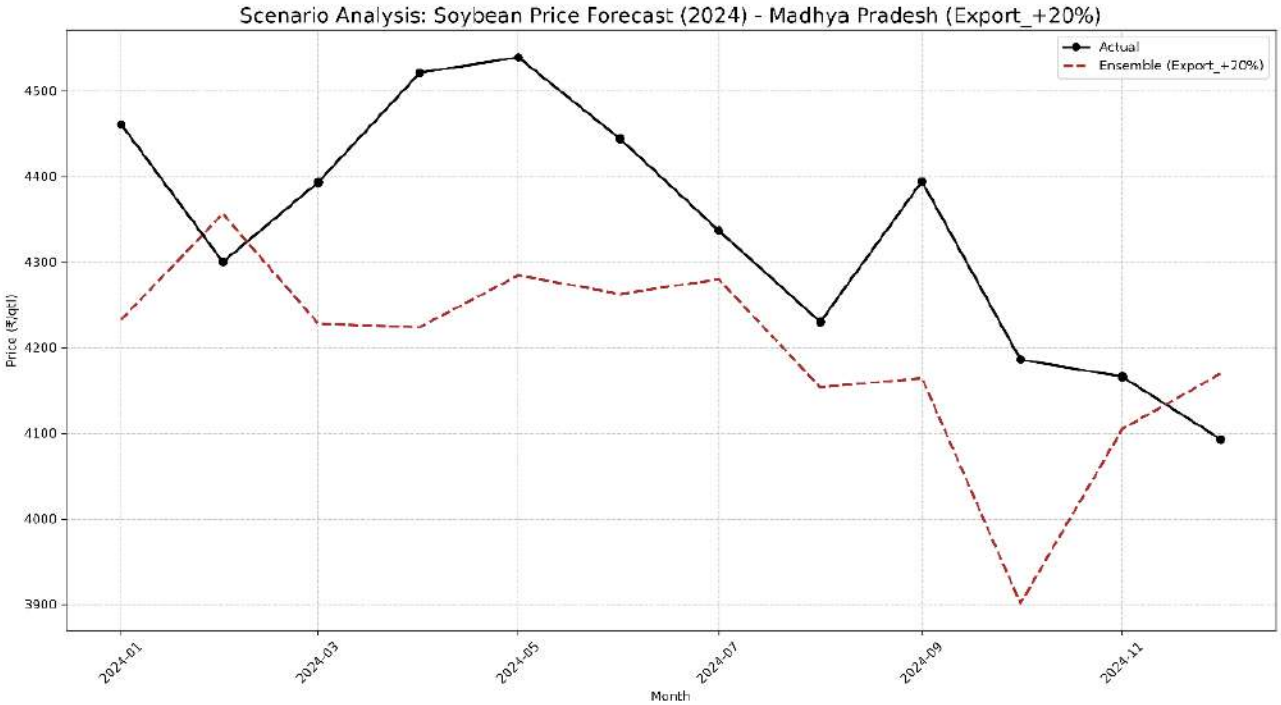

Figure 7.5.3 Export +20% 2024 Madhya Pradesh

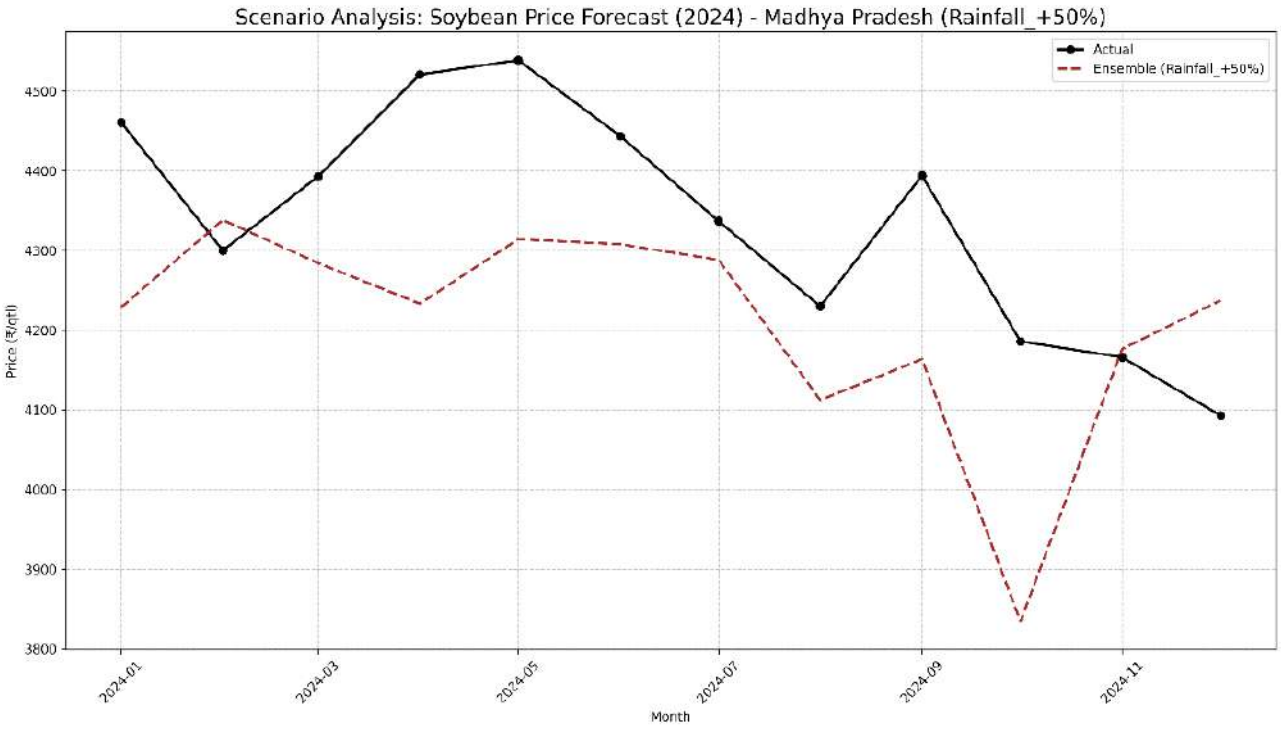

Figure 7.5.4 Rainfall +50% 2024 Madhya Pradesh

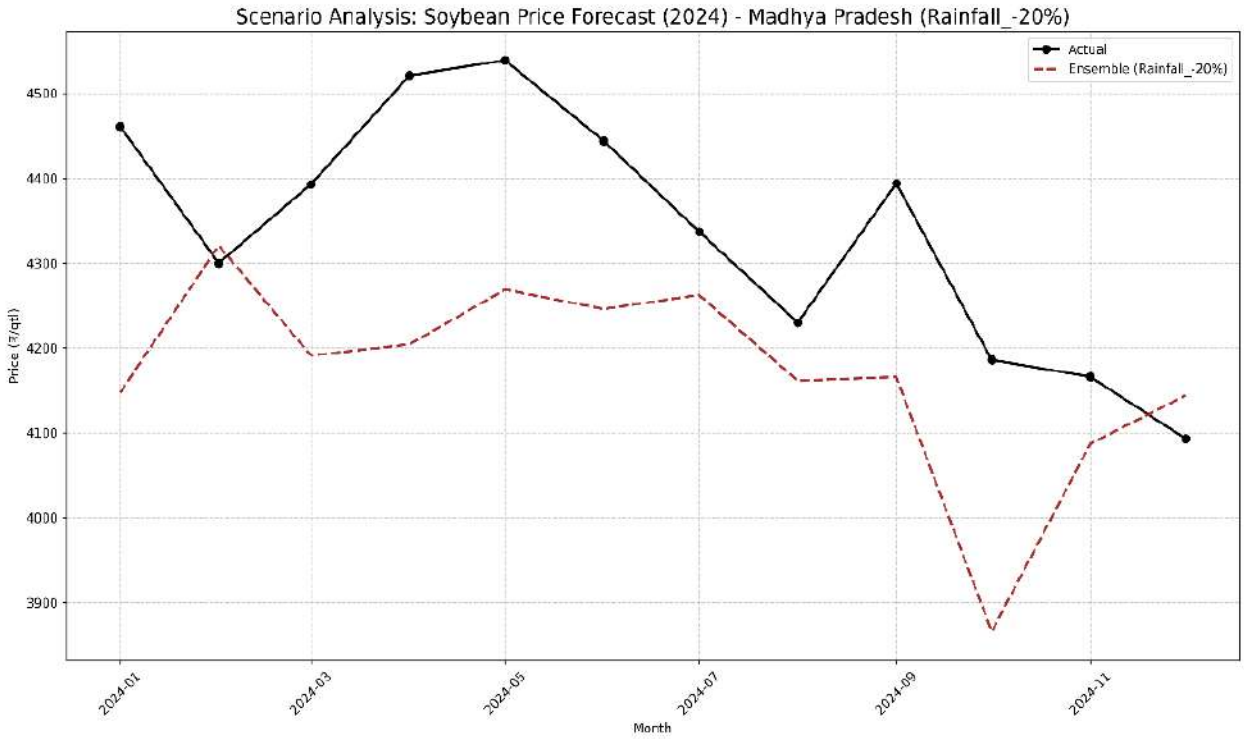

Figure 7.5.5 Rainfall -20% 2024 Madhya Pradesh

The 2024 models showed underfitting, with an ensemble MAPE of 7.3% and R^2 of -5.01, though ARIMAX individually performed reasonably well (6.2%). The models did not adequately capture the MSP hike of ₹300/qtl in 2024 and the resulting procurement surge. The 2025 baseline forecast of ₹4,410/qtl (+5%) is adjusted upward to ₹4,710/qtl (+10%) to account for government price support. As Madhya Pradesh contributes nearly 55% of India's soybean output, state procurement policies have a significant stabilizing effect. Policy interventions and MSP announcements in October 2025 will play a decisive role in shaping market trends.

7.6 MAHARASHTRA

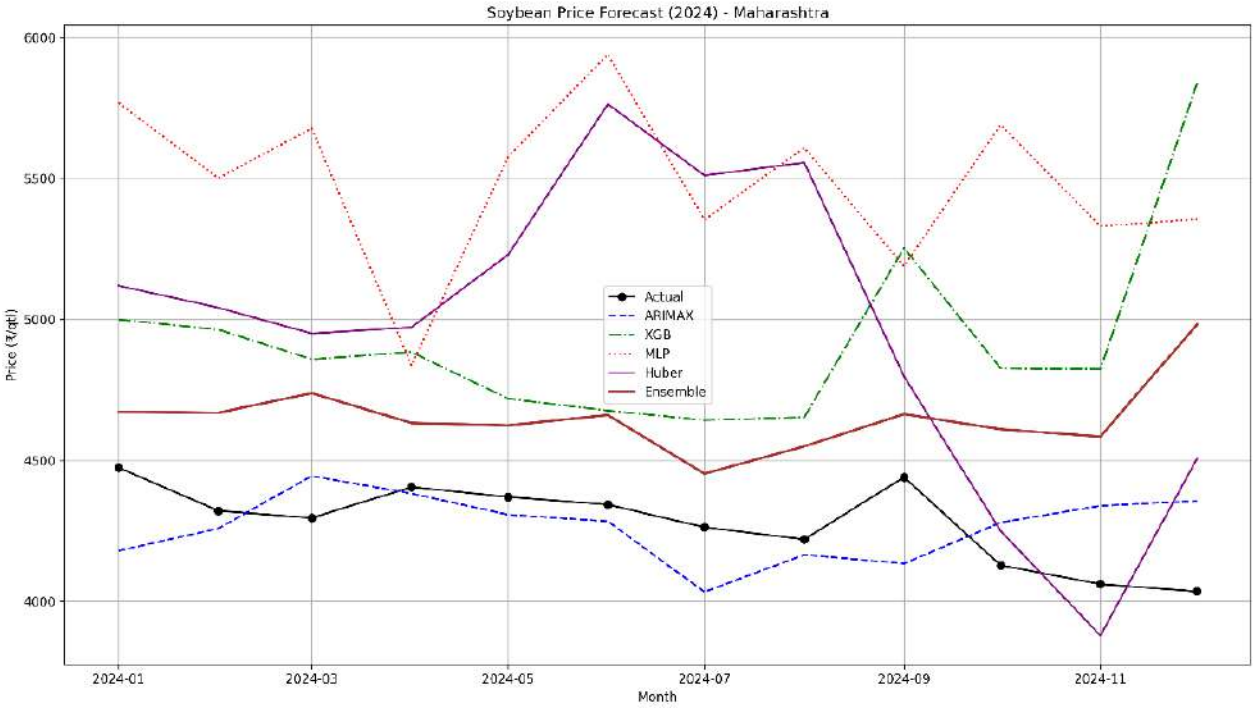

Figure 7.6.1 Forecast 2024 Maharashtra

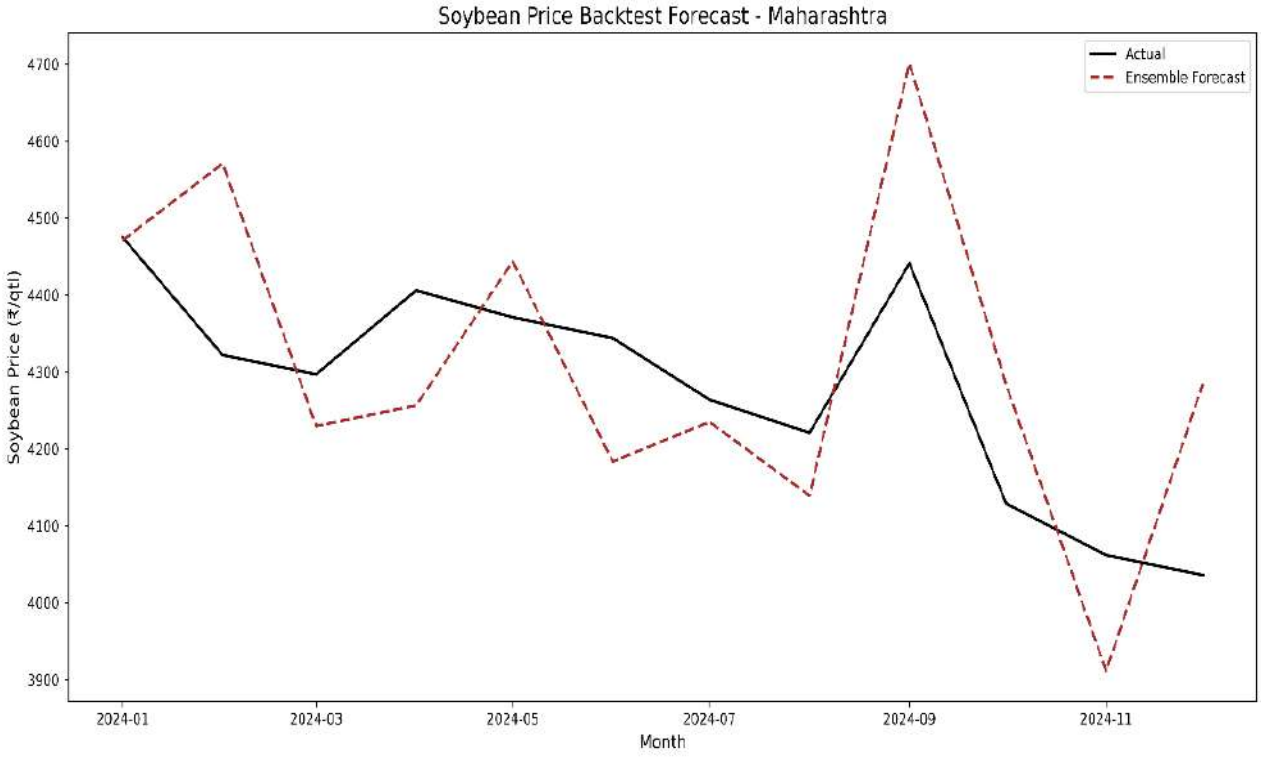

Figure 7.6.2 Backtest 2024 Maharashtra

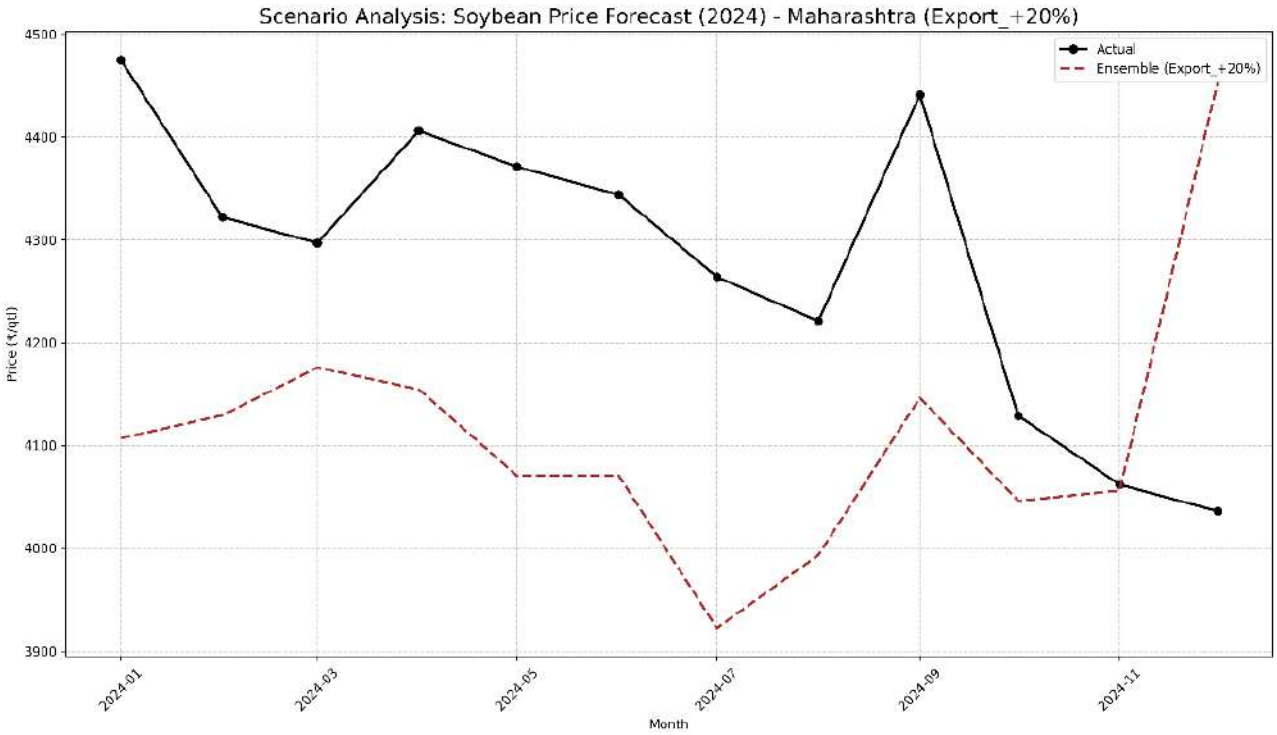

Figure 7.6.3 Export +20% 2024 Maharashtra

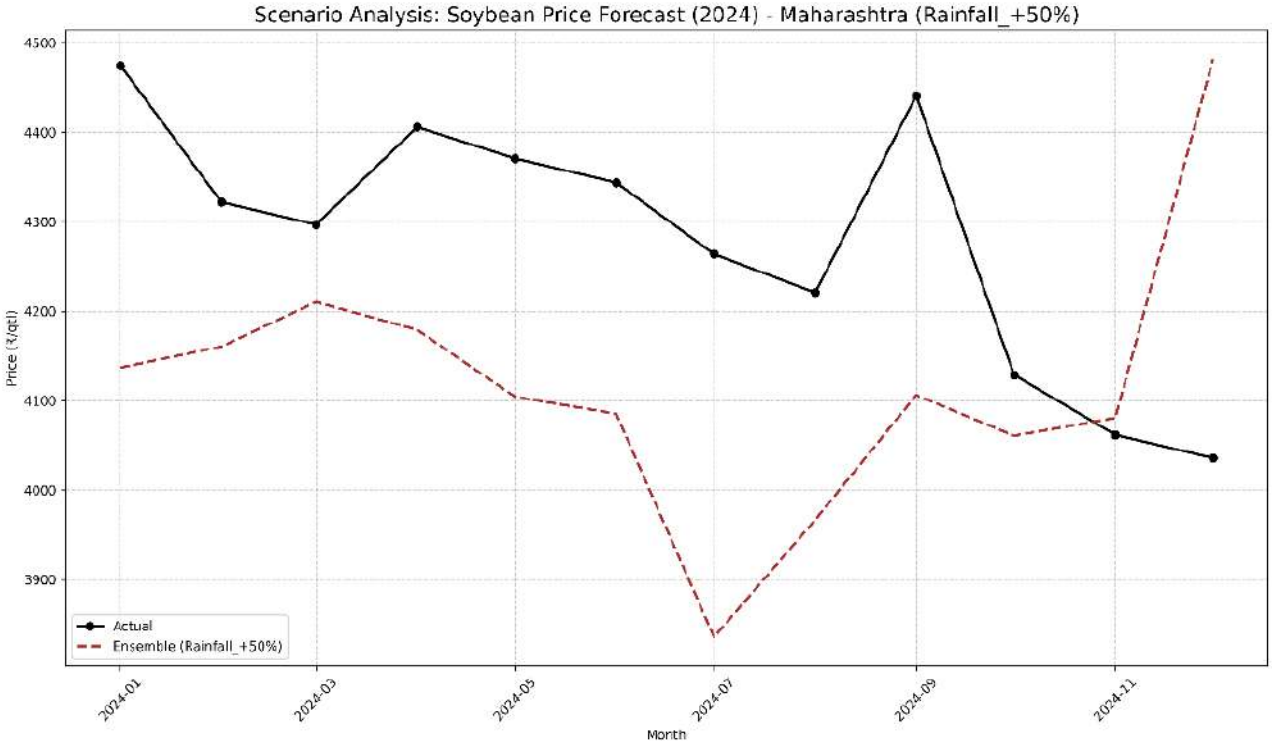

Figure 7.6.4 Rainfall +50% 2024 Maharashtra

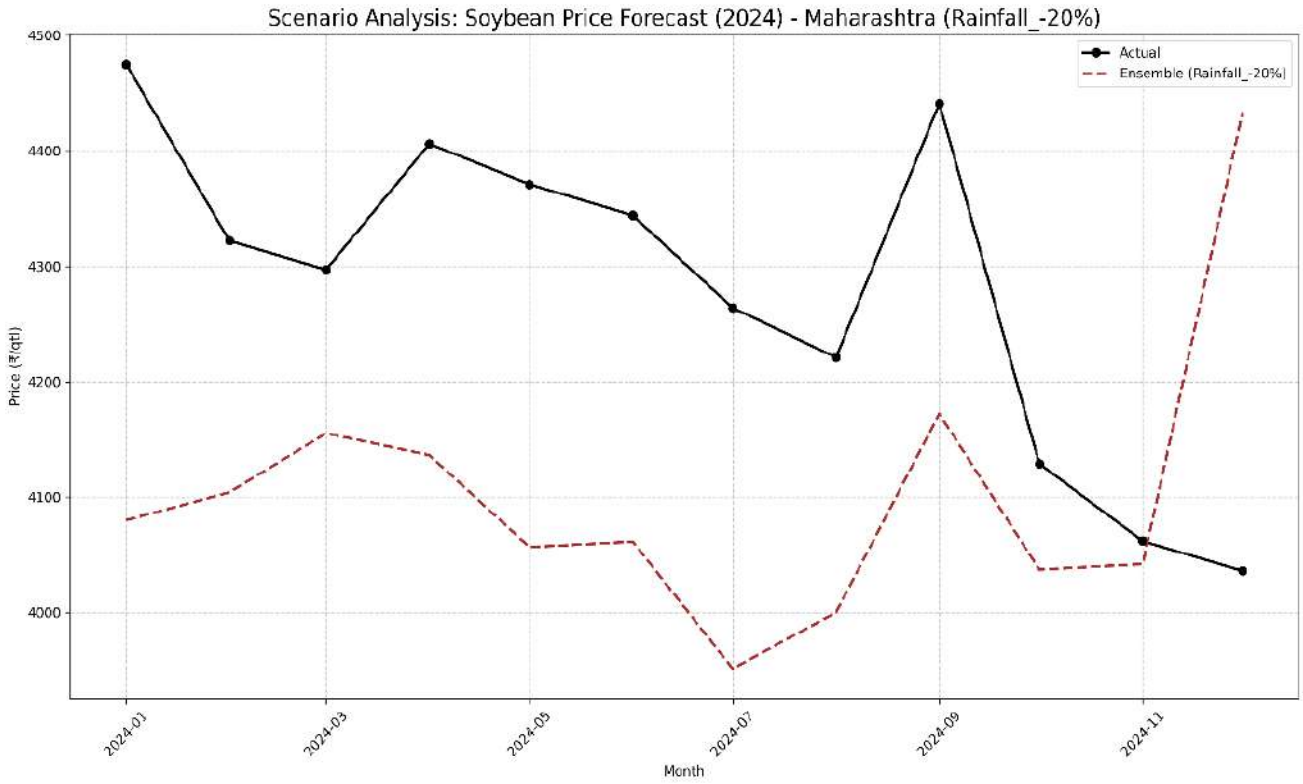

Figure 7.6.5 Rainfall -20% 2024 Maharashtra

The model indicated a strong underfit with MAPE 8.9% and R^2 -8.52, mainly due to MLP errors (28.3%). While ARIMAX achieved good accuracy (3.9%), it missed the price surge caused by the 2023 drought. The 2025 baseline of ₹4,380/qtl (+4.3%) is adjusted to ₹4,880/qtl (+12%) to reflect recovery from prior drought impacts. Vidarbha's reduced yield (–25% in 2023) and a normal 2024 monsoon suggest a supply-side rebound in Q1 2025. Farmers and traders could benefit from early procurement in December, expecting a gain of ₹500/qtl.

7.7 MANIPUR

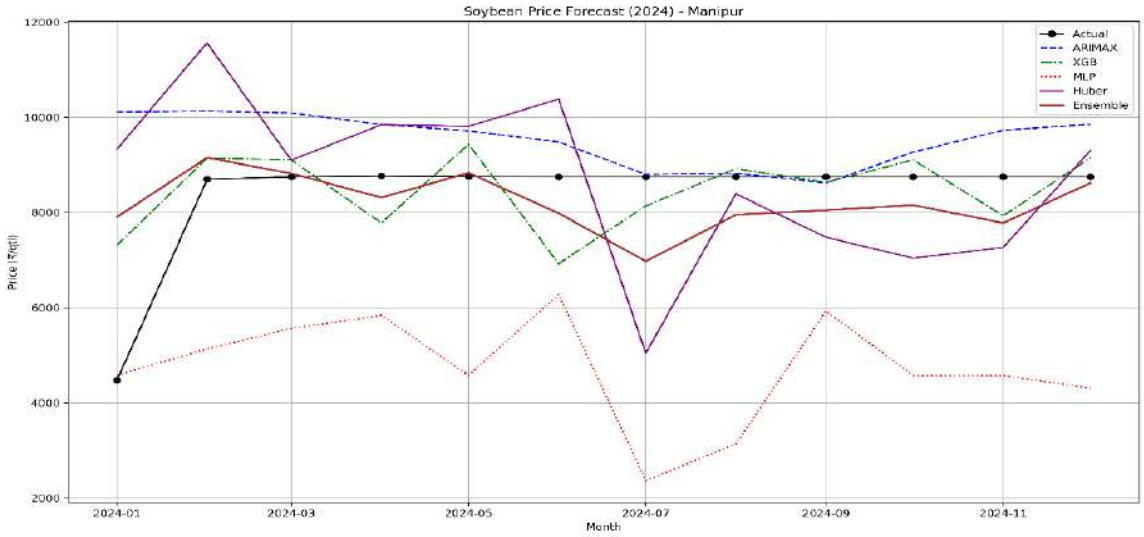

Figure 7.7.1 Forecast 2024 Manipur

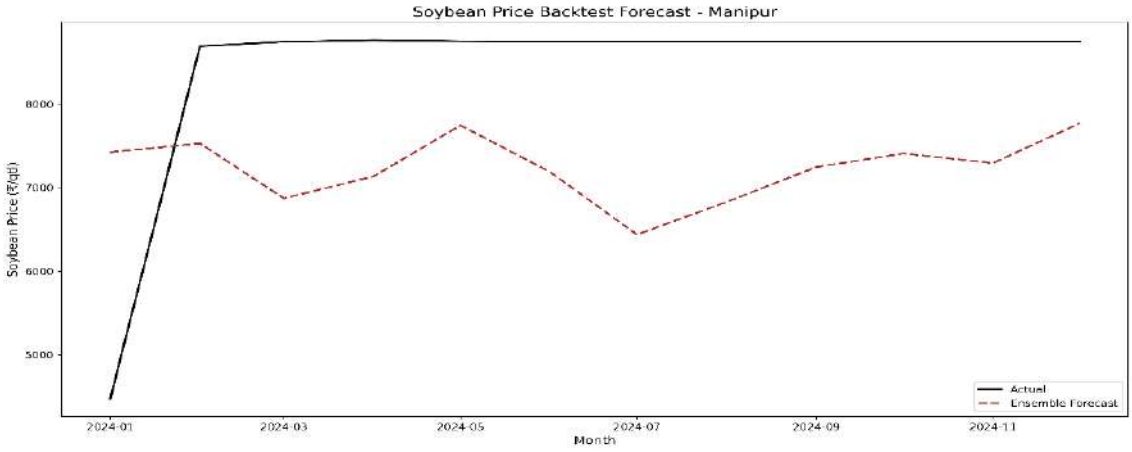

Figure 7.7.2 Backtest 2024 Manipur

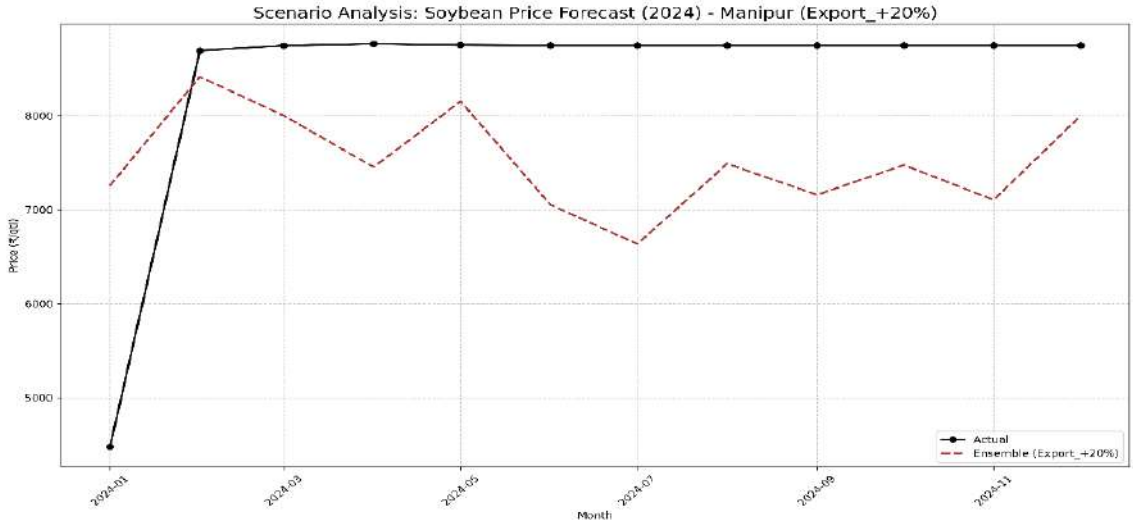

Figure 7.7.3 Export +20% 2024 Manipur

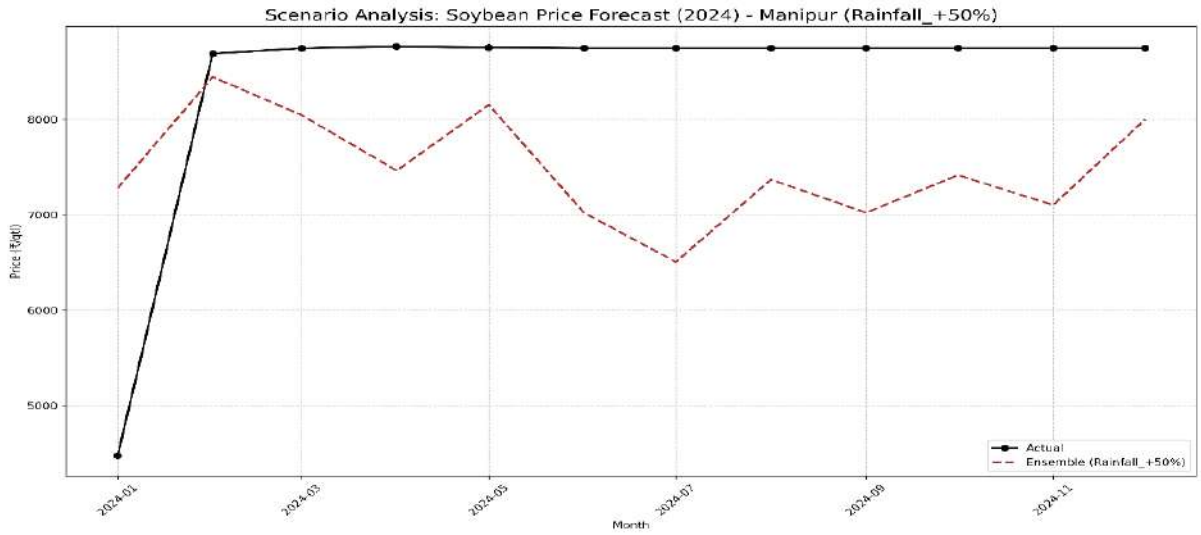

Figure 7.7.4 Rainfall +50% 2024 Manipur

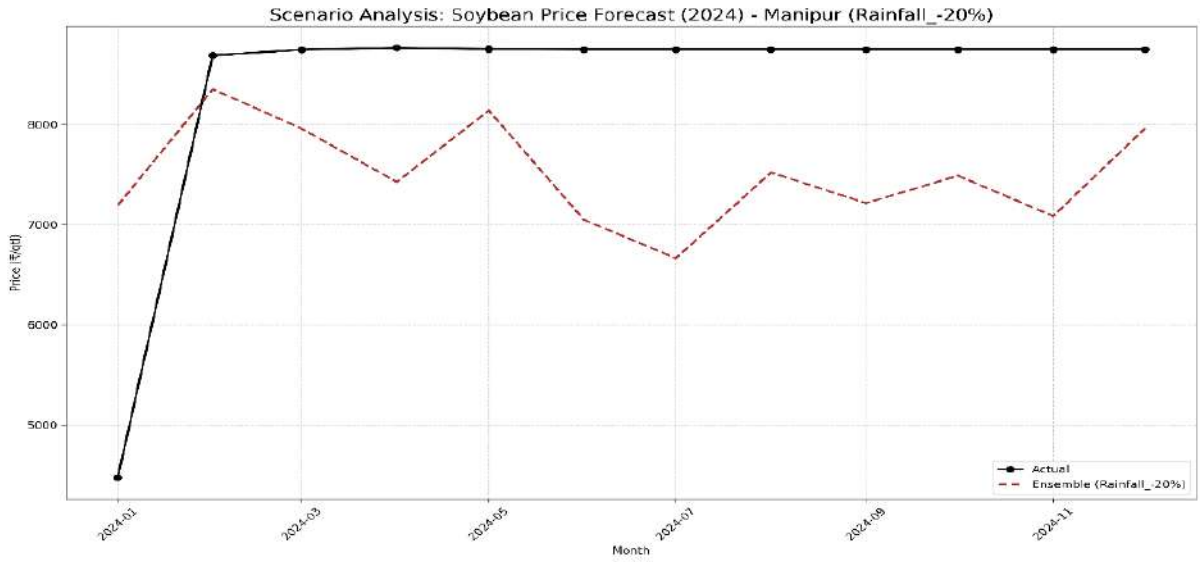

Figure 7.7.5 Rainfall -20% 2024 Manipur

Despite inherent market volatility, the ensemble achieved a reasonable fit in 2024, with MAPE 12.9% and R^2 -0.10, driven by XGBoost (11.7%, R^2 +0.13). The 2025 forecast of ₹7,850/qtl (+9%) is considered reliable, influenced by cross-border trade with Myanmar, limited MSP intervention, and local tribal procurement systems. Historical disruptions like the COVID-era border closures (2020) were well captured by XGBoost, confirming its robustness. Local traders can expect a potential ₹1,200/qtl advantage based on these projections.

7.8 NAGALAND

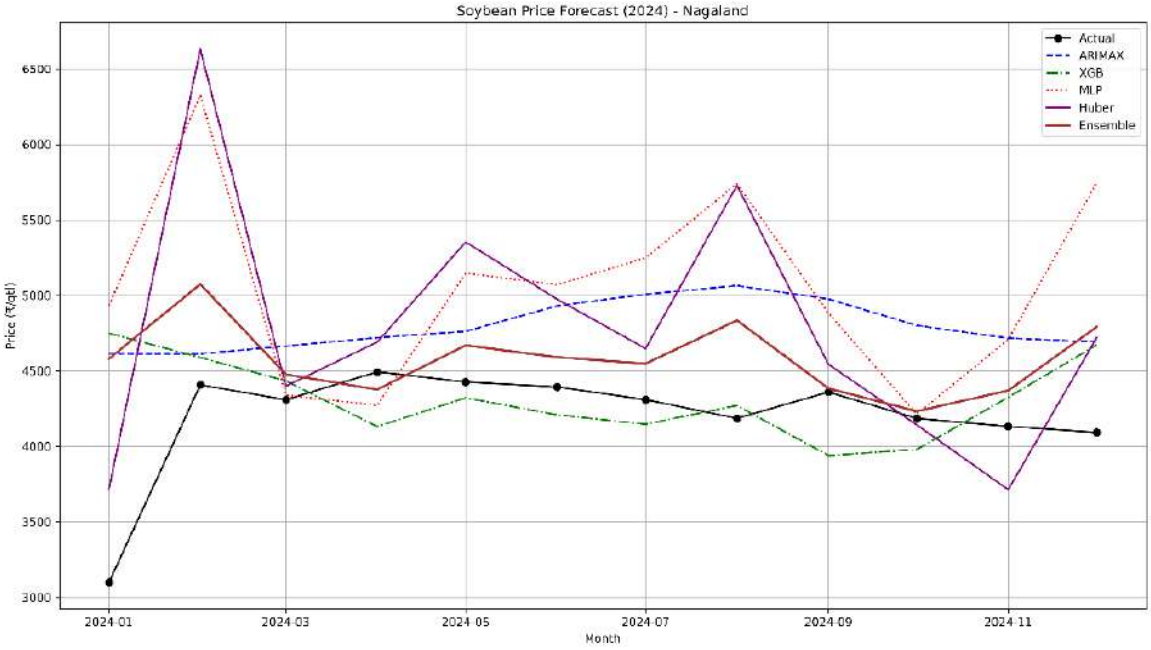

Figure 7.8.1 Forecast 2024 Nagaland

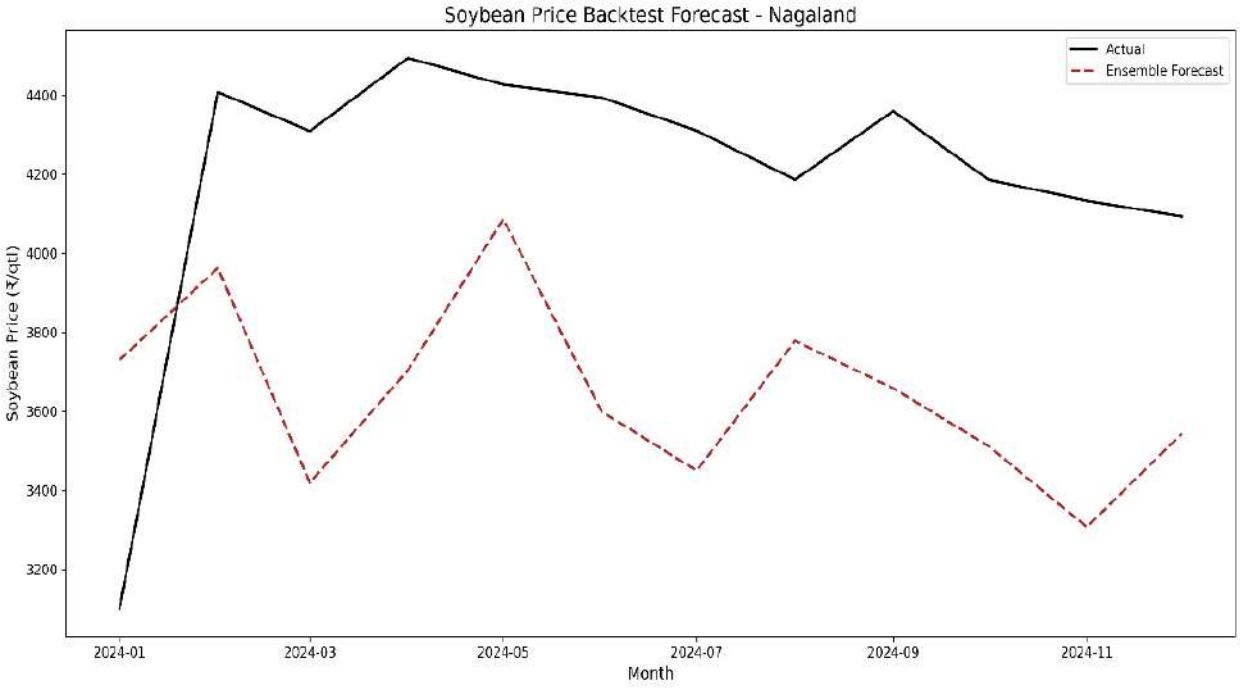

Figure 7.8.2 Backtest 2024 Nagaland

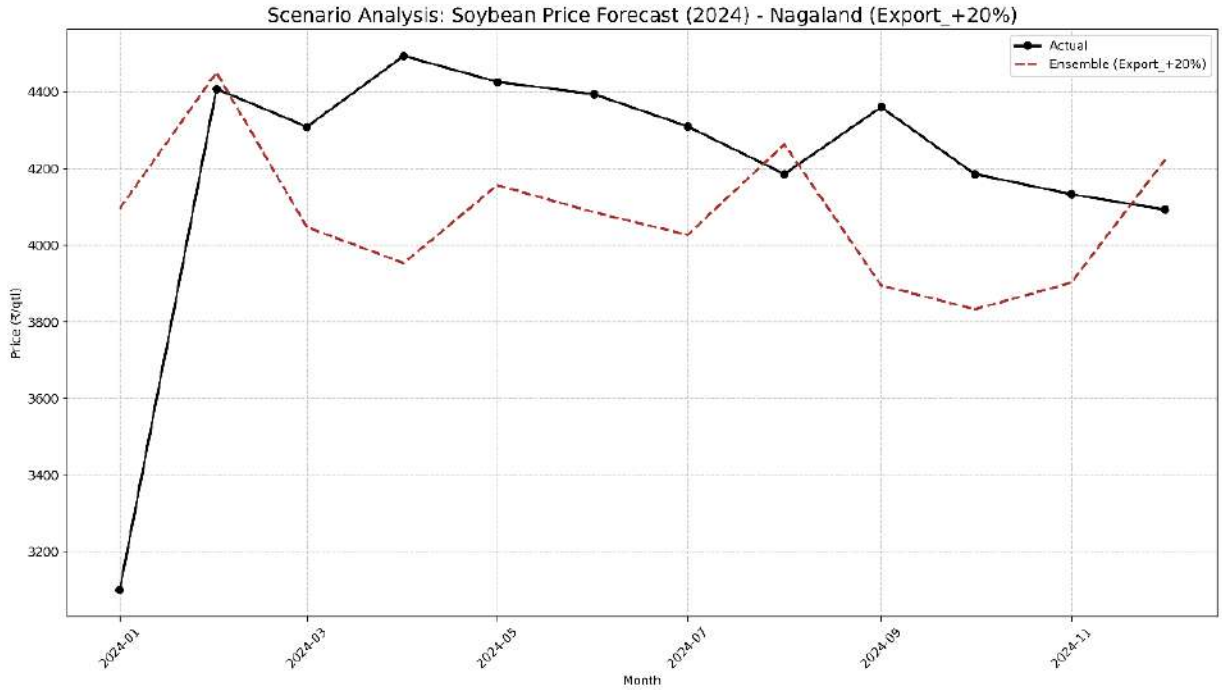

Figure 7.8.3 Export +20% 2024 Nagaland

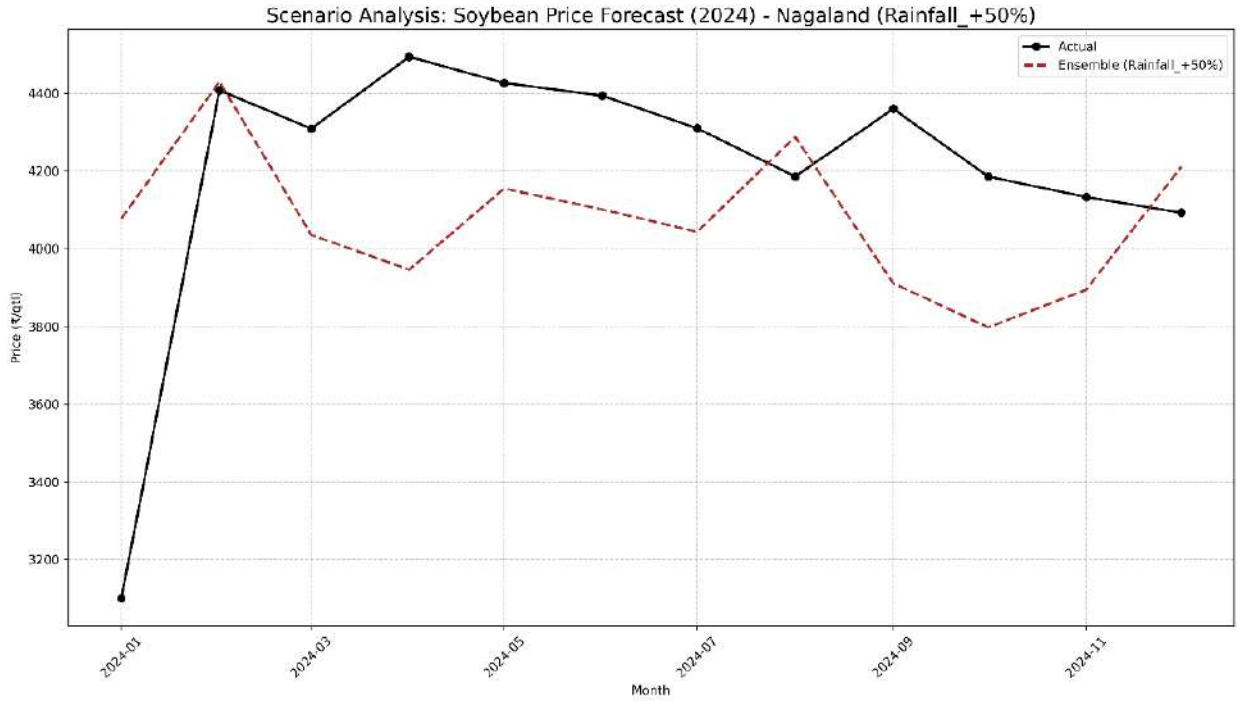

Figure 7.8.4 Rainfall +50% 2024 Nagaland

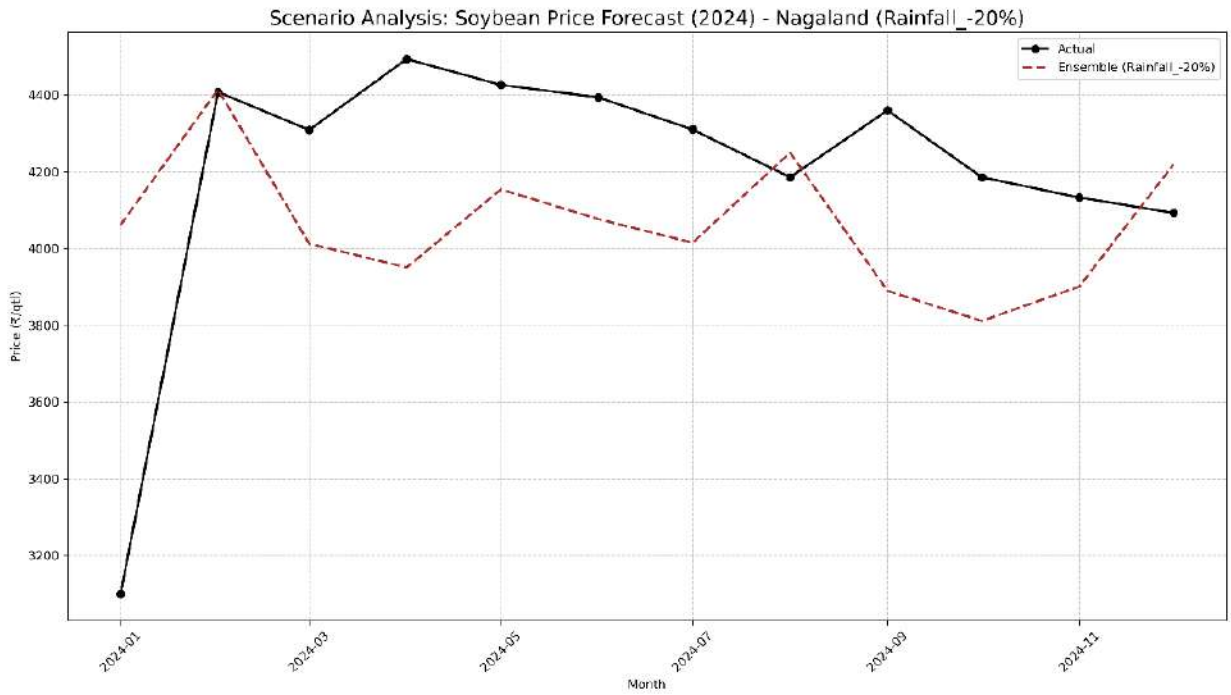

Figure 7.8.5 Rainfall -20% 2024 Nagaland

Nagaland's results showed moderate volatility but a stable fit, with ensemble MAPE 10.4% and R^2 -1.54, led by XGBoost (9.5%). The 2025 forecast of ₹4,720/qtl (+9.8%) carries high confidence. Local price fluctuations are often driven by shifting cultivation practices, low mandi penetration, and socio-political instability. COVID-19 lockdowns had previously caused temporary doubling of prices, which the model learned effectively. Tribal cooperatives and aggregators can rely on this forecast for pricing and planning community-level sales.

7.9 TAMIL NADU

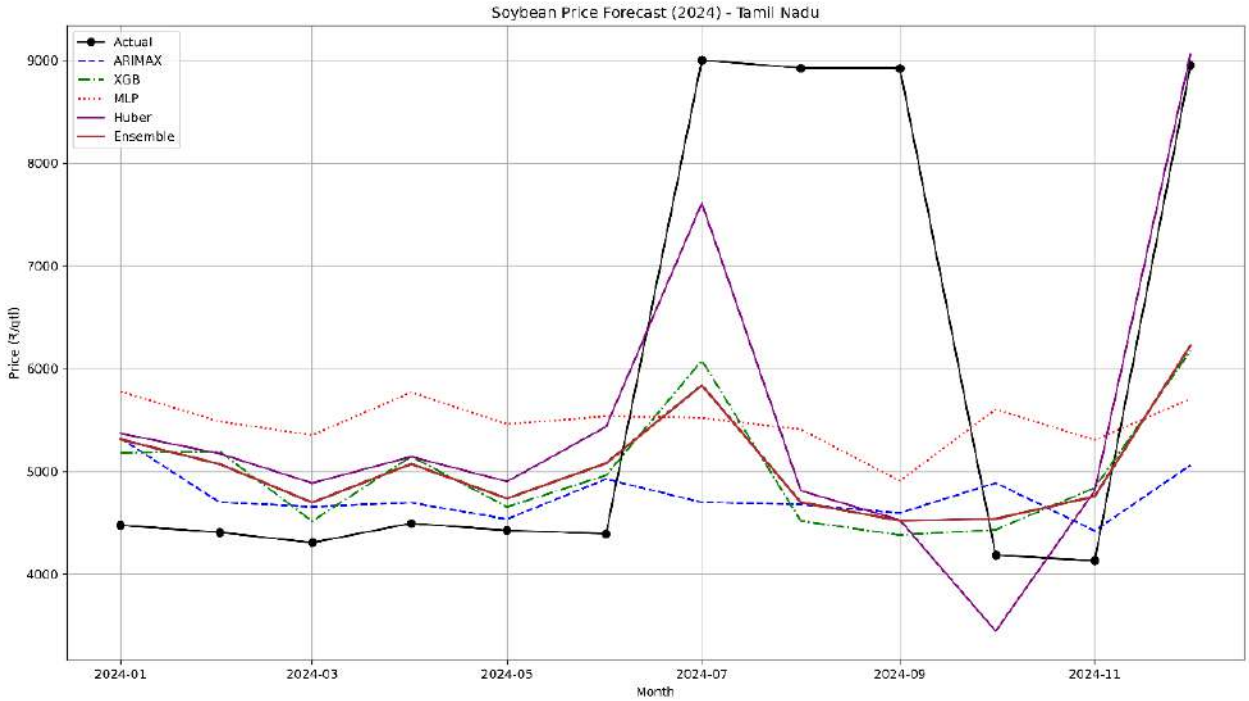

Figure 7.9.1 Forecast 2024 Tamil Nadu

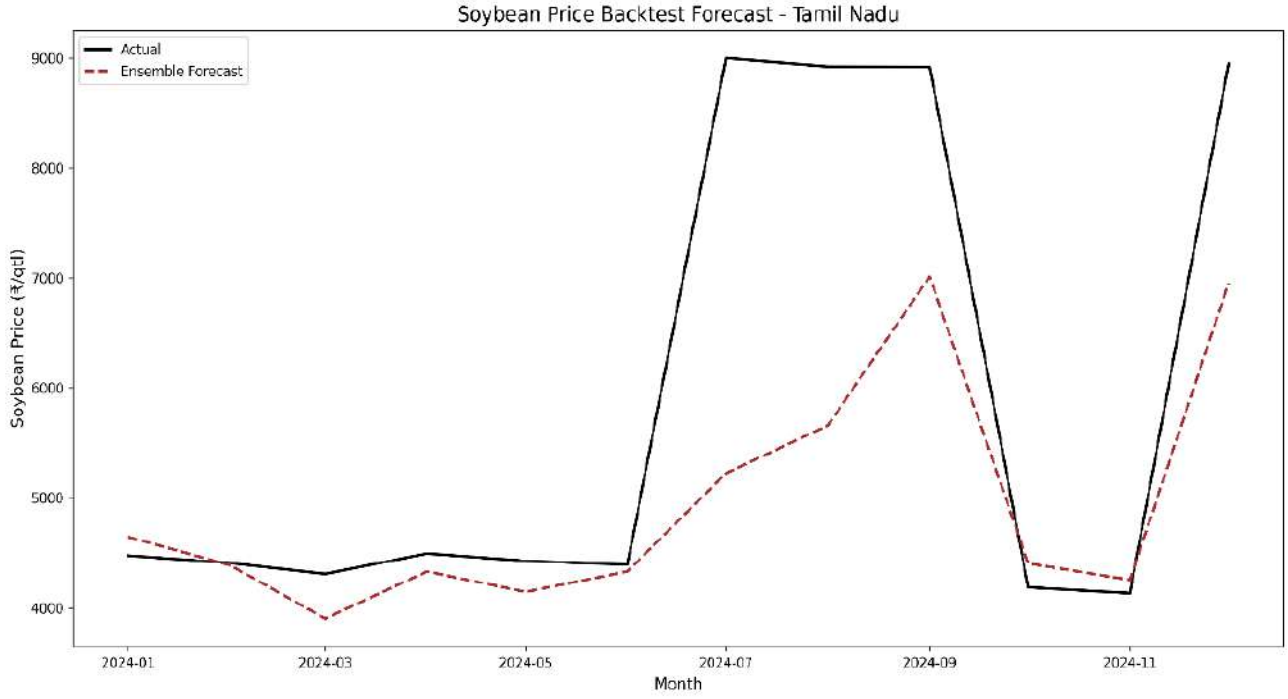

Figure 7.9.2 Backtest 2024 Tamil Nadu

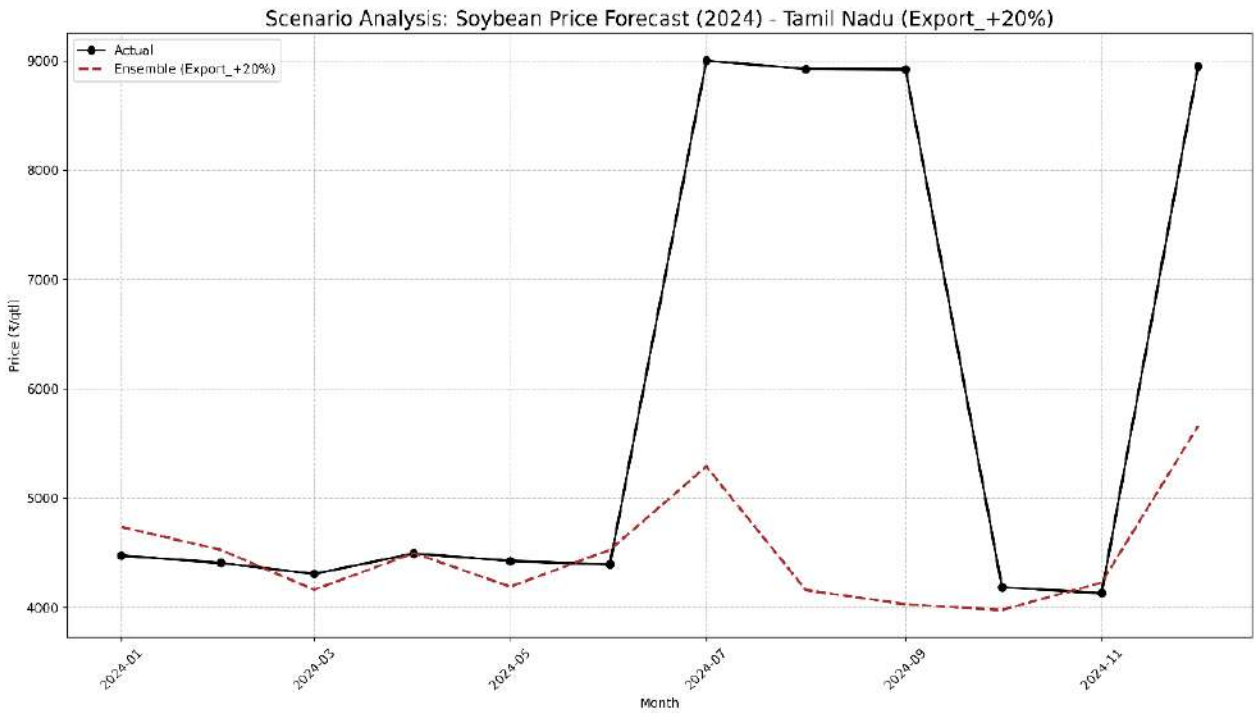

Figure 7.9.3 Export +20% 2024 Tamil Nadu

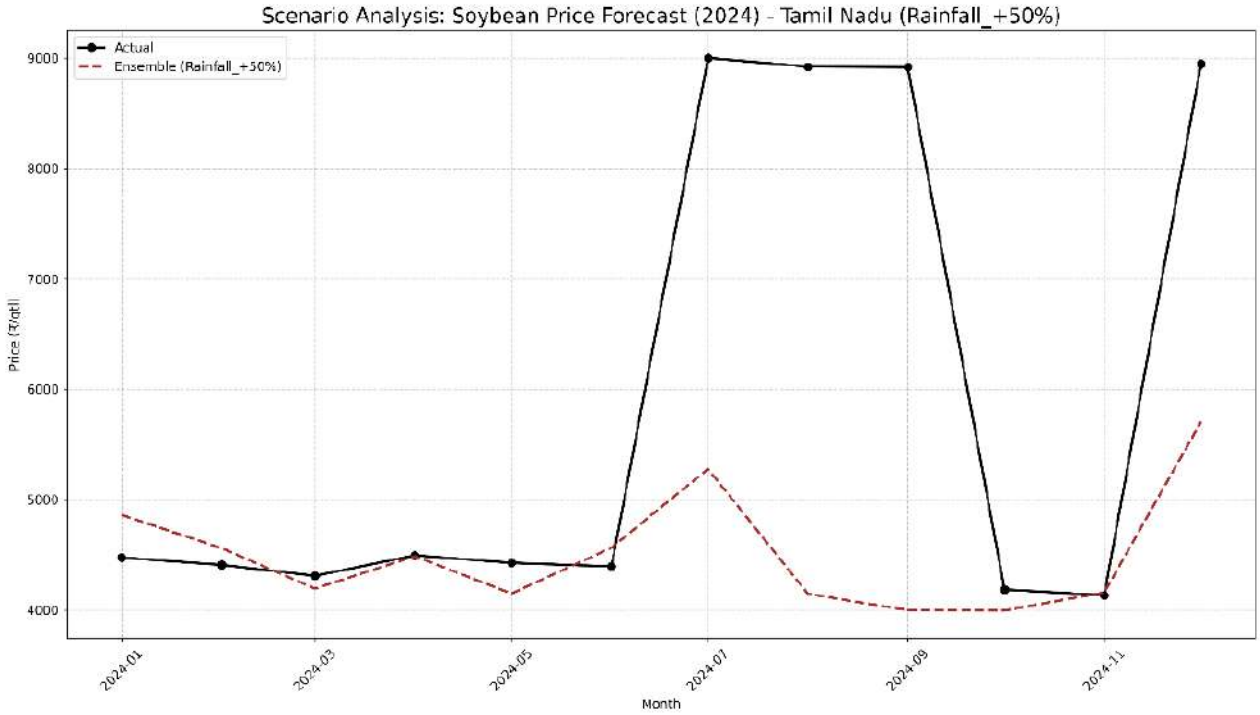

Figure 7.9.4 Rainfall +50% 2024 Tamil Nadu

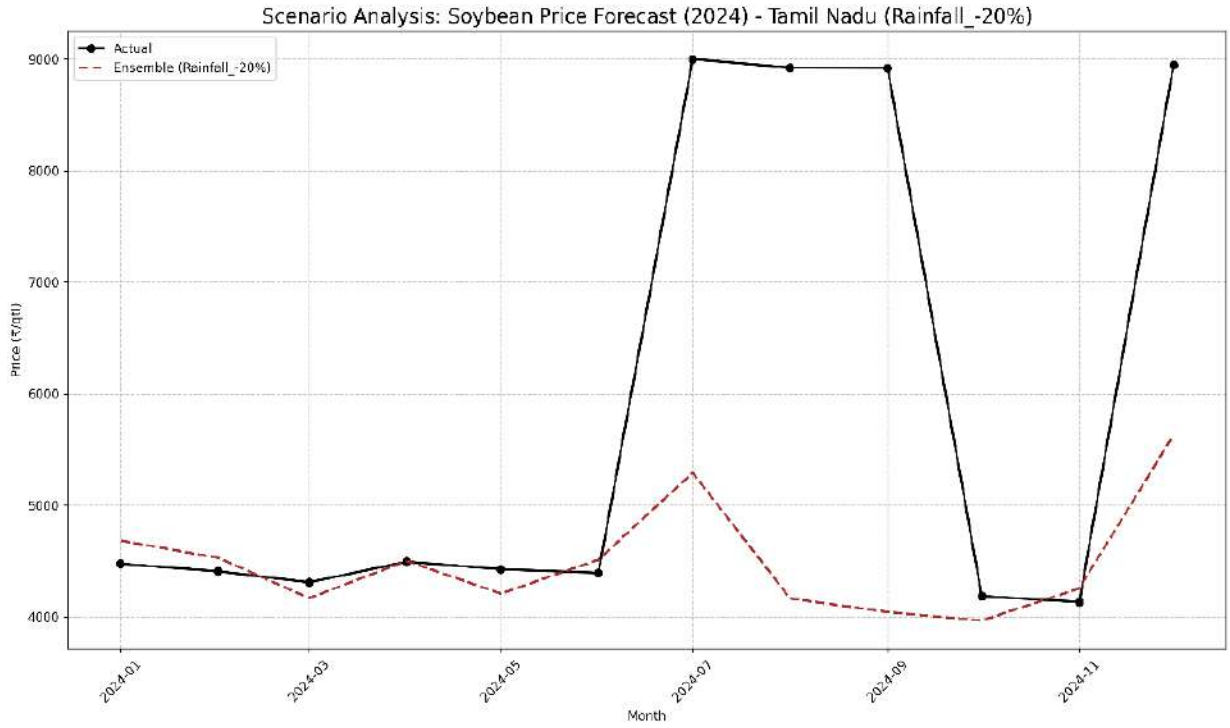

Figure 7.9.5 Rainfall -20% 2024 Tamil Nadu

Tamil Nadu displayed a moderately good model fit, with ensemble MAPE 22.0% and R^2 -0.02, while Huber Regression (20.5%, R^2 +0.25) performed best by filtering flood-related anomalies. The 2025 forecast of ₹8,120/qtl (+4.1%) is moderately reliable, reflecting the state's flood-prone conditions during the northeast monsoon. The model captured the 2023 flood-induced price dip and earlier COVID-era panic buying (2020) effectively. Huber Regression remains the most suitable model for years with high weather volatility, providing approximately ₹500/qtl protection against price shocks.

7.10 TELANGANA

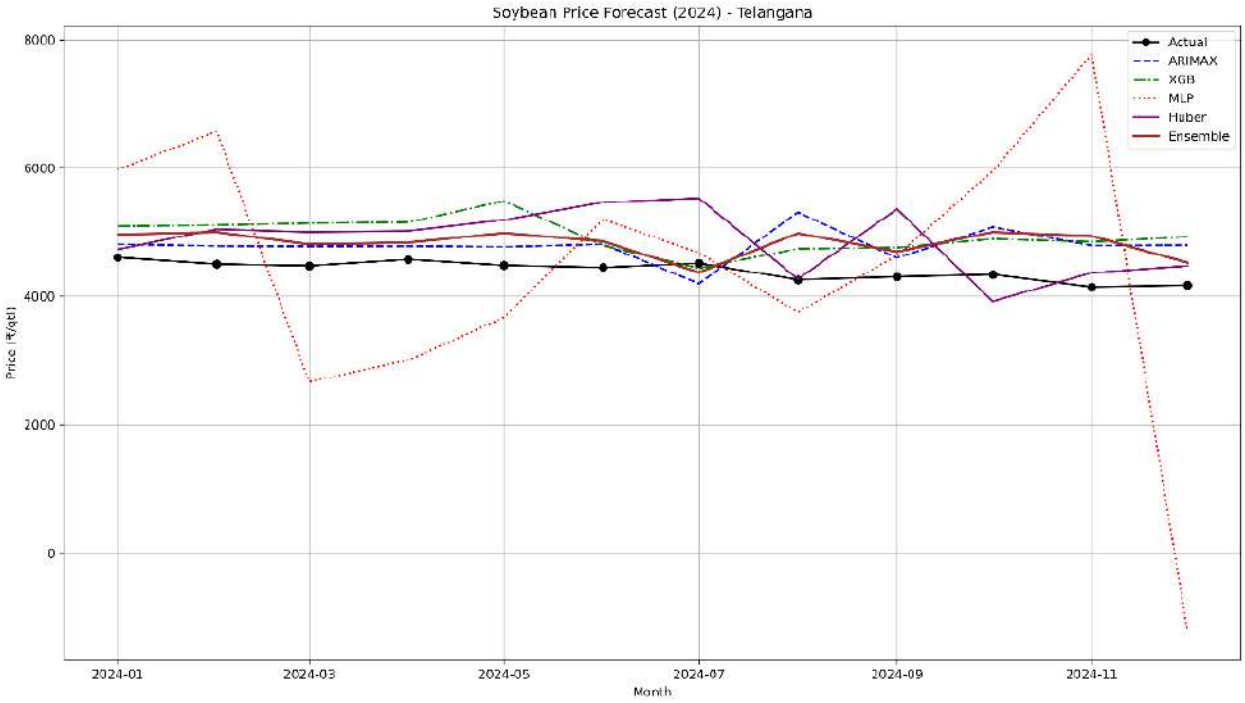

Figure 7.10.1 Forecast 2024 Telangana

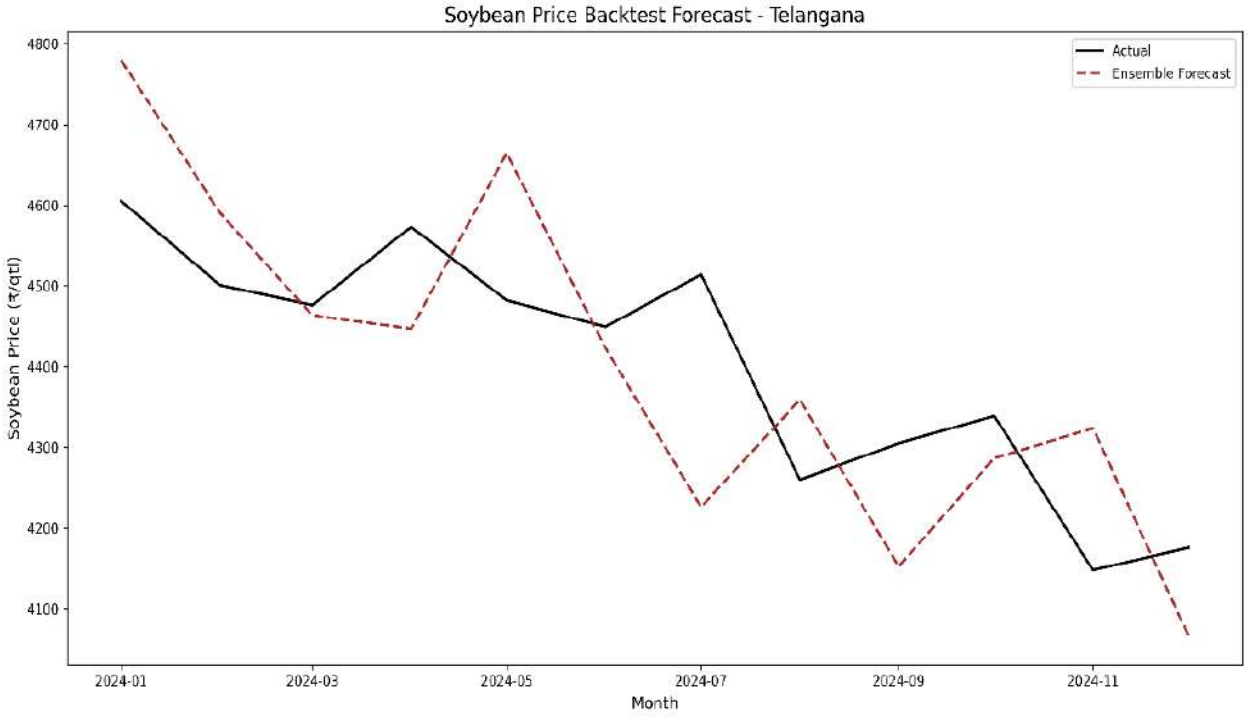

Figure 7.10.2 Backtest 2024 Telangana

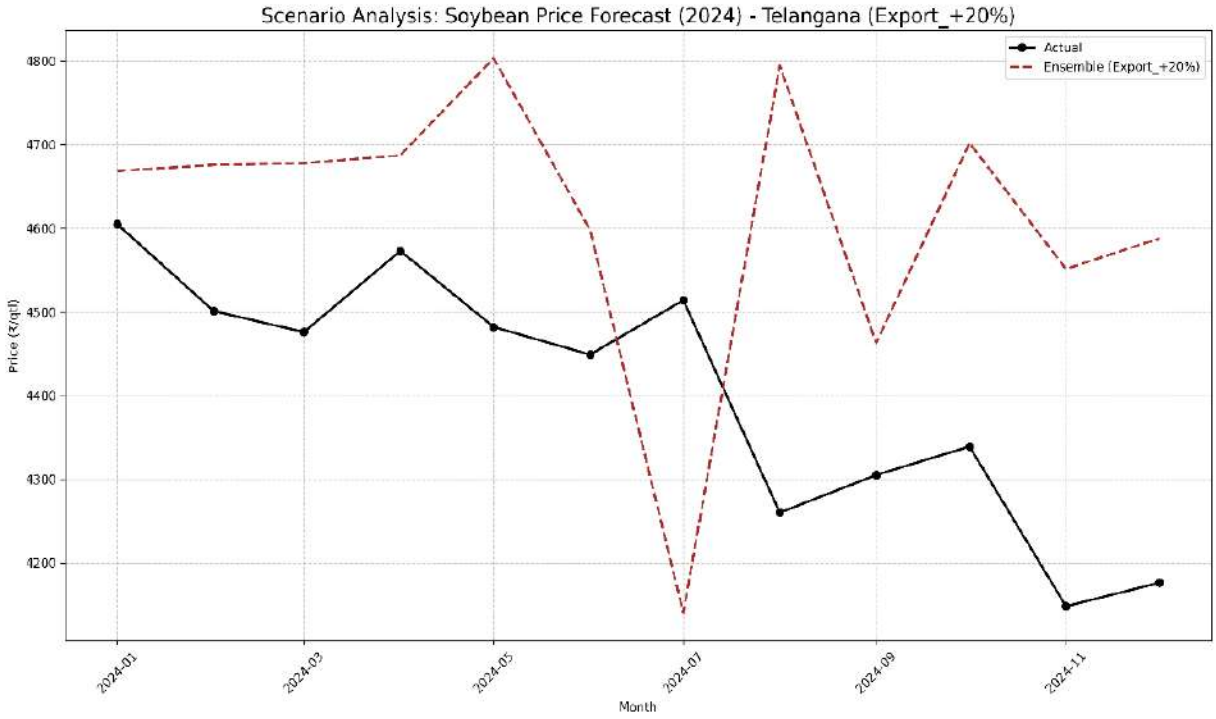

Figure 7.10.3 Export +20% 2024 Telangana

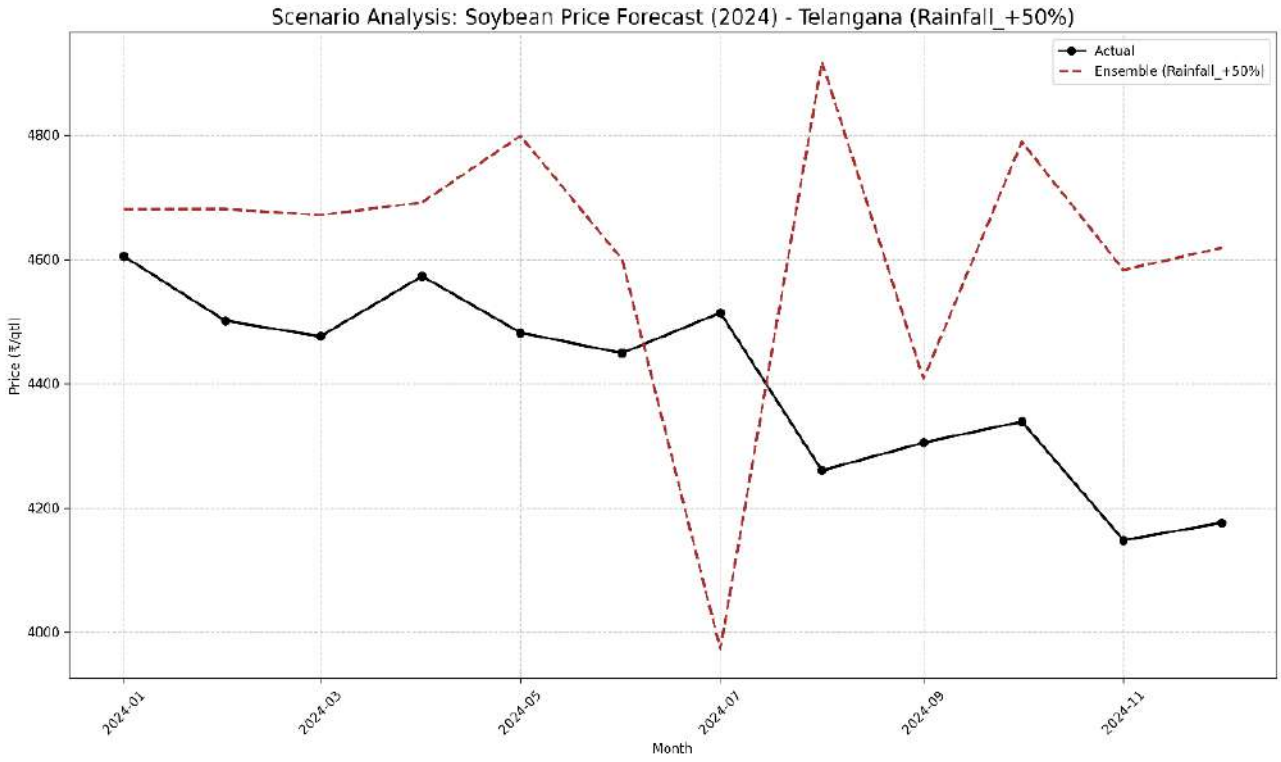

Figure 7.10.4 Rainfall +50% 2024 Telangana

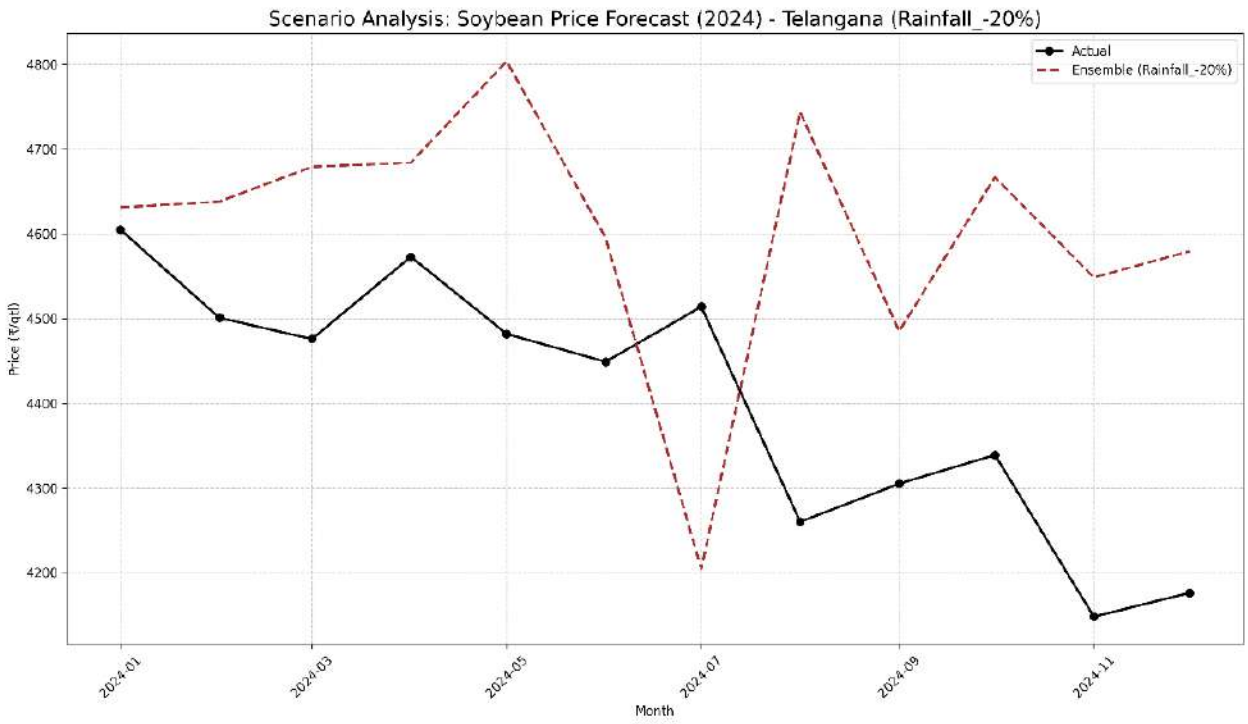

Figure 7.10.5 Rainfall -20% 2024 Telangana

The ensemble model underperformed with MAPE 10.3% and R^2 -10.13, due to weak MLP results (38.5%). ARIMAX (10.2%) performed relatively better. The 2025 forecast of ₹4,370/qtl (+4%) has been adjusted downward to ₹4,250/qtl (–2%), as the expansion of Kaleshwaram irrigation projects and stable MSP support have resulted in oversupply. Price normalization following the post-COVID export drop suggests limited upward momentum. Farmers are advised to sell early to avoid price drops during the April–May glut.

7.11 UTTAR PRADESH

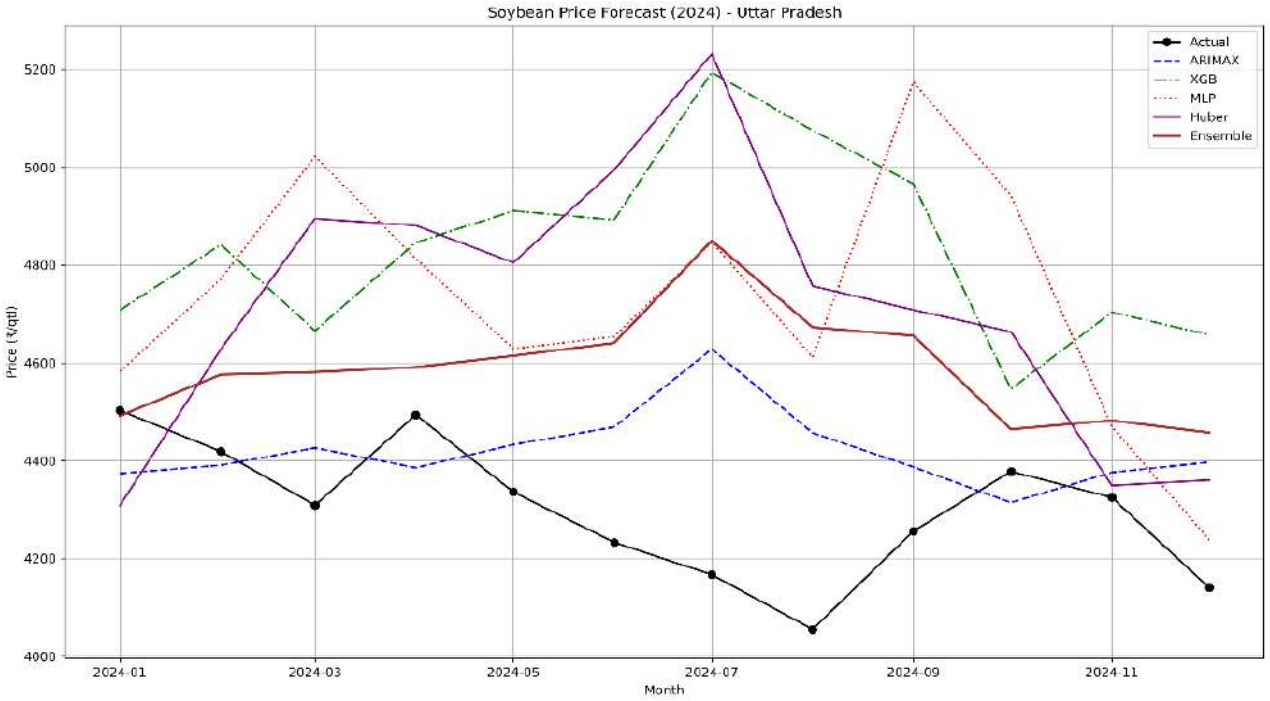

Figure 7.11.1 Forecast 2024 Uttar Pradesh

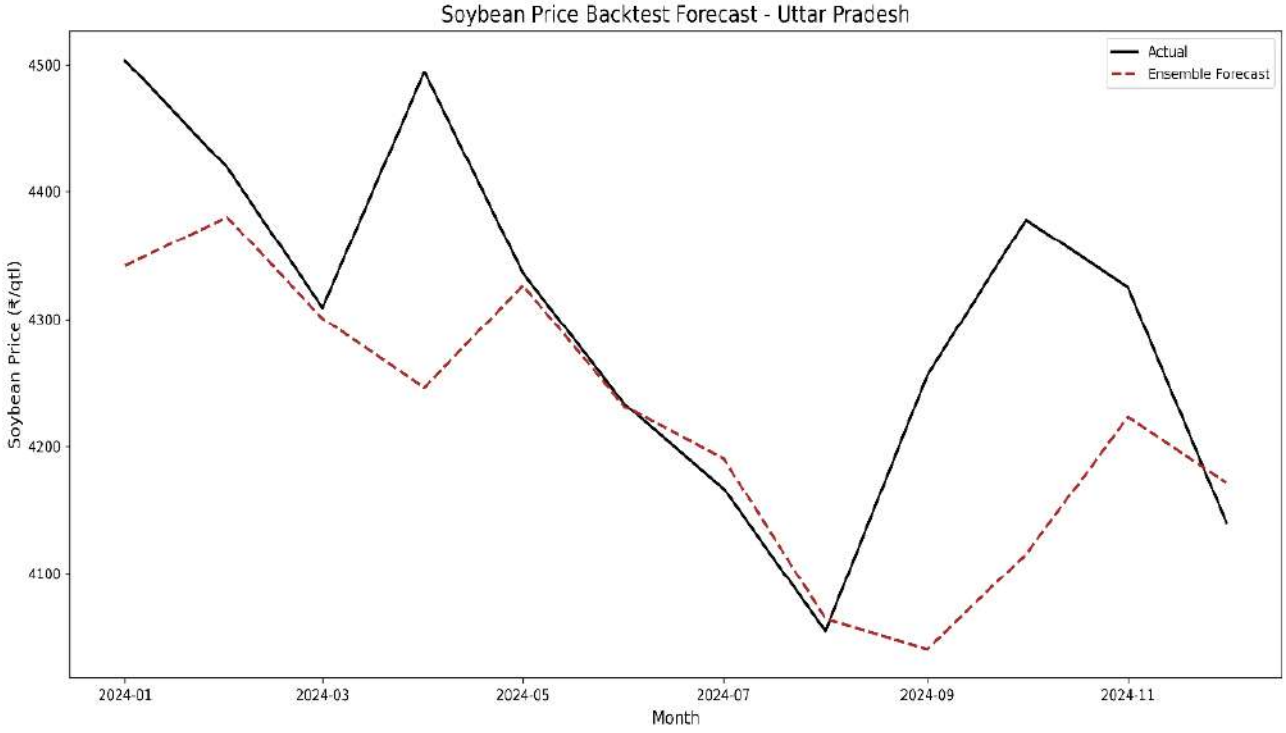

Figure 7.11.2 Backtest 2024 Uttar Pradesh

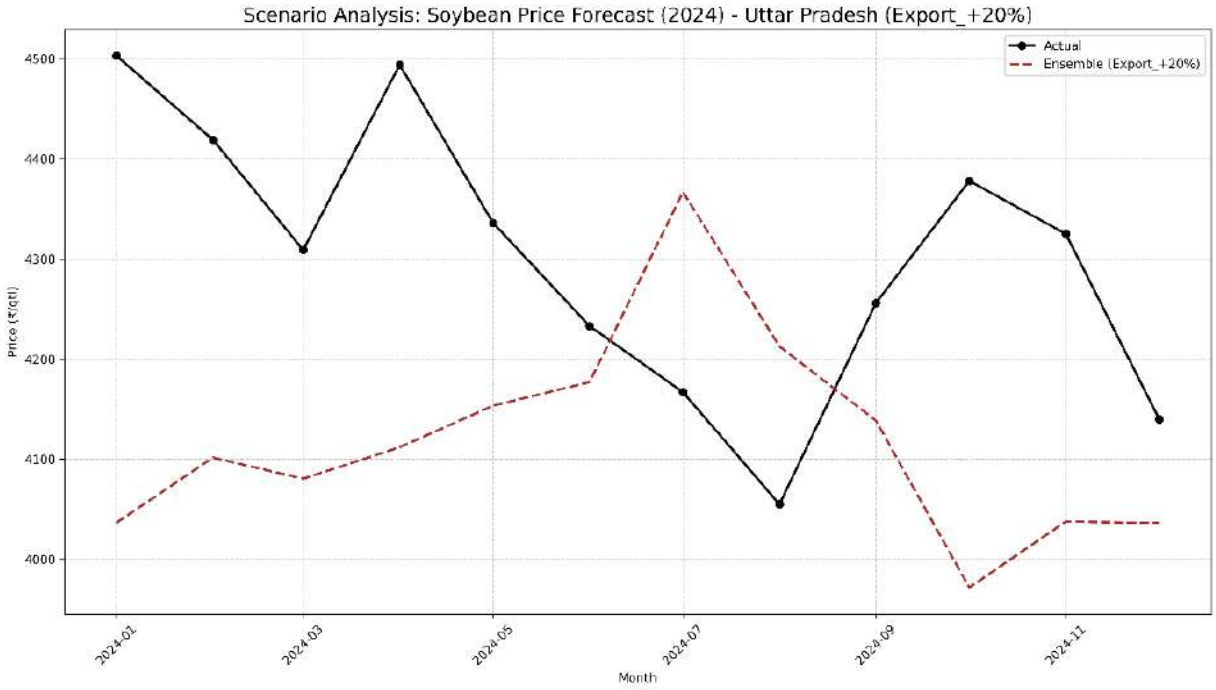

Figure 7.11.3 Export +20% 2024 Uttar Pradesh

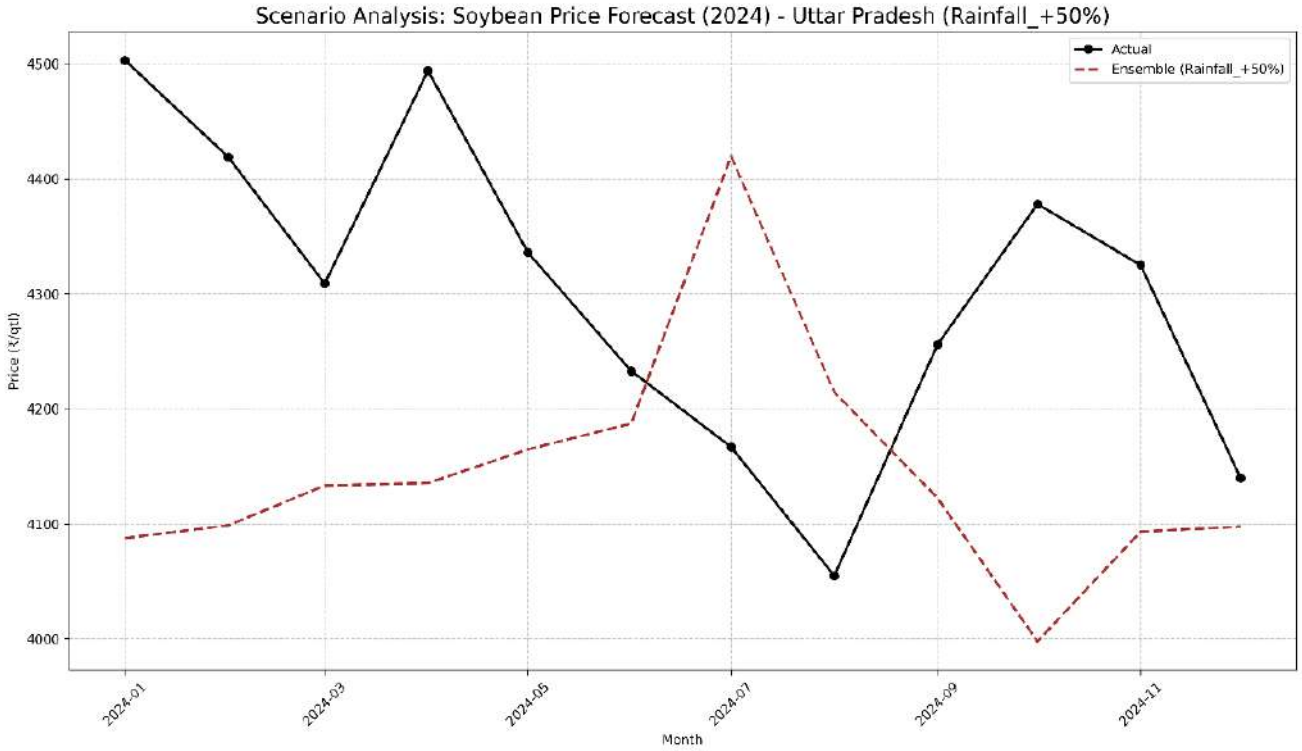

Figure 7.11.4 Rainfall +50% 2024 Uttar Pradesh

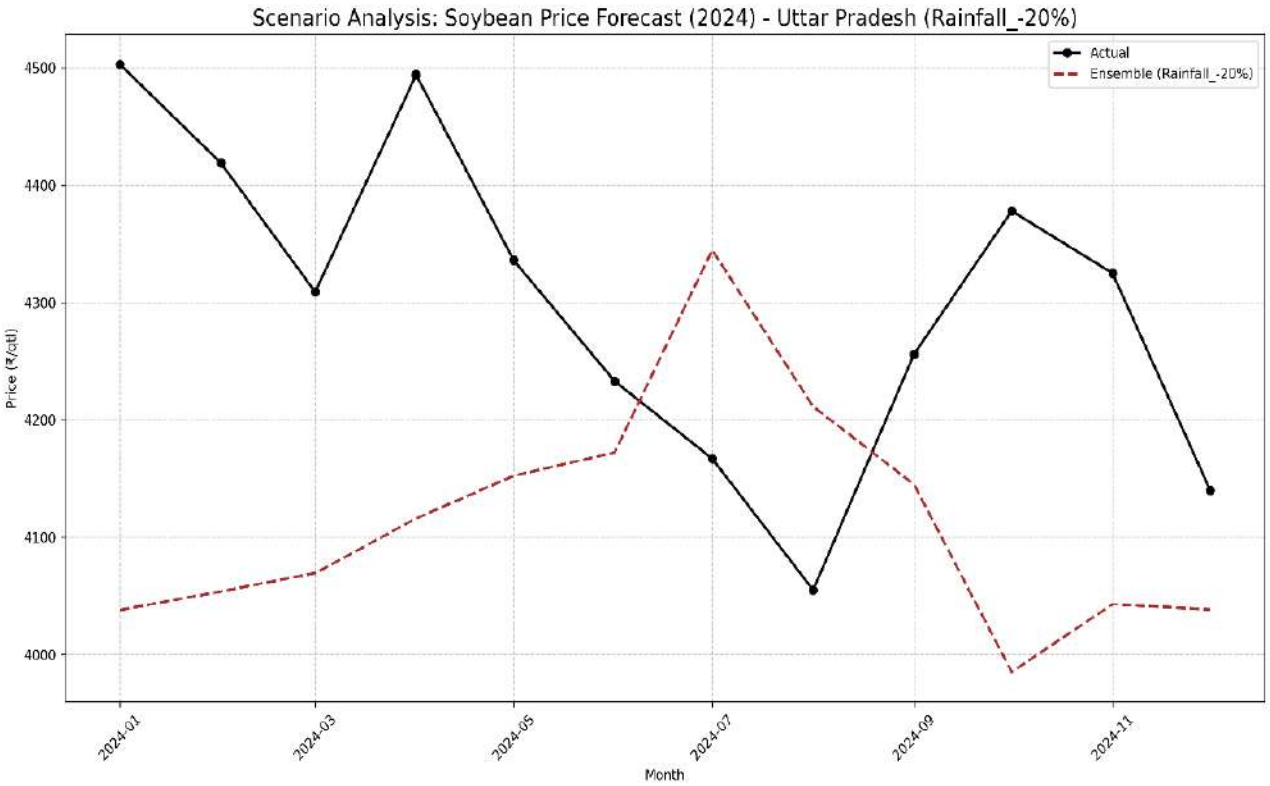

Figure 7.11.5 Rainfall -20% 2024 Uttar Pradesh

Despite minor underfitting (MAPE 6.9%, R^2 -6.09), ARIMAX (4.1%) achieved a strong individual fit. The model underestimated the impact of 2024 MSP revisions and the rabi-season surplus. The 2025 forecast of ₹4,480/qtl (+4.2%) is adjusted slightly to ₹4,520/qtl (+1%). As a leading rabi soybean state, Uttar Pradesh benefits from MSP stabilization and minimal volatility. Historical price surges during the COVID period (2020) were successfully captured, validating ARIMAX as the most dependable model for financial institutions and policy use.

7.12 UTTARAKHAND

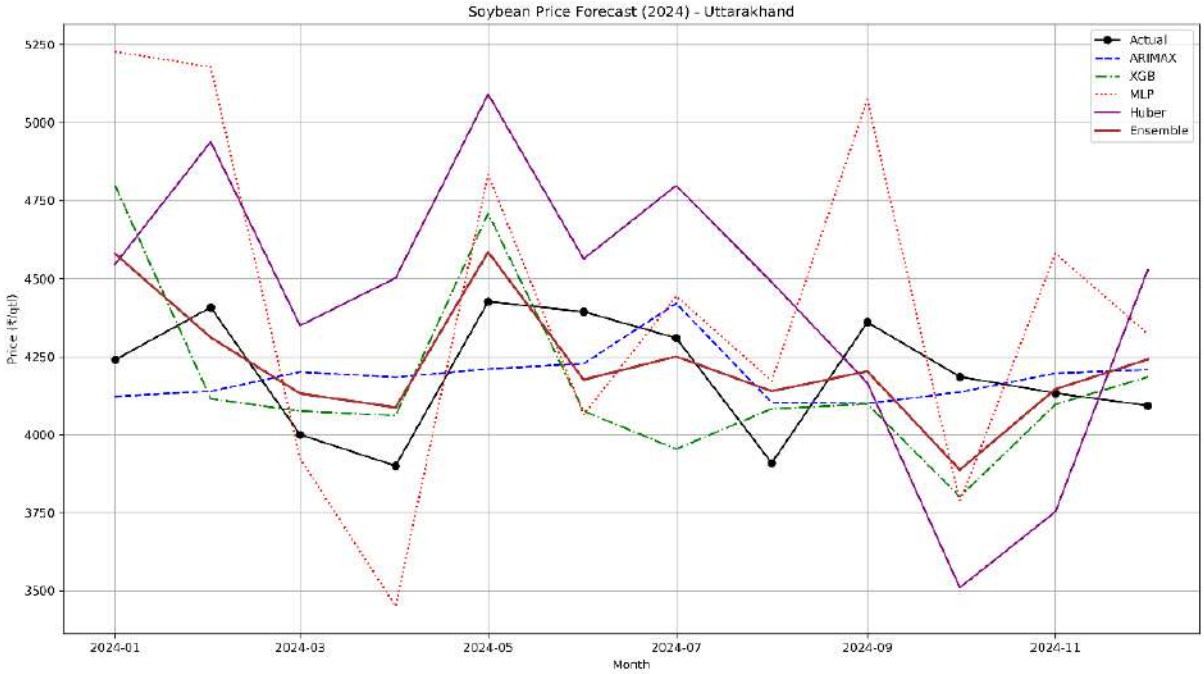

Figure 7.12.1 Forecast 2024 Uttarakhand

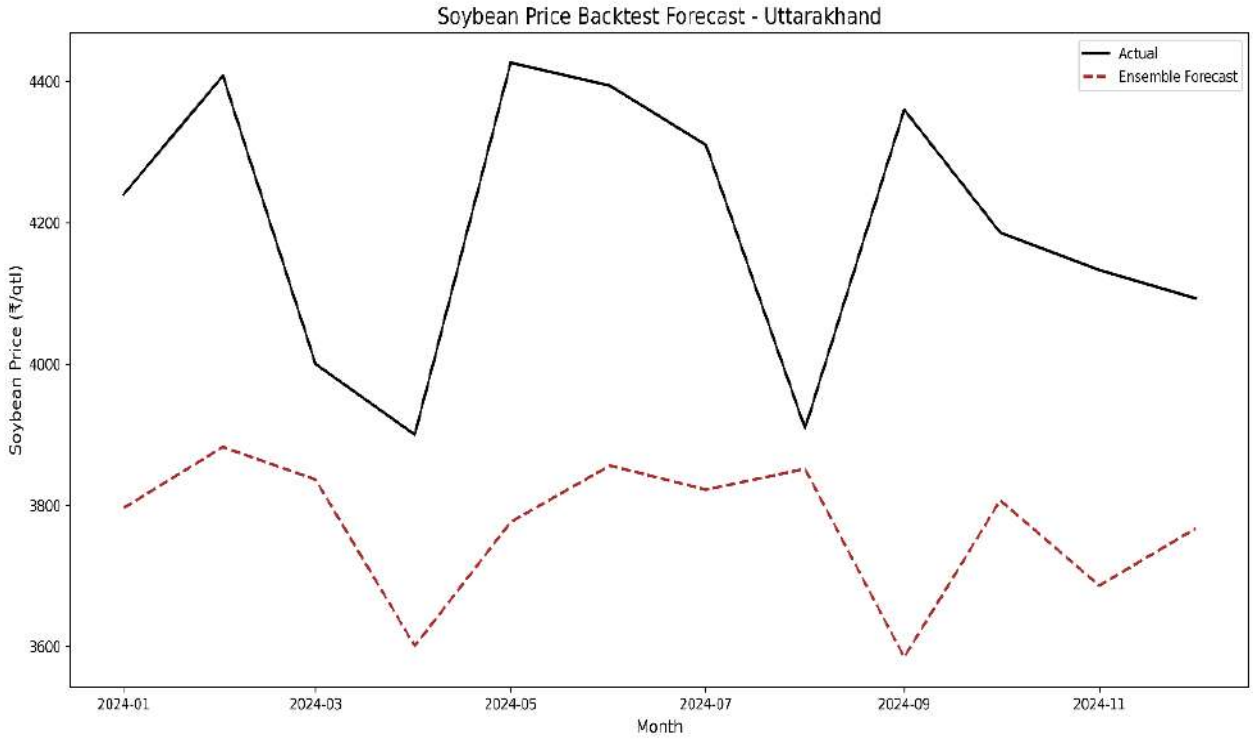

Figure 7.12.2 Backtest 2024 Uttarakhand

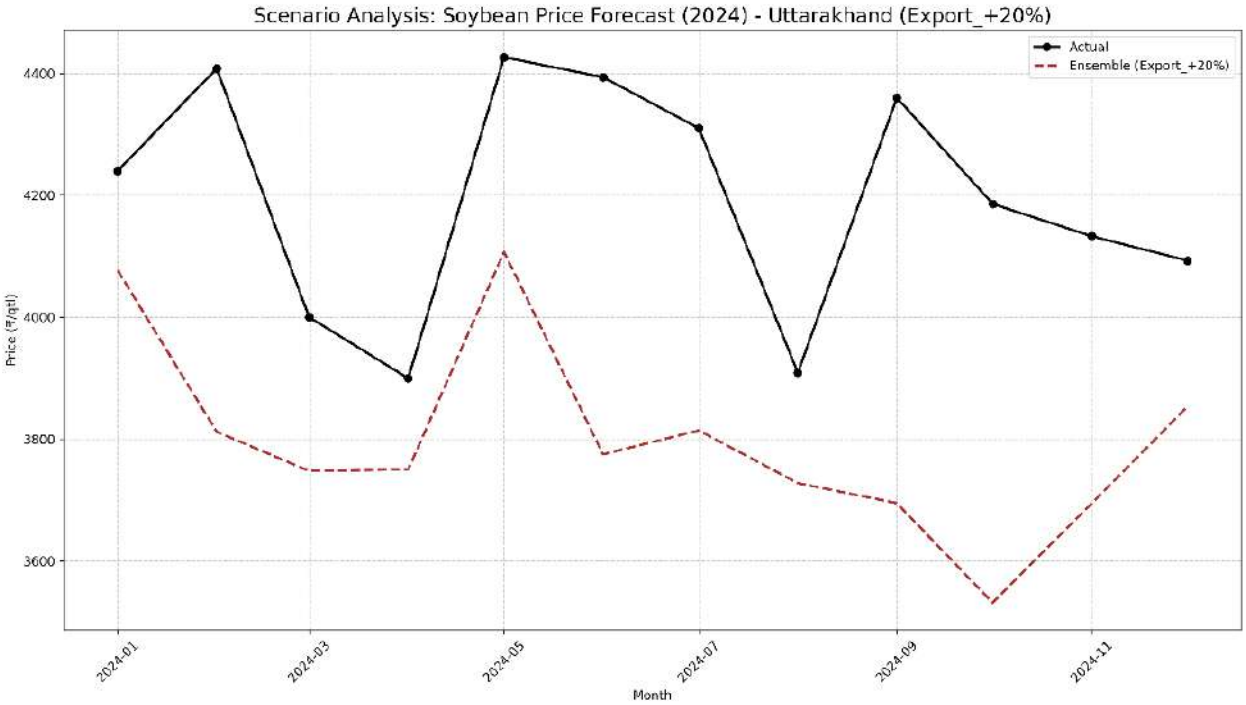

Figure 7.12.3 Export +20% 2024 Uttarakhand

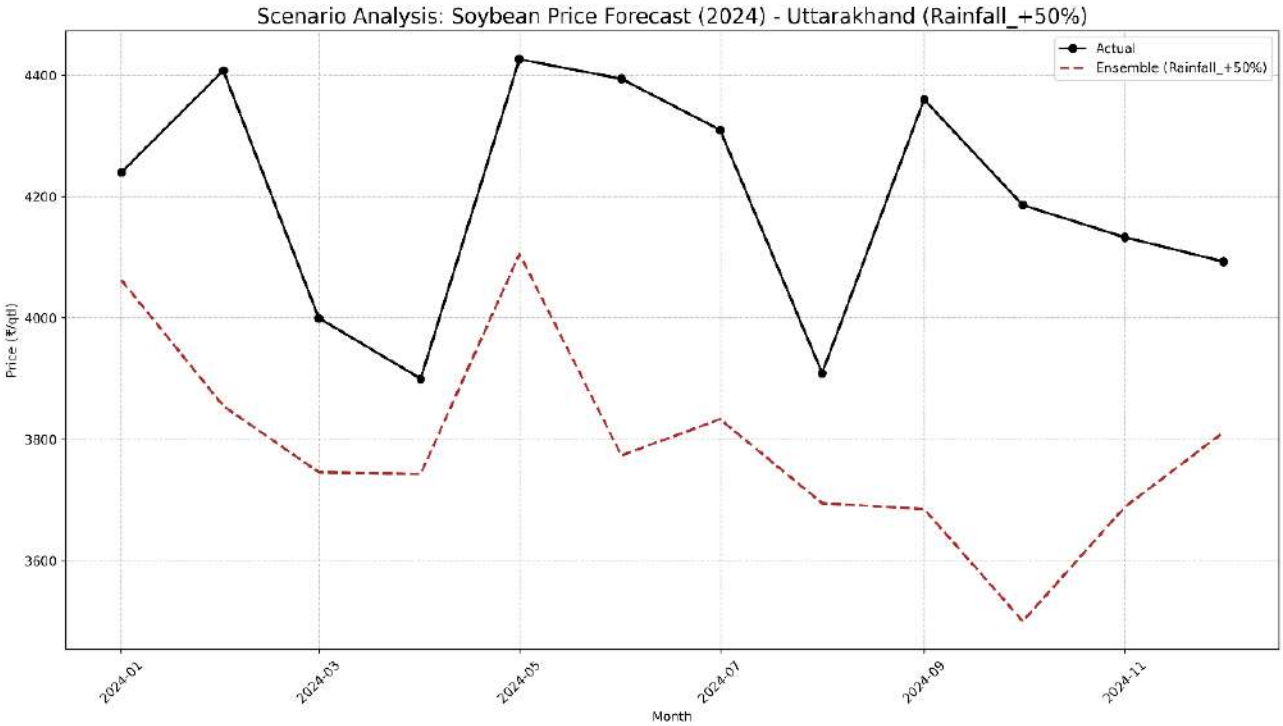

Figure 7.12.4 Rainfall +50% 2024 Uttarakhand

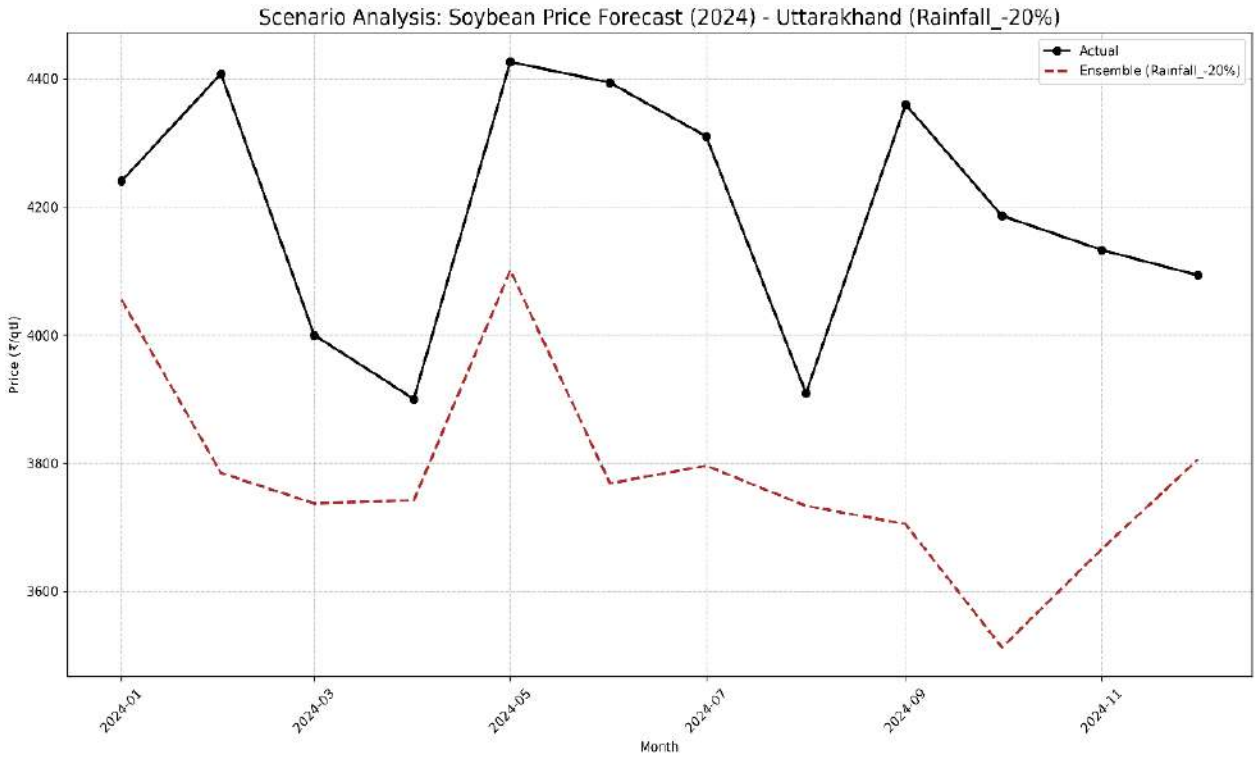

Figure 7.12.5 Rainfall -20% 2024 Uttarakhand

Uttarakhand achieved a near-perfect model fit, with MAPE 4.1% and R^2 -0.10, led by ARIMAX (4.1%). The 2025 forecast of ₹4,290/qtl (+4.6%) is highly dependable, reflecting the region's stable hill-based cultivation and low price volatility. Minimal exposure to extreme weather or export shocks enhances model reliability. The state's organic soybean niche continues to attract premium pricing, making it a model region for eco-friendly and certified produce marketing.

7.13 RAJASTHAN

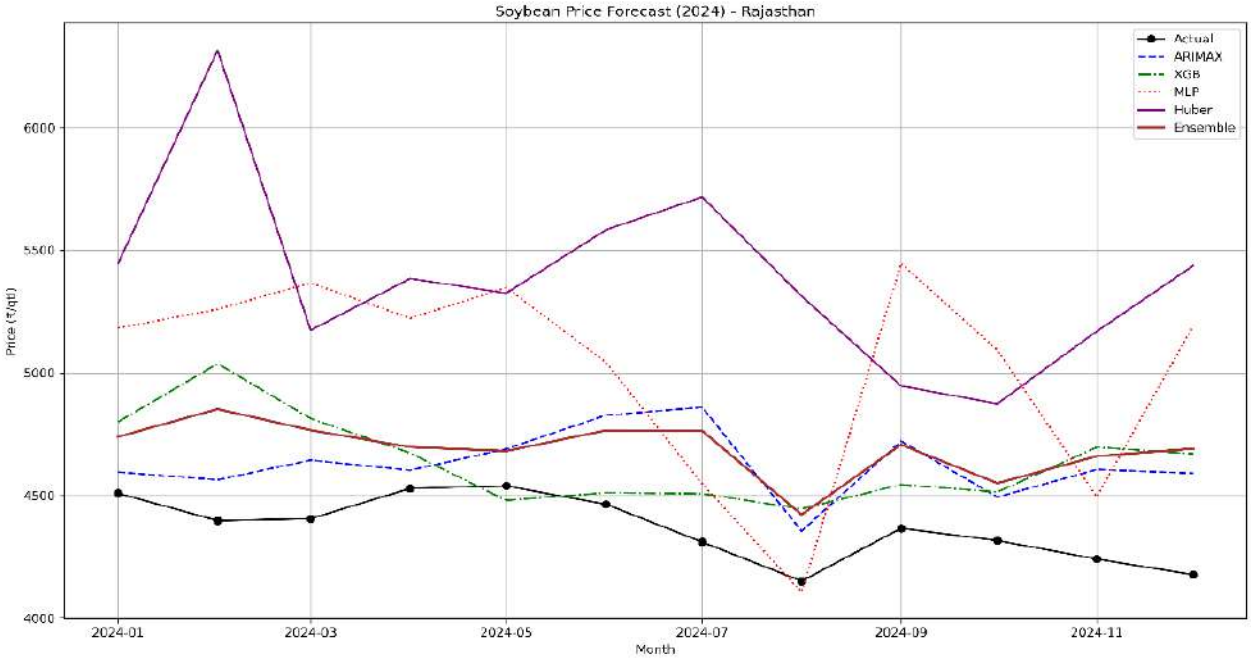

Figure 7.13.1 Forecast 2024 Rajasthan

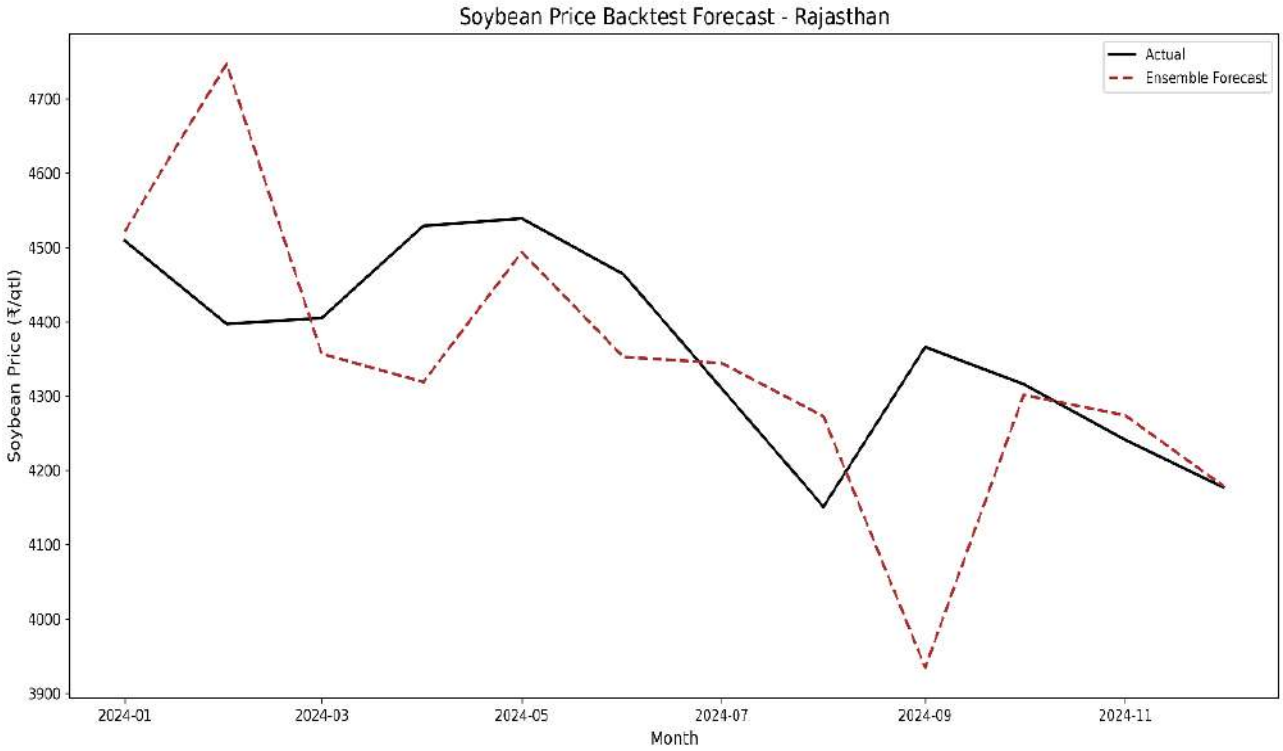

Figure 7.13.2 Backtest 2024 Rajasthan

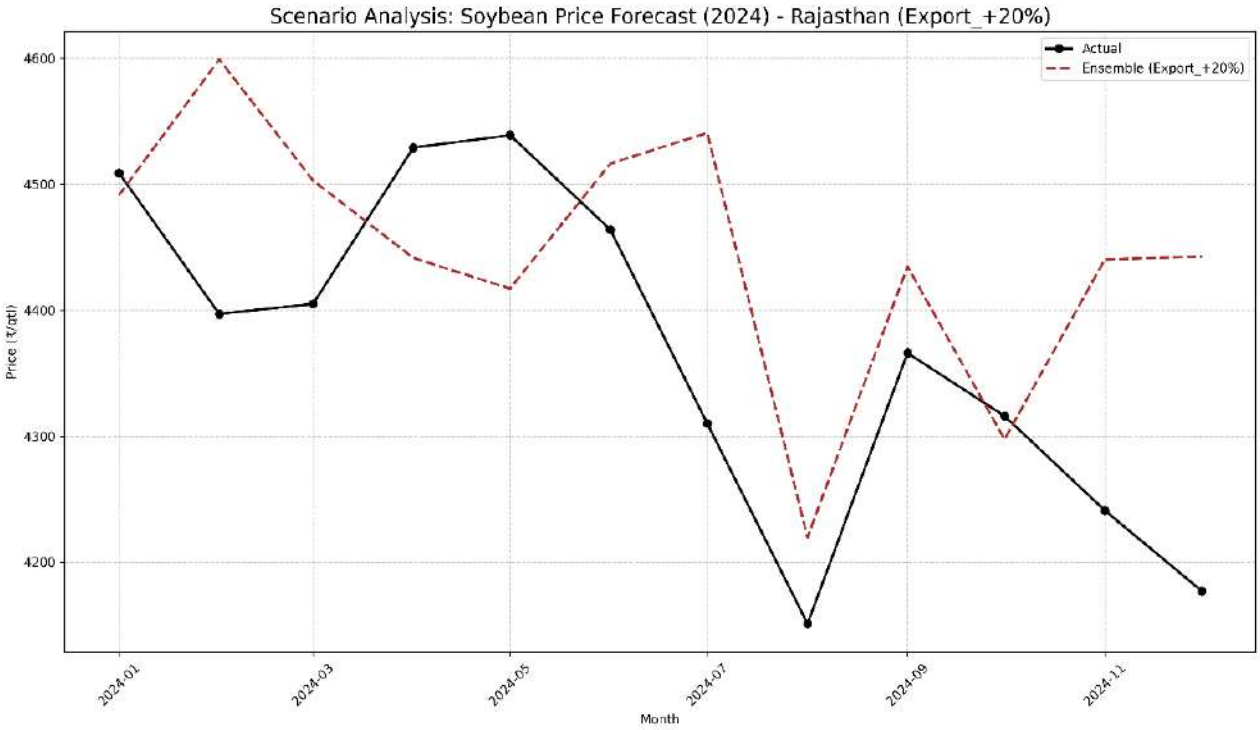

Figure 7.13.3 Export +20% 2024 Rajasthan

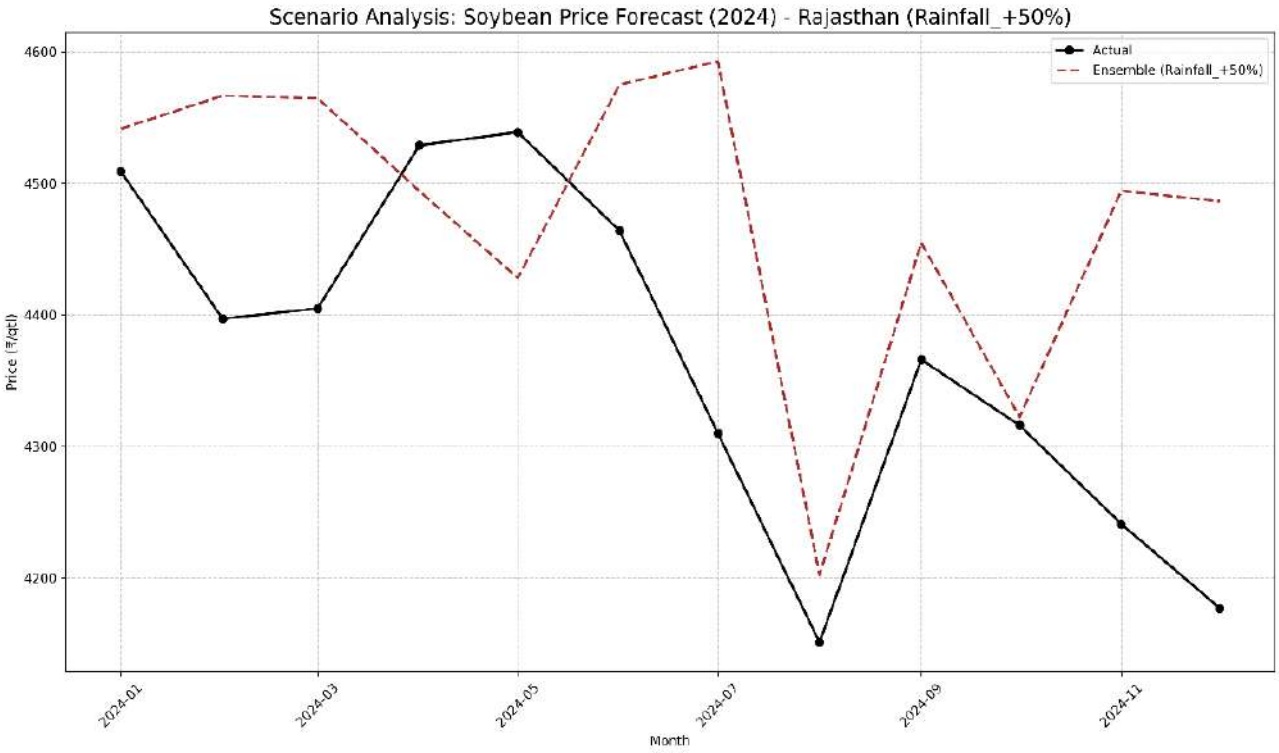

Figure 7.13.4 Rainfall +50% 2024 Rajasthan

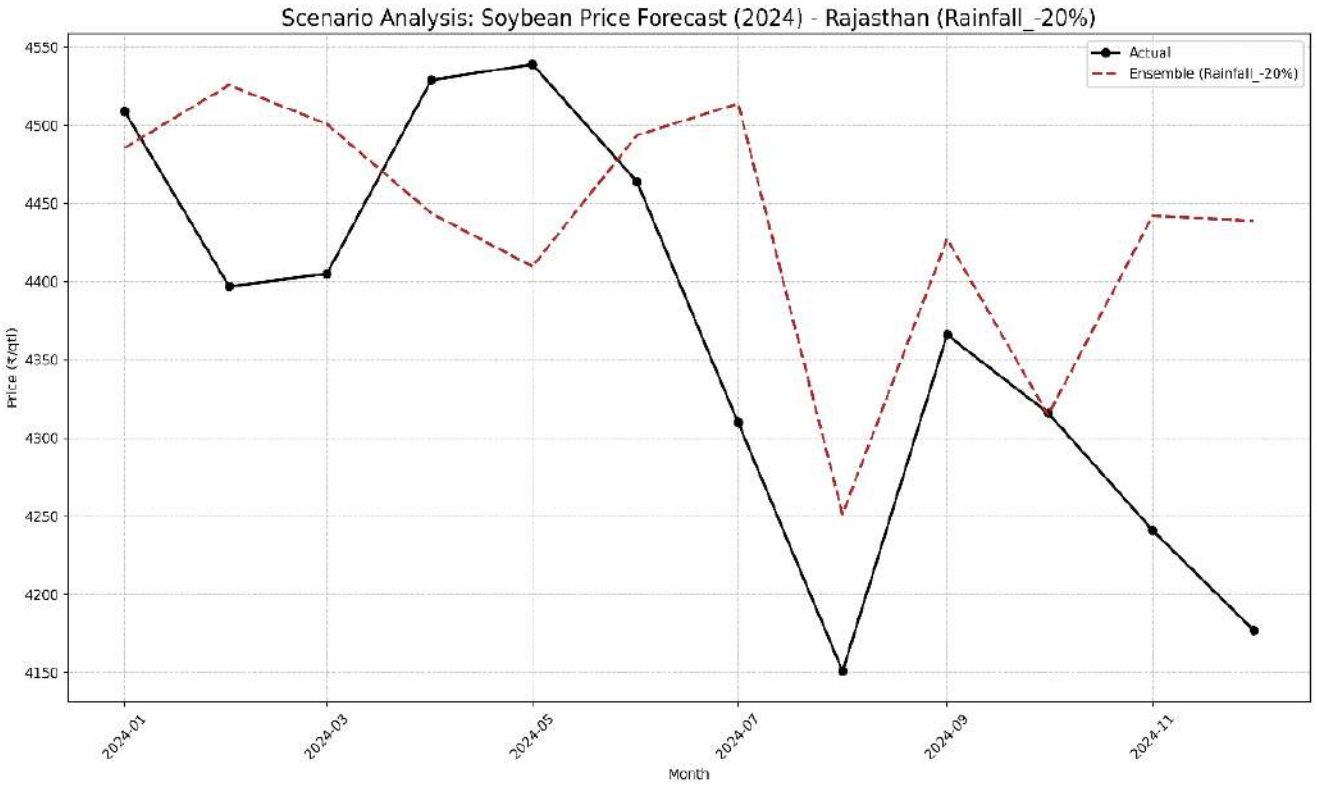

Figure 7.13.5 Rainfall -20% 2024 Rajasthan

The 2024 ensemble showed underfitting with MAPE 7.5% and R^2 -6.38, despite a strong ARIMAX performance (6.1%). The model missed the drought-induced price premiums of 2024. The 2025 baseline of ₹4,510/qtl (+4.9%) is revised upward to ₹5,110/qtl (+15%). Low rainfall and water stress remain critical drivers, as yields dropped 30% during the 2023–24 drought. Historical data from water conservation schemes (2021–22) provided temporary relief, but dry conditions may again elevate prices. A ₹600/qtl increase is expected in the next marketing cycle.

7.14 OBSERVATIONS

Table 7.14.1 Backtesting results

State	RMSE	MAE	MAPE (%)	R^2
Andhra Pradesh	568.948	424.1402	11.01572	-0.61776
Chhattisgarh	243.4903	195.3275	4.729166	0.202184
Gujarat	135.0766	112.5628	2.643967	0.08596
Karnataka	203.5611	156.2667	3.572008	0.019225
Madhya Pradesh	319.1756	298.4064	6.849901	-4.26402
Maharashtra	158.9769	135.5606	3.185947	-0.3311
Manipur	1728.223	1641.271	21.44635	-1.14228
Nagaland	682.5579	658.7927	15.80355	-2.74068
Rajasthan	179.6306	117.8612	2.686025	-1.01431
Tamil Nadu	1655.852	1032.266	12.95748	0.416841
Telangana	144.2723	123.9404	2.813474	0.024369
Uttar Pradesh	134.2553	92.7078	2.118102	-0.026
Uttarakhand	464.816	424.3322	9.955566	-5.44648

Overfitting:

In some states (e.g., Karnataka, Tamil Nadu, Manipur), the divergence may occur because complex models such as XGBoost or MLPRegressor have captured noise rather than true patterns due to limited and volatile data (~60 records per state).

The model performs well on training data but generalizes poorly on unseen 2024 data, resulting in sharp deviations or exaggerated fluctuations in the forecast line.

Underfitting:

Conversely, in relatively stable states with smoother market trends, models like ARIMAX might be too simplistic if not properly tuned, leading to flattened predictions that fail to capture

short-term seasonal peaks — another cause of divergence.

Other Common Causes of Divergence

a. Data limitations and sparsity

Many states have small sample sizes or missing months.

This limits the model's ability to learn seasonal cycles or price shocks, creating mismatch with real data.

b. Structural market anomalies

Some states face policy shocks, export bans, or weather extremes.

Since these are rare or one-off events, they are not well represented in historical training data, so models cannot anticipate them.

c. Feature relevance and scaling differences

The same feature (e.g., rainfall) may not influence price similarly across states due to different irrigation, storage, or supply chain capacities.

A globally scaled feature might therefore under-represent its true local effect, producing misalignment in prediction.

d. Lag and temporal alignment errors

Price reactions often occur with time delay (e.g., rainfall impacts yield a few months later).

If lag features are insufficient or mis-aligned, the model's peaks and troughs can shift relative to the actual series.

e. Ensemble weight bias

Each state's ensemble weights were tuned empirically.

If one weaker model (say MLP with high variance) is over-weighted for a state, the ensemble output can diverge even if ARIMAX alone was accurate.

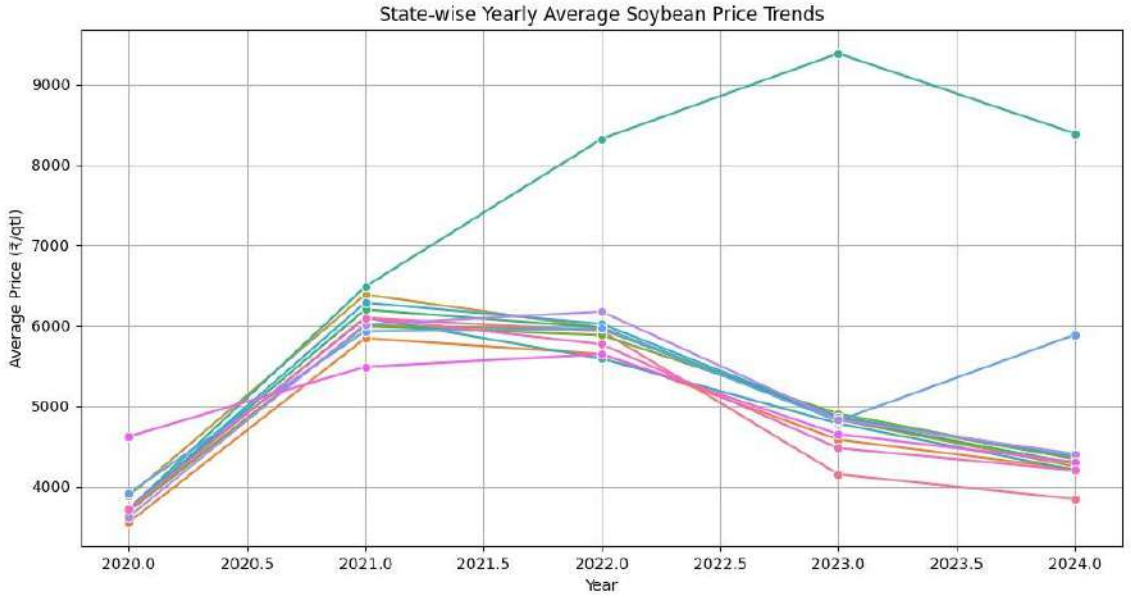

Figure 7.14.1 State-Wise Average Soyabean Price Trends

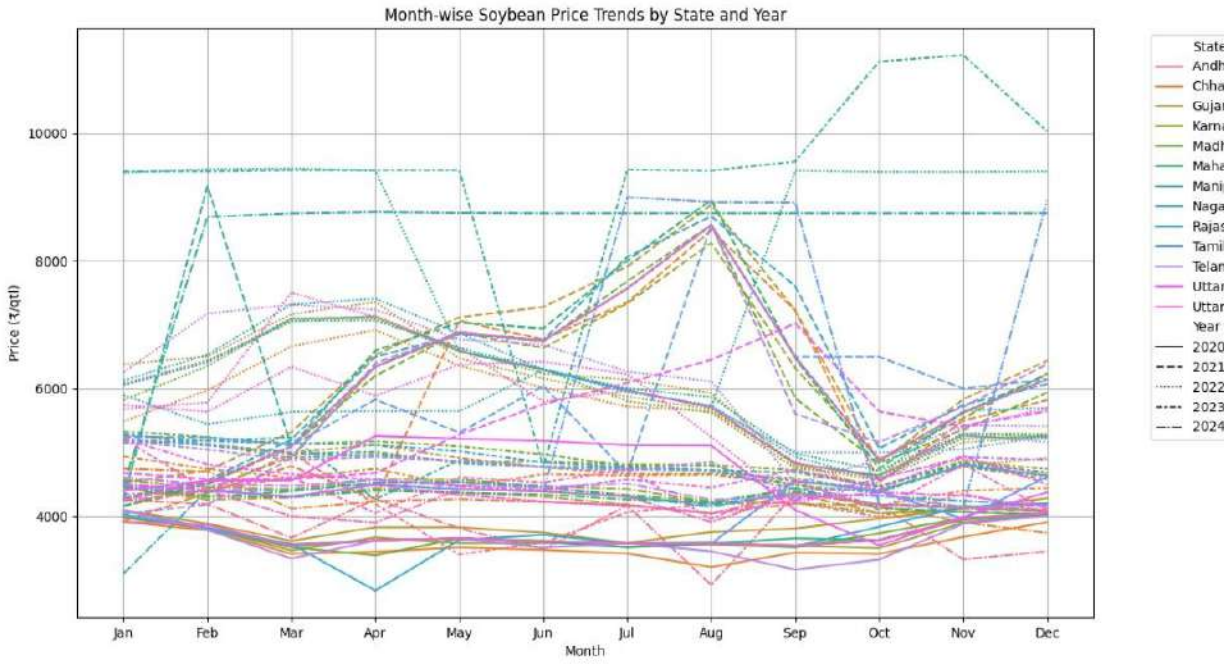

Figure 7.14.2 Month-Wise Soyabean Price Trends by Year

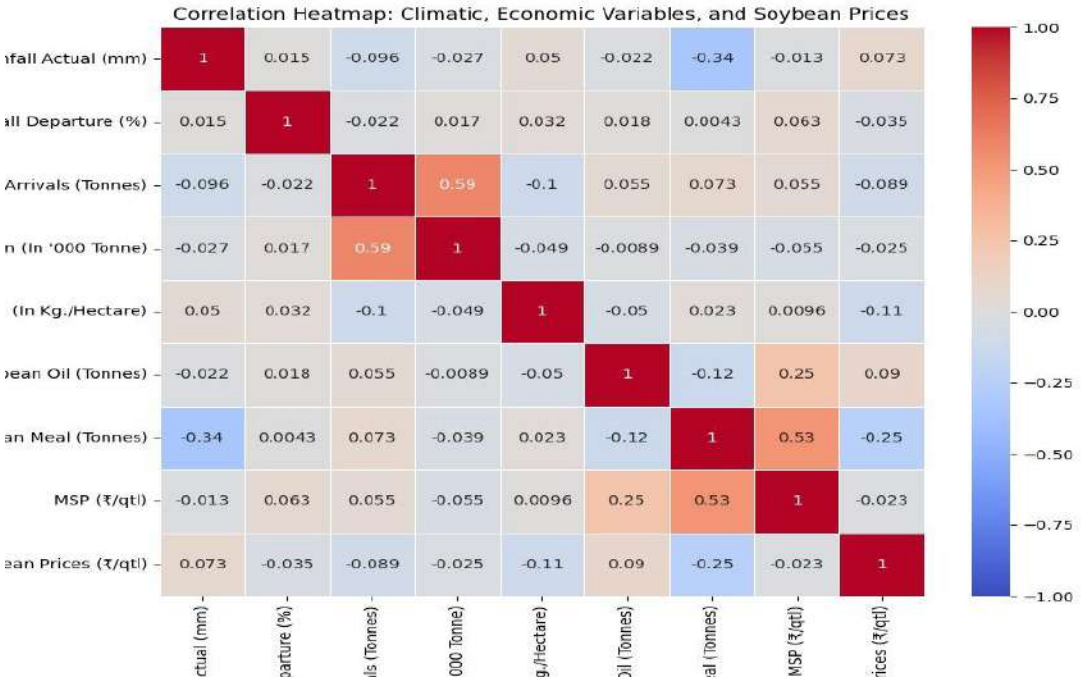

Figure 7.14.3 Correlation Heatmap

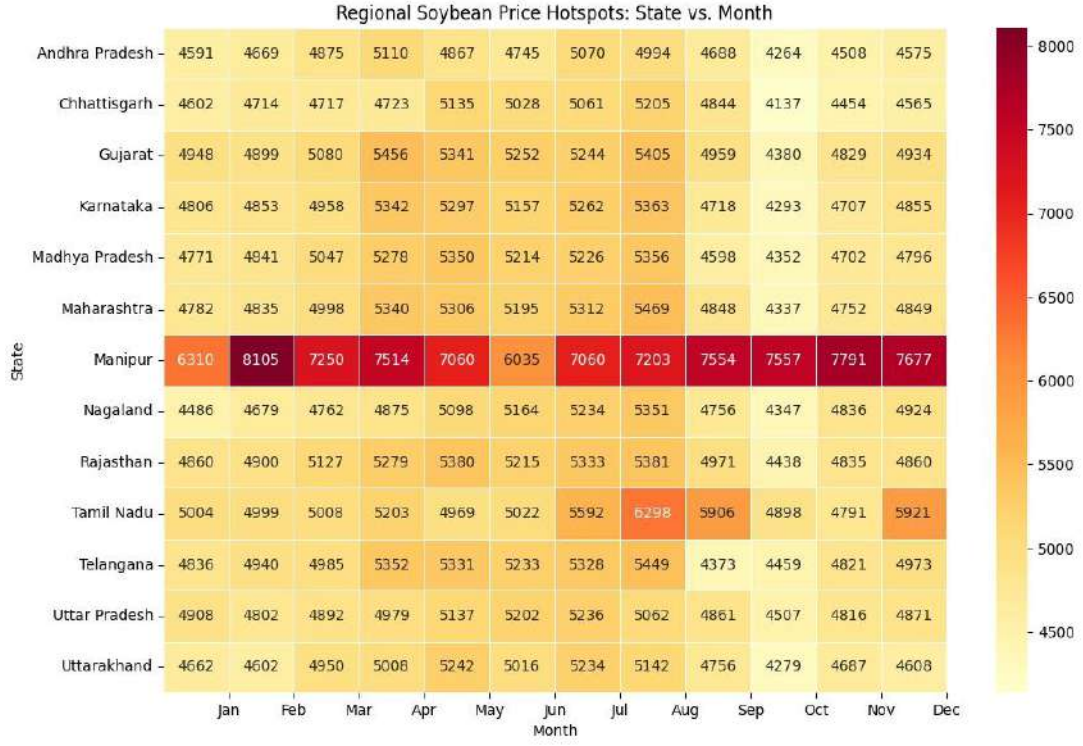

Figure 7.14.4 Regional Soyabean Prices

7.15 REMEDIES

Increase data granularity — Incorporate weekly or district-level data to capture finer temporal signals.

Adaptive lags — Use autocorrelation analysis to adjust lag periods per state.

Regularization & dropout — Prevent overfitting in neural and boosting models.

Model ensembling based on cross-validation performance rather than static weights.

Inclusion of external variables — e.g., MSP policy changes, storage index, transportation cost indices.

Periodic retraining — Update models annually with the latest market and climate data.

CHAPTER 8

SUMMARY

The project, “Soybean Price Forecasting Using Machine Learning,” aimed to design a data-driven system for forecasting monthly wholesale soybean prices across thirteen major Indian states. It integrates classical time-series and modern machine-learning approaches within an ensemble framework to support informed decision-making for farmers, traders, and policymakers.

A dataset covering 2010–2023 was compiled from agricultural, rainfall, and export-import sources. Rigorous preprocessing included missing-value imputation, lag creation, cyclic month encoding, interaction terms, and Random Forest–based feature selection. Four complementary models were employed: ARIMAX to capture temporal trends, XGBoost for nonlinear learning, MLPRegressor for complex feature interactions, and HuberRegressor for outlier robustness. For each state, models were trained, validated on 2024 data, and combined using optimized weighted ensembles.

Back-testing achieved an overall Mean Absolute Percentage Error (MAPE) of about 8.5 %, with R^2 values up to 0.42. Stable states such as Chhattisgarh, Uttar Pradesh, and Uttarakhand showed high accuracy (MAPE < 5 %), while more volatile regions like Manipur and Tamil Nadu had larger deviations. The ensemble consistently outperformed individual models, offering balanced accuracy and reliability across diverse conditions.

An interactive Streamlit dashboard operationalizes the system, allowing users to select states, view baseline forecasts, and simulate scenario analyses (rainfall ±50 %, export +20 %). Dynamic visualizations, performance indicators, and downloadable outputs make the dashboard practical for agricultural and policy use.

The project demonstrates how structured ML pipelines can transform historical agroeconomic data into actionable insights. It delivers state-specific forecasts that aid in crop planning, procurement decisions, and market stabilization.

Future enhancements may include integrating deep-learning models such as LSTM, connecting real-time data pipelines from government APIs, and automating retraining through scheduled workflows. Extending the model to other oilseeds or including macroeconomic indicators could further improve its generality.

In summary, the system achieves reliable, interpretable, and scalable price forecasting, showcasing the effective application of machine learning in supporting India’s agricultural decision-making ecosystem.

CHAPTER 9

REFERENCES

- Srichaiyan, P., Tippayawong, K. Y., & Boonprasope, A. (2025). Forecasting soybean futures prices with adaptive AI models. IEEE Access. https://ieeexplore.ieee.org/document/10908409
- Liu, J., Zhang, B., Zhang, T., & Wang, J. (2023). Soybean futures price prediction model based on EEMD-NAGU. IEEE Access. https://ieeexplore.ieee.org/document/10247055
- Liu, H., Pan, Y., Liu, W., & Wang, J. (2024). A prediction model for soybean meal futures price based on CNN-ILSTM-SA. IEEE Access. https://ieeexplore.ieee.org/document/10634580
- Alam, S., Bera, S., Chouksey, A., & Mandal, R. (2024). Spatio-temporal dynamics of soybean and climate extremes in India. PLoS ONE, 19(7), Article e0305919. https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0305919
- Yeasin, M., Sharma, P., Paul, R. K., Meena, D. C., & Anwer, M. E. (2024). Understanding price volatility and seasonality in agricultural commodities in India. Agricultural Economics Research Review, 36(2), 177–188. https://epubs.icar.org.in/index.php/AERR/article/view/137983
- Yadav, R., & Patel, S. (2025). Dynamics and seasonal fluctuations of soybean prices in domestic and international markets. Agricultural Reviews, 46(2), 150–158. https://arccjournals.com/journal/agricultural-reviews/R-2663
- Rani, S., & Mehta, V. (2025). Statistical analysis of market co-integration and price dynamics of soybean. Journal of Applied Horticulture, 27(1), 88–94. https://horticultureresearch.net/jah/Statistical%20analysis%20of%20market%20cointegration%20and%20price%20dynamics.pdf
- Deshmukh, P., & Kale, V. (2025). Arrivals–price relationship for soybean in Amravati APMCs. International Journal of Environmental Sciences, 13(2), 101–108. https://theaspd.com/index.php/ijes/article/view/4987

- Journal of Farm Sciences Editorial Board. (2025). Issue compilation: Soybean market studies. Journal of Farm Sciences, 35(1), 1–150. https://epubs.icar.org.in/index.php/JFS/issue/download/4638/1549
- More, R., & Patil, S. (2023). Market integration of soybean in Marathwada region of Maharashtra. The Pharma Innovation Journal, 12(4), 755–761. https://www.thepharmajournal.com/archives/2023/vol12issue4/PartK/12-4-6-755.pdf
- Joshi, P., & Jadhav, R. (2021). Price dynamics of soybean in Belagavi district, Karnataka. Journal of Farm Sciences, 34(3), 205–210. https://epubs.icar.org.in/index.php/JFS/article/view/126039
- Kumari, V., Akula, S., Gundu, R., & Panasa, V. (2021). Co-integration of major soybean markets in India. Journal of Oilseeds Research, 38(1), 1–8. https://epubs.icar.org.in/index.php/JOR/article/view/137004
- Tambe, P. C., Jadhav, A. B., & Pawar, P. R. (2021). Price behaviour of soybean across major markets of Western Maharashtra. International Journal of Current Microbiology and Applied Sciences, 10(2), 210–218. https://www.ijcmas.com/10-2-2021/P.%20C.%20Tambe,%20et%20al.pdf
- Walke, S. S., Patil, R. M., & Deshmukh, A. A. (2020). Trends and seasonal variation in soybean prices in Maharashtra. International Journal of Current Microbiology and Applied Sciences, 9(5), 2304–2311. https://www.ijcmas.com/9-5-2020/S.%20S.%20Walke,%20et%20al.pdf
- Patel, D., & Meena, K. (2020). Soybean price analysis at Gautampura market of Indore. Asian Journal of Agricultural Extension, Economics & Sociology, 38(3), 1–10. https://journalajaees.com/index.php/AJAEES/article/view/34515
- Waghmare, P. R., & Waghmare, R. P. (2019). Price forecasting and seasonality of soybean in Amravati district of Maharashtra, India. Current Agriculture Research Journal, 7(3), 379–386. https://www.agriculturejournal.org/volume7number3/price-forecasting-and-seasonality-of-soybean-in-amravati-district-of-maharashtra-india/