

PREDICTIVE ANALYTICS IN NEPALESE AGRICULTURE

Sandesh Kumar Kafle¹, František Dařena¹

¹Department of Informatics, Faculty of Business and Economics, Mendel University in Brno, Zemědělská 1, 613 00 Brno, Czech Republic

DOI: https://doi.org/10.11118/10.11118/ecos2026_010

ABSTRACT

This study addresses climate vulnerability and the lack of data-driven tools in Nepalese agriculture by building predictive crop-yield models for 14 districts in the Koshi Province. Applying a data-science pipeline to historical agricultural records and NASA POWER agroclimatic data spanning 11 crops over 11 fiscal years (2011/12–2021/22), the research evaluates Random Forest and Gradient Boosting Regressors across five distinct weather-feature datasets. The results demonstrate that the Random Forest Regressor consistently outperforms Gradient Boosting, explaining 74% to 84% of yield variance. The highest predictive accuracy was achieved using engineered growing-season averages (Baseline $R^2 = 0.83$, Post-tuning $R^2 = 0.84$), confirming that crop-specific seasonal summaries are the most informative features. Temperature emerged as the strongest single predictor of crop yield ($R^2 = 0.82$), while standalone precipitation exhibited the weakest predictive power ($R^2 = 0.74$). These findings offer an empirical decision-support foundation for farmers and policymakers to quantify climate impacts and enhance sustainable agricultural planning in Nepal.

Keywords: predictive analytics, machine learning, crop yield prediction, Nepalese agriculture, weather impact

JEL Code: Q10, C54, C63

1 INTRODUCTION

Agriculture is the backbone of Nepal's economy, employing around 65% of the population and contributing approximately 27% of GDP. The country's diverse agroecological zones and rich biodiversity create significant potential for food and nutritional security; however, the sector faces persistent challenges, including unpredictable weather events, outdated irrigation infrastructure, and a lack of evidence-based decision-making tools.

Predictive analytics—using historical data to forecast future outcomes—offers a path toward more resilient and productive agriculture. Advances in machine learning (ML), the Internet of Things (IoT), and open data platforms now make it feasible to develop yield-prediction models



This article is licensed under CC BY-SA 4.0 Attribution-ShareAlike 4.0 International license. <https://creativecommons.org/licenses/by-sa/4.0/deed.en>

tailored to Nepal's conditions. Technology can thus assist in transforming the global food production system while mitigating its impact on the climate and environment (Lim, 2022).

Nepalese farmers lack the knowledge and tools to leverage past weather and yield data for decision-making. Existing crop-yield prediction models do not adequately capture the specific impact of weather variability on Nepalese agriculture. This study addresses the gap by building and evaluating predictive models using historical agroclimatic and agricultural data.

The objective is to identify problems in Nepal's agricultural sector that can be addressed through advanced analysis of historical data and to provide a decision-support foundation for farmers and policymakers. The study specifically pursues the following questions:

- *Can a predictive model capture the effect of weather factors (temperature, humidity, precipitation) on crop yields?*
- *Which weather parameter exerts the greatest influence on agricultural yields?*
- *Is there a significant relationship between growing-season weather conditions and crop yield?*
- *What actionable recommendations can improve the future potential of Nepalese agriculture?*

The study focuses on 14 districts of Koshi Province, Nepal, covering 11 crops (6 cereals, 3 cash crops, 2 pulses) over 11 fiscal years (2011/12–2021/22). Key limitations include sampling bias toward Koshi Province, data quality constraints, reliance on publicly available weather records, exclusion of soil and socioeconomic variables, and the difficulty of modelling unpredictable extreme weather events.

2 CURRENT STATE OF RESEARCH

2.1 Predictive Analytics in Agriculture

Predictive analytics has gained widespread adoption in agriculture globally, enabling more accurate yield forecasting, resource optimization, and risk management. Machine learning models—particularly ensemble methods such as Random Forest and Gradient Boosting—have demonstrated strong performance in non-linear regression tasks involving weather-crop relationships. Studies across East Africa (Aworka *et al.*, 2022), Europe (Paudel *et al.*, 2022), India (Sabu *et al.*, 2020), and Nepal (Fernández-Beltrán *et al.*, 2021) confirm the adaptability of these approaches across diverse agricultural contexts.

2.2 Weather and Crop Yield in Nepal

Nepal's agriculture is highly monsoon-dependent. Rice, the dominant cereal, requires timely and adequate rainfall; wheat is sensitive to heat and water stress during grain-filling. Climate studies confirm that increased temperature variability and erratic rainfall reduce yields for both crops, with effects varying by altitude and growing season. A 2016 CBS survey found that nearly 100% of households observed increased drought over the prior 25 years, and 97.7% noted a rise in disease incidence and irregular rainfall.

Climate change is projected to intensify these pressures through rising temperatures, altered monsoon patterns, and more frequent extreme events—flooding in 2017 alone damaged 10% of total paddy area (FAO, 2023).

2.3 Key Models and Evaluation Approaches

The literature documents several modelling frameworks relevant to Nepal: CNN-based models using Sentinel-2 satellite imagery for rice yield (Fernández-Beltrán *et al.*, 2021), ARIMA and LSTM for price/yield time-series forecasting, Random Forest and Support Vector Regression for multi-feature yield prediction (Ramu, K. & Priyadarsini, K., 2021); and interaction regression models that can outperform standard ML algorithms (Ansarifar *et al.*, 2021). Common evaluation metrics include R^2 (explanatory power), Mean Squared Error (MSE), and Mean Absolute Error (MAE), validated primarily via k-fold cross-validation (Oikonomidis *et al.*, 2022; Burdett & Wellen, 2022).

2.4 Research Gap

Despite a growing global body of work, predictive models specifically calibrated to Nepal's diverse agroecological conditions, fiscal-year data structure, and smallholder context remain scarce. This study addresses that gap by building district-level, multi-crop models from publicly available Nepalese data.

3 METHODOLOGY AND DATA

3.1 Research Approach and Design

The study combines quantitative modelling with qualitative contextualization. Quantitative analysis is central: numerical weather and yield data are used to train and evaluate regression-based ML models predicting yield. Qualitative insights from the literature inform feature selection, model interpretation, and recommendations. The workflow follows a standard data-science pipeline: data collection → preprocessing → feature engineering → model training → evaluation → tuning.

3.2 Data Sources

Two primary data sources were used: (1) annual crop yield records for 11 crops across 14 Koshi districts, obtained from the Ministry of Agriculture and Livestock Development (MOALD); and (2) monthly weather data—surface temperature (TS), 2-meter temperature (T2M), specific humidity (QV2M), relative humidity (RH2M), and corrected precipitation (PRECTOTCORR)—obtained from NASA POWER. The combined dataset spans 11 years, 14 districts, and 11 crops, and contains the following features: Year (Fiscal year), District (Name of the district), Crop (Name of the crop), Yield (Metric ton production over the area in Hectare), TS (Earth Skin Temperature/Surface Temperature (°C)), T2M (Temperature at 2 Meters (°C)), QV2M (Specific Humidity at 2 Meters (g/kg)), RH2M (Relative Humidity at 2 Meters (%)), PRECTOTCORR (Precipitation Corrected (mm/day)).

3.3 Data Preprocessing

After merging the two sources by fiscal year and district, the dataset was cleaned by removing duplicate records, replacing missing numerical values with column means, and dropping remaining null entries. All features were then normalised using the Yeo-Johnson power transformation (to approximate Gaussian distributions) and standardised with StandardScaler (zero mean, unit variance).

3.4 Feature Engineering

Raw monthly weather data produced a large feature space detrimental to model performance. Following Eck *et al.* (2020), crop-specific growing-season averages were calculated for each weather variable, reducing the feature set to five engineered variables: GSavgTS (Growing season average for Temperature Surface), GSavgT2M (Growing season average for Temperature at 2 meters), GSavgQV2M (Growing season average for Specific Humidity at 2 meters), GSavgRH2M (Growing season average for Relative Humidity at 2 meters), and GSavgPREC (Growing season average for Precipitation Corrected). Five distinct datasets were constructed to test the individual and combined predictive power of temperature, humidity, precipitation, and growing-season weather: Dataset 1 (Temperature + Humidity + Precipitation + Crop + Yield), Dataset 2 (Humidity + Crop + Yield), Dataset 3 (Temperature + Crop + Yield), Dataset 4 (Precipitation + Crop + Yield), Dataset 5 (Growing Season Averages + Crop + Yield).

3.5 Models

Two ensemble regression models were selected after initial experiments: Random Forest Regressor and Gradient Boosting Regressor (both from scikit-learn). Each model was trained on 75% of the data and tested on the remaining 25%. Key hyperparameters include `n_estimators`, `max_depth`, `learning_rate` (GBM only), `min_samples_split`, `min_samples_leaf`, and `max_features`.

3.6 Evaluation and Tuning

Model performance was measured using R^2 , MSE, and MAE on the held-out test set. Generalisation was assessed with 5-fold cross-validation (KFold). Hyperparameters were optimised via two approaches: RandomizedSearchCV (faster, samples random combinations) and GridSearchCV (exhaustive but computationally expensive). Both tuning results are reported and compared against baseline models.

4 RESULTS AND DISCUSSION

4.1 Baseline Model Performance

Table 1 summarises the performance of the two base models across all five datasets. Random Forest consistently outperformed Gradient Boosting in R^2 , MSE, and MAE. The highest baseline performance was achieved on Dataset 5 (Growing Season Averages) with Random Forest ($R^2 = 0.83$,

Tab. 1 Base Model Performance

Dataset	RF R^2	RF MSE	RF MAE	GB R^2	GB MSE	GB MAE
1: All weather	0.80	29.09	1.62	0.77	32.66	2.14
2: Humidity	0.79	29.83	1.70	0.79	30.93	2.11
3: Temperature	0.82	25.62	1.60	0.78	31.09	2.10
4: Precipitation	0.74	37.66	2.02	0.74	37.31	2.21
5: Growing Season Avg	0.83	25.03	1.76	0.81	27.80	2.03

MSE = 25.03) and Dataset 3 (Temperature only) also performing strongly ($R^2 = 0.82$, MSE = 25.62). Precipitation alone (Dataset 4) yielded the weakest results for both models ($R^2 = 0.74$).

4.2 Cross-Validation Results

5-fold cross-validation confirmed the robustness of both models. Average R^2 values ranged from 0.78–0.80 for both models across all datasets, with standard deviations of MSE between 8.99 and 11.09—indicating acceptable and consistent variance. The two models performed nearly identically in cross-validation, suggesting neither is prone to overfitting on this dataset.

4.3 Model Tuning Results

After hyperparameter optimization, performance improved modestly. Table 2 shows the post-tuning results. The best overall result was achieved by Random Forest on Dataset 5 using both RandomizedSearch and GridSearch ($R^2 = 0.84$, MSE = 23.24 and 23.01, respectively), a meaningful improvement over the baseline MSE of 25.03. RandomizedSearch generally yielded comparable or better results than GridSearch at far lower computational cost.

Tab. 2 Performance After Hyperparameter Tuning

Dataset	Model	Rand. R^2	Rand. MSE	Grid R^2	Grid MSE
1: All weather	Random Forest	0.81	28.03	0.80	28.15
1: All weather	Gradient Boosting	0.78	31.50	0.77	32.66
3: Temperature	Random Forest	0.83	25.06	0.83	24.60
5: Growing Season	Random Forest	0.84	23.24	0.84	23.01
5: Growing Season	Gradient Boosting	0.80	28.63	0.81	28.00

4.4 Key Findings

Several key findings can be identified after detailed analysis

- Temperature is the strongest single predictor of crop yield. Dataset 3 (temperature only) achieved a near-top $R^2 = 0.82$, outperforming humidity and precipitation individually.
- Growing-season averages are the most informative engineered feature. Dataset 5 achieved the best performance across both models and both tuning methods, confirming that crop-specific seasonal weather summaries are more predictive than raw monthly records.
- Precipitation has the weakest standalone predictive power. Dataset 4 consistently recorded the lowest R^2 (~ 0.74) and highest MSE, though it still contributes in combination with other variables.
- Random Forest Regressor is the preferred model. It outperformed Gradient Boosting on all five datasets in both baseline and tuned evaluations, with higher R^2 and lower MSE/MAE in most cases.
- Broad weather feature sets improve accuracy. Combining temperature, humidity, and precipitation (Dataset 1) produced better results than any single variable except growing-season averages, suggesting complementary predictive value.

4.5 Discussion and Recommendations

The findings confirm that weather—particularly temperature and growing-season conditions—is a significant driver of agricultural yield variability in Nepal’s Koshi Province. The Random Forest model’s robustness and its ability to capture non-linear feature interactions make it well-suited for this domain.

Recommendations for practitioners and policymakers include:

- Incorporate comprehensive weather variables in yield models, prioritising temperature and growing-season averages.
- Use `RandomizedSearchCV` for hyperparameter tuning as a cost-effective alternative to `GridSearchCV`.
- Extend the model with additional features (soil properties, fertiliser data, planting and harvest season averages) to improve precision.
- Establish feedback loops with farmers and policymakers to validate predictions and iteratively refine the model.
- Design seasonal climate-resilience policies that specifically account for temperature fluctuations during crop growing seasons.

5 CONCLUSIONS

This study demonstrates that machine learning—in particular, the Random Forest Regressor—can effectively model the relationship between weather conditions and crop yields in Nepal’s Koshi Province, achieving R^2 values up to 0.84 after hyperparameter tuning. Across five datasets representing different combinations of temperature, humidity, and precipitation, the models consistently explained 74–84% of yield variance, confirming the significant impact of weather on agricultural productivity.

Among all weather variables examined, temperature and growing-season average weather emerge as the most influential predictors of crop yield. Precipitation, while important in combination, contributes less when used in isolation. These results align with the broader climate-agriculture literature and provide empirical grounding specific to Nepal’s Koshi region.

The developed framework offers a practical decision-support foundation: it can predict yields based on historical data, quantify how specific weather patterns drive productivity, and guide resource allocation and risk management. With continuous data updates, integration of additional features, and collaboration between data scientists, agronomists, and policymakers, this approach can evolve into a robust tool for sustainable agricultural planning in Nepal.

Future research should broaden the geographic scope beyond Koshi Province, incorporate more crops and non-weather variables (e.g., soil chemistry, fertilizer application), and explore real-time data integration via IoT sensors and satellite imagery to further enhance predictive accuracy and practical utility.

Acknowledgements

This paper was supported by the project CZ.02.1.01/0.0/0.0/16_017/0002334 Research Infrastructure for Young Scientists, which is co-financed from Operational Programme Research, Development and Education.

REFERENCES

- ANSARIFAR, J., WANG, L., ARCHONTOULIS, S. V. 2021. An interaction regression model for crop yield prediction. *Scientific Reports*. 11(1).
- AWORKA, R. *et al.* 2022. Agricultural decision system based on Advanced Machine Learning models for yield prediction: Case of East African countries. *Smart Agricultural Technology*. 2, 100048.
- BURDETT, H., WELLEN, C. 2022. Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. *Precision Agriculture*. 23(5), 1553–1574.
- ECK, M. A. *et al.* 2020. Influence of growing season temperature and precipitation anomalies on crop yield in the southeastern United States. *Agricultural and Forest Meteorology*. 291, 108053.
- FAO. 2023. *Country Brief – Nepal*. Food and Agriculture Organization of the United Nations. <https://www.fao.org/giews/countrybrief/country.jsp?code=NPL>.
- FERNÁNDEZ-BELTRÁN, R., BAIDAR, T., KANG, J., PLA, F. 2021. Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sensing*. 13(7), 1391.
- KUMAR KAFLE, S. 2024. *Predictive Analytics in Nepalese Agriculture*. Diploma Thesis. Brno: Mendel University in Brno.
- LIM, X. 2022. Agriculture Needs Technology for Resilient Food Production. *World Economic Forum*. [Accessed: August 2023] <https://www.weforum.org/agenda/2022/03/how-technology-can-help-address-challenges-in-agriculture/>
- MINISTRY OF AGRICULTURE AND LIVESTOCK DEVELOPMENT (MOALD). [n.d.]. Welcome to Ministry of Agriculture and Livestock Development. *MOALD*. [Accessed: August 2023] <https://moald.gov.np/welcome-to-ministry-of-agriculture-and-livestock-development/>
- NASA. 2024. *NASA POWER. Data Access Viewer*. [Accessed: March 2024] <https://power.larc.nasa.gov/data-access-viewer/>
- OIKONOMIDIS, A., CATAL, C., KASSAHUN, A. 2022. Hybrid deep learning-based models for crop yield prediction. *Applied Artificial Intelligence*. 36(1).
- PAUDEL, D. *et al.* 2022. Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*. 276, 108377.
- RAMU, K., PRIYADARSINI, K. 2021. A Review on Crop Yield Prediction Using Machine Learning Methods. In: *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. pp. 1239-1245.
- SABU, K. M., KUMAR, T. K. M. 2020. Predictive analytics in agriculture: Forecasting prices of arecanuts in Kerala. *Procedia Computer Science*. 171, 699–708.

Contact information

Sandesh Kumar Kafle: e-mail: 89995@node.mendelu.cz

František Dařena: e-mail: darena@mendelu.cz