remote sensing

MDPI

Article

Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning

Md Didarul Islam [1], Liping Di [1,*], Faisal Mueen Qamer [2], Sravan Shrestha [2], Liying Guo [1], Li Lin [1], Timothy J. Mayer [3,4] and Aparna R. Phalke [3,4]

[1] Center for Spatial Information Science and Systems, George Mason University, Fairfax, VA 22030, USA; mislam25@gmu.edu (M.D.I.); lguo2@gmu.edu (L.G.); llin2@gmu.edu (L.L.)
[2] International Center for Integrated Mountain Development, Lalitpur 44700, Nepal; faisal.qamer@icimod.org (F.M.Q.)
[3] Earth System Science Center, The University of Alabama in Huntsville, 320 Sparkman Drive, Huntsville, AL 35805, USA; timothy.j.mayer@nasa.gov (T.J.M.)
[4] SERVIR Science Coordination Office, NASA Marshall Space Flight Center 320 Sparkman Drive, Huntsville, AL 35805, USA
* Correspondence: ldi@gmu.edu; Tel.: +1-703-993-6114

Abstract: This study developed a rapid rice yield estimation workflow and customized yield prediction model by integrating remote sensing and meteorological data with machine learning (ML). Several issues need to be addressed while developing a crop yield estimation model, including data quality issues, data processing issues, selecting a suitable machine learning model that can learn from few available time-series data, and understanding the non-linear relationship between historical crop yield and remote sensing and meteorological factors. This study applied a series of data processing techniques and a customized ML model to improve the accuracy of crop yield estimation at the district level in Nepal. It was found that remote sensing-derived NDVI product alone was not sufficient for accurate estimation of crop yield. After incorporating other meteorological variables into the ML models, estimation accuracy improved dramatically. Along with NDVI, the meteorological variables of rainfall, soil moisture, and evapotranspiration also exhibited a strong association with rice yield. This study also found that stacking multiple tree-based regression models together could achieve better accuracy than benchmark linear regression or standalone ML models. Due to the unique and distinct physio-geographical setting of each district, a variation in estimation accuracy from district to district could be observed. Our data processing and ML model workflow achieved an average of 92% accuracy of yield estimation with RMSE 328.06 kg/ha and MAE 317.21 kg/ha. This methodological workflow can be replicated in other study areas and the results can help the local authorities and stakeholders understand the factors affecting crop yields as well as estimating crop yield before harvesting season to ensure food security and sustainability.

Keywords: precision agriculture; crop yield estimation; crop yield; crop phenology; multisource remotely sensed data

Citation: Islam, M.D.; Di, L.; Qamer, F.M.; Shrestha, S.; Guo, L.; Lin, L.; Mayer, T.J.; Phalke, A.R. Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning. *Remote Sens.* **2023**, *15*, 2374. https://doi.org/10.3390/rs15092374

Academic Editor: Wenjiang Huang

Received: 27 March 2023
Revised: 20 April 2023
Accepted: 29 April 2023
Published: 30 April 2023

1. Introduction

Rice is one of the most consumed staple foods for nearly a half of the world's population. By 2050, the population is projected to grow to 9.8 billion, and to keep up with this growth, there will be a 60% rise in food consumption [1]. In addition, the livelihood of millions of farmers depends directly or indirectly on crop production. Over the past few years, the world's total arable land has been diminishing as a result of rising urbanization, which has an impact on overall production and results in a persistent inability to meet the world's demand for agricultural products [2–4]. To satisfy this anticipated future demand, ending hunger, establishing food security, and promoting sustainable agriculture are expressly listed as the top priorities in the 2030 Agenda for Sustainable Development

Goals (SDG) of the United Nations (UN) [5]. In recent years, a great deal of effort has been initiated to increase rice production with modern technology. However, it is generally considered that rice production is associated with the immediate and dynamic nature of global anthropogenic changes, such as population growth and climate changes, and also technological advancement [6–8]. Extreme natural and manmade events such as drought, flood, and fire/war frequently harm food production [9,10]. It is anticipated that both the rates and patterns of total precipitation amount will keep changing, along with the rise in global temperatures, which is expected to have mild to severe consequences on agriculture around the world [11]. The shortage of water and energy supply in the agriculture sector will further impose constraints on food production [12]. In this regard, more precise crop monitoring, dependable mapping tools, and early forecasting of rice production, before the harvest time, can substantially assist the decision makers in minimizing losses and achieving desired yields [6,7,13,14].

Recently, remote sensing became a popular tool to monitor crop health, growth, measurement, and to determine the optimal time for harvesting and rapid near real-time crop yield estimation with minimal cost [6,7,13–18]. Remote sensing-based techniques have already been successfully applied for mapping rice-cultivated areas and demonstrated promising results in delineating accurate cultivated yield [3,7,9,13,19,20]. Furthermore, the vegetation index derived from satellite images was also successfully applied for predicting rice yield before harvesting [21–23]. One of the popular remote sensing products is MODIS NDVI data, which has the benefits of decadal archives and high spatiotemporal resolution and has been widely employed for regional agricultural yield assessment and forecast [24]. The most widely used vegetation indices are the normalized difference vegetation index (NDVI), enhanced vegetation index (EVI), soil-adjusted vegetation index (SAVI), leaf area index (LAI), and the fraction of absorbed photosynthetically active radiation index (FPAR) [13,21,23,25,26]. In [27,28], the authors found a substantial link between NDVI and LAI and green biomass yield before rice harvest. In this study, MODIS NDVI product along with MODIS LAI and FPAR product and several meteorological factors were used in the experimental rice yield estimation model development. Since meteorological factors such as monthly rainfall, land surface temperature, potential evapotranspiration, and soil moisture content have a substantial influence on rice crop productivity, it is crucial to take into account this information in the yield estimation workflow [29–31]. Based on the literature, we initially selected NDVI, LAI, FPAR as the primary factors, then added soil moisture, rainfall, total precipitation amount, land surface temperature (LST), and evapotranspiration (ET) as auxiliary factors.

The selection of a suitable ML model is also an important part of rice yield estimation model development. As the yearly crop yield dataset extends only over a few years (the years 2001 to 2019 in this study), an efficient ML model that can learn from only a few multivariate time-series data points and provide a satisfactory estimation is necessary. ML models such as random forests (RF), decision trees, artificial neural networks (ANN), and support vector machines (SVM) are effective for building the non-linear relationship between predictors and an observed phenomenon, which has led researchers to apply these techniques to crop yield estimation and prediction [20,24,32–37]. ML models can handle an infinite number of complicated, multi-dimensional datasets and exhibit promising results while predicting yields that are more accurate than those from traditional statistical regression models [38–43]. Several studies applied ML to develop a rice yield estimation model by integrating meteorological variables such as LST, ET, pressure, solar radiation, soil moisture, total precipitation amount, and relative humidity along with NDVI, with satisfactory results [24,27,29]. However, these models produced excellent results for the most part, although they were hampered by the optimization of variables and overfitting due to limited data points that affected crop yield estimation. Considering data availability, multicollinearity, and overfitting issues of ML models, this study constructed a hybrid model by stacking four different simple tree-based regressor models.

Initially, we used linear regression as a benchmark model in this study and trained several ML models to see how they improved accuracy over benchmark models. This study also applied one of the popular tree-based regressor models XGBoost and compared the result with the benchmark model. Finally, we proposed a customized stack-ensemble model by combining four different tree-based regressors: XGBoost [44], LightGBM [45], Gradient Boost [46], and random forest model [47]. The main limitation for accurate crop yield estimation model development for Nepal was lacking a sufficient amount of training data. As the yearly crop yield training data available was from the years 2001 to 2019 for the study area, there were only 19 time-series data points available for each district, which was not good enough for learning patterns from data. A low amount of training data with a limited number of predictors may cause poor performance, and if multicollinearity presents among the predictors, it will severely influence the model's performance. To overcome these issues, we carefully selected tree-based regressors approaches which are not affected by multicollinearity. To improve the crop yield estimation accuracy, we also employed a series of data processing techniques, which are described in the methodology section. Initially, we carried out a few experiments with LR, RF, LightGBM, Gradient Boost, XGBoost, and stack-ensemble model with only NDVI vs. yield in order to understand how NDVI alone is associated with rice yield. Later, the meteorological factors were included in the models to improve the accuracy.

2. Materials and Methods

The entire workflow of crop yield estimation was implemented in several steps, including study area selection, data preprocessing, combining multiple datasets, ML model development, and model performance evaluation. Figure 1 displays the simplified workflow of the entire crop yield estimation model development process.

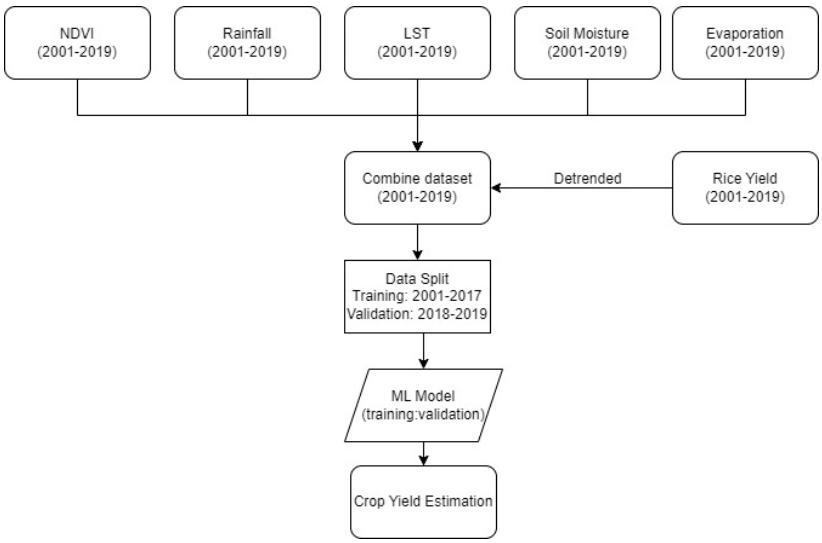


Figure 1. Crop yield estimation workflow.

2.1. Study Area

The study area (Figure 2) is the Terai belt region of southern lowland Nepal that lies south of the outer foothills of the Himalayas. Nepal has unique variation in its landforms and climate conditions. Because of this diversity, Nepal also has a wide range of natural vegetation, from the subtropical evergreen forests of the Terai region to the pine forests, alpine grasslands, and tundra types of the Himalayan Region. This indicates that Nepal's landform conditions include a variety of climate and vegetation. The monsoon has a significant impact on Nepal's climate. People there cultivate tropical and subtropical crops such as paddy, mustard, tobacco, and sugarcane due to the easier irrigation facilities. Each year, cultivation of up to three crop types is possible due to the hot temperature (Figure 3).

The agricultural yield is high when monsoon rains arrive on schedule. When the monsoon rains are unreliable or excessive, farmers in Terai and hilly areas suffer from droughts, floods, and landslides.

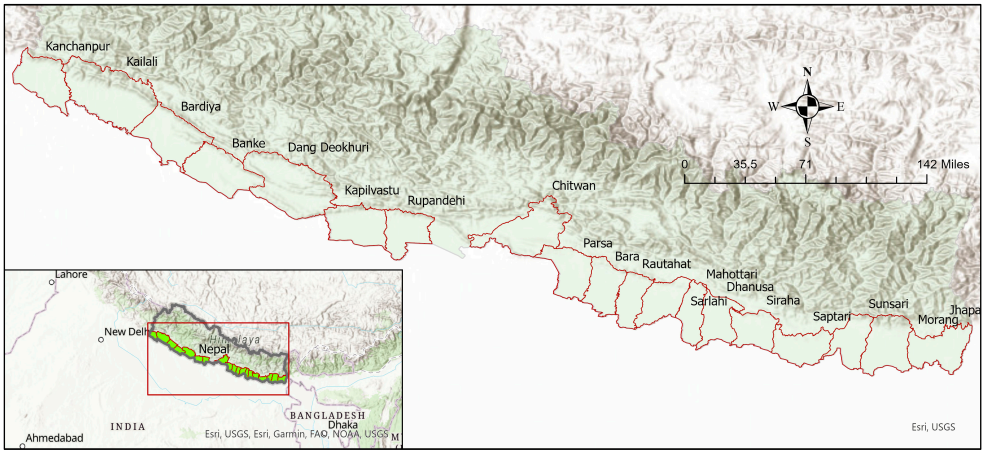


Figure 2. The study area map showing nineteen districts of the Terai belt region in Nepal.

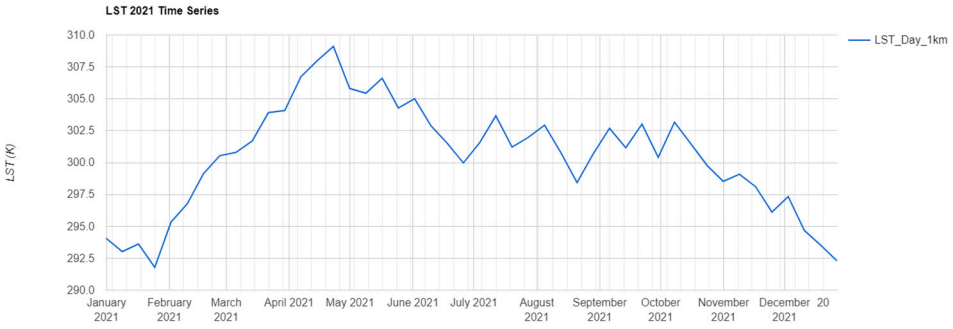


Figure 3. Average land surface temperature (LST) of the Terai belt region of Nepal in the year 2021.

Rice is one of the most consumed foods in Nepal. The Terai regions produce most of the rice, which is about 73% of the total rice production, while hills and high hill areas produce 24% and 4%, respectively (Tripathi et al., 2019). For a country such as Nepal, an effective rice yield estimation tool can help to identify the factors affecting yield and determine optimal harvesting time and early or near-real-time estimation of rice yield. This vital information can help to reduce crop loss and ensure food security by helping governments, decision makers, and other stakeholders to formulate appropriate policy measures for sustainable development.

2.2. Data Collection and Processing

2.2.1. Rice Yield

The rice-growing season in Nepal is from July to October and the harvesting season is from October to mid-December. The yearly crop yield data was collected by partners at The International Centre for Integrated Mountain Development (ICIMOD). The dataset contains rice yield information for 19 districts from the years 2000 to 2019 and 10 m spatial raster data of rice cultivation field for the years 2020 and 2021. The yearly yield data was detrended by applying a polynomial curve fitting to make it stationary as it allows for any potential sub-trends in the data to be observed, which is common for time-series statistical modeling [48]. The yield delta (i.e., the difference between the actual reported yield from the polynomial curve) was used as the ground truth data in the ML models. We also made a combined rice field mask by merging the raster layer of rice field of the years 2020 and 2021. The rice mask layer was used to exclude non-agricultural pixels from NDVI and other data layers. Figure 4 displays a district-wise yearly rice yield.

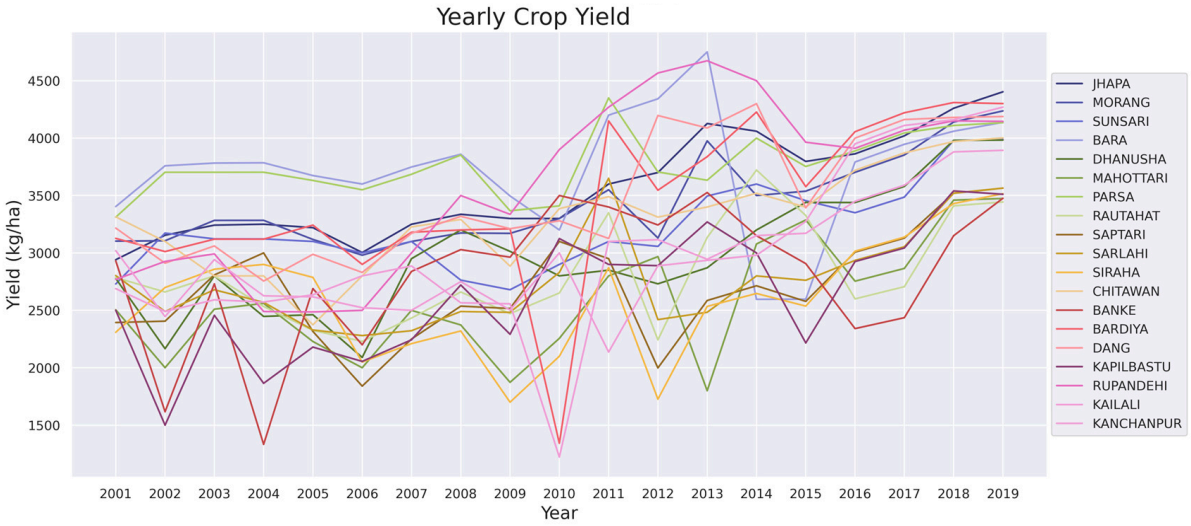


Figure 4. Yearly crop yield (kg/ha) in each district.

The spatiotemporal rice yield maps (Figure 5) exhibit yearly rice yield patterns that vary from district to district and time to time. A slightly increasing trend of rice yield in all districts in the recent years can be observed from the spatiotemporal rice yield maps.

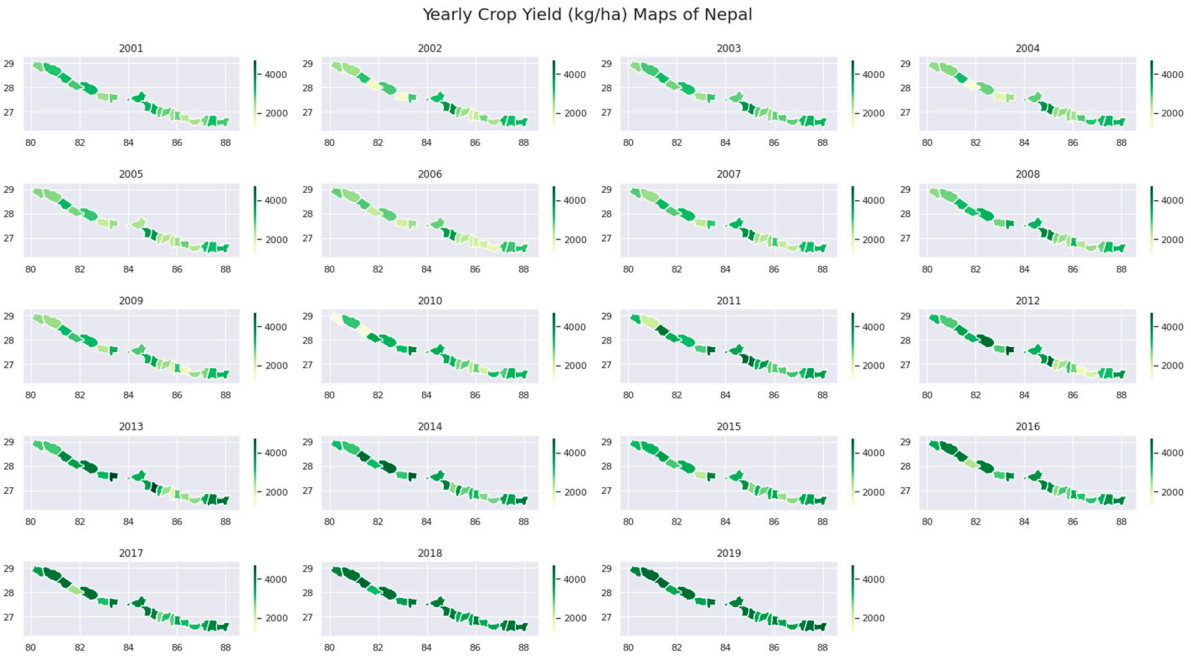


Figure 5. Yearly rice yield (kg/ha) maps for 21 rice-growing districts in Nepal (X-axis: longitude and Y-axis: latitude).

2.2.2. NDVI

Rice growing and harvesting can be monitored from remotely sensed data. It can be used to make relationships with yearly yield trends in order to make the estimation of the yield. However, there are several limitations of remotely sensed data. We noticed the presence of a significant amount of cloud cover on Landsat/Sentinel-2 data over Nepal during the rice-growing season. Therefore, we were not able to derive high spatial resolution NDVI products from Landsat/Sentinel-2 satellite imageries; instead, the team used coarser spatial resolution but finer temporal resolution MODIS products where images with a significant presence of cloud cover were filtered out. Initially, we selected the peak NDVI image for each district during crop growing and harvesting season for each year and

then we extended the selection algorithm backward and forward to collect a few more peak NDVI images for a particular year. Later, we limited the selection criteria only between August and September, based on local expert opinion, in order to avoid noisy NDVI data. We collected a total of eight NDVI images for August and September, and the average NDVI values for all pixels within agricultural lands for each district was calculated. We also used MODIS LAI and FPAR products, but we found that NDVI, LAI, and FPAR were highly correlated with each other where the correlation coefficient between NDVI vs. LAI = 0.99 and NDVI vs. FPAR = 0.97, and our experiment showed that using both products together did not improve the yield estimation accuracy. Therefore, we excluded the LAI and FPAR products and only used the NDVI product as the latter had a better spatial resolution (250 m) than the following image products (500 m). NDVI values were smoothened by using the Savitzky–Golay filter (Savitzky and Golay, 1964). The Savitzky–Golay (SAVGOL) filter is used to eliminate noise and improve the smoothness of time-series NDVI values. The filter calculates a polynomial fit of each window based on polynomial degree and window size. Figure 6 displays a smoother time-series pattern of NDVI values each year for all rice-growing districts in Nepal.

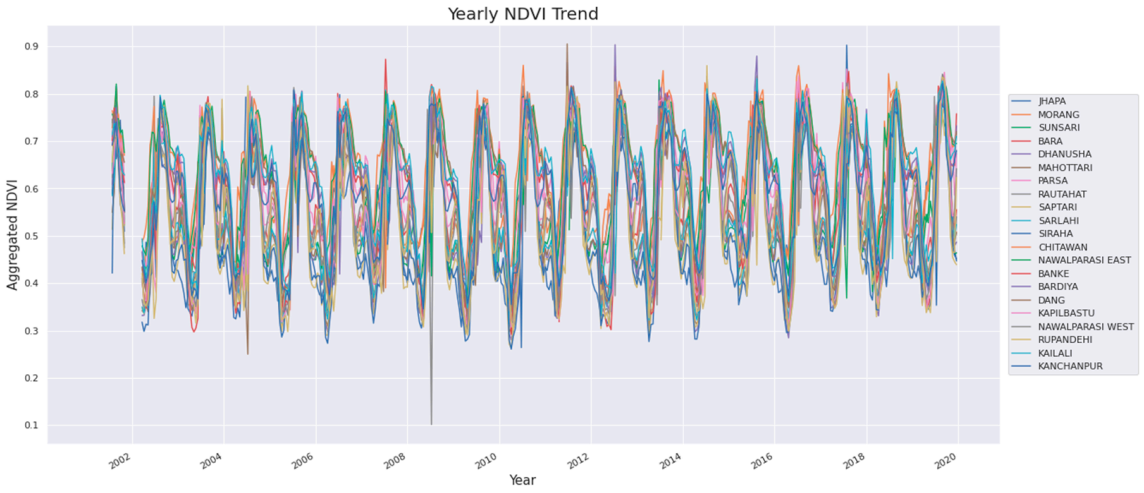
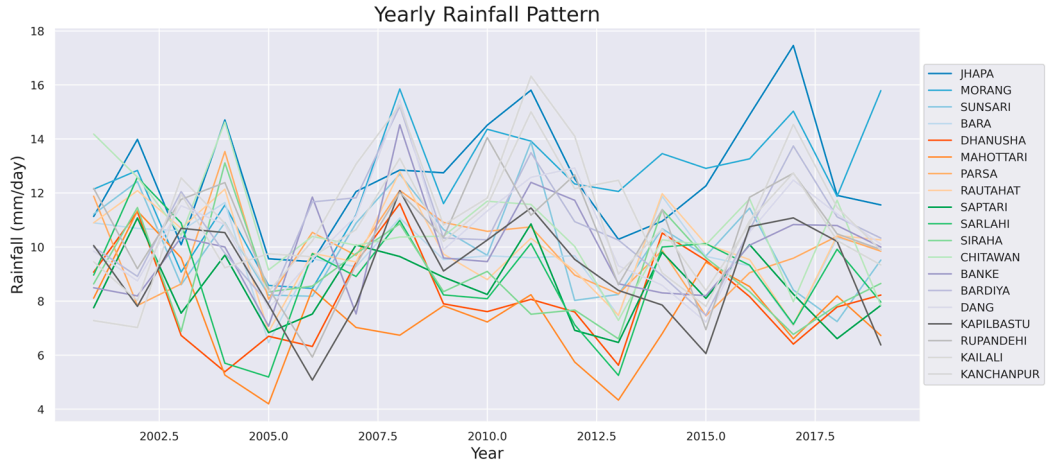


Figure 6. Time-series graph of NDVI for rice-growing districts in Nepal.

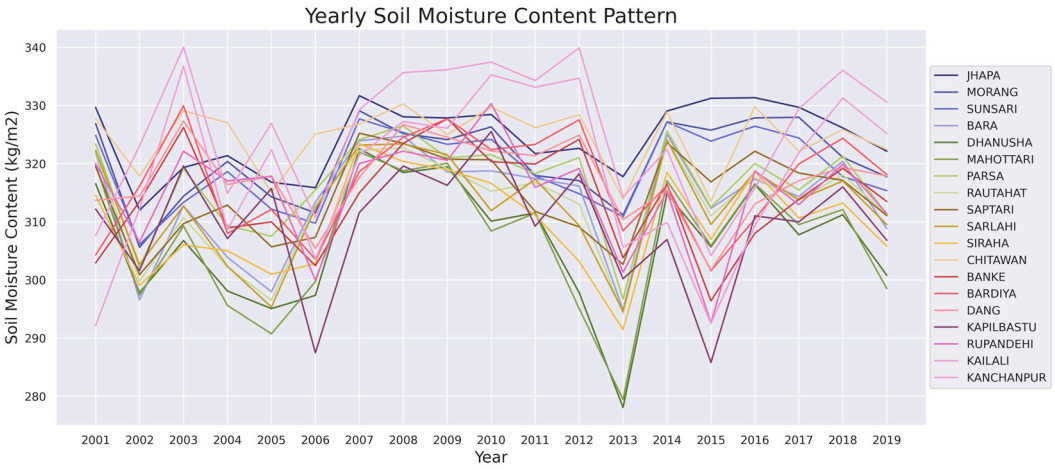
2.2.3. Auxiliary Variables

Apart from NDVI, we also used several auxiliary variables, including rainfall (Figure 7a), soil moisture content (Figure 7b), LST (Figure 7c), ET (Figure 7d), and total precipitation amount (Figure 7e). The yearly dataset was collected from ICIMOD. The following figures show an average yearly pattern of each five auxiliary variables only for the corresponding months of the NDVI images. Then, the mean of yearly values for each district was calculated and used in ML models as a factor.

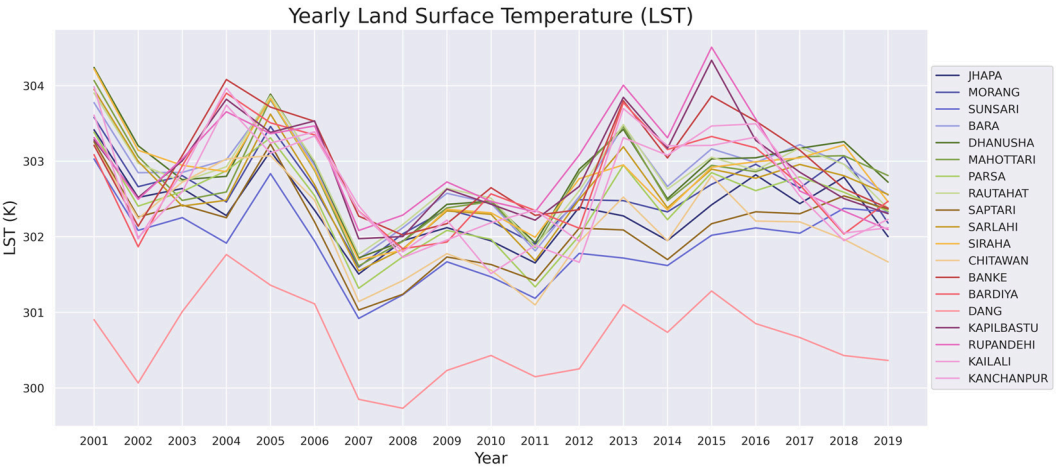
The district-wise processed dataset was split into two sets—the training set which contains data points from the years 2001 to 2017 (17 data points for training) and the validation set which contains data points from the years 2018 and 2019 (2 data points for validation) for each district. Each dataset has a total of six features—NDVI, rainfall, total precipitation amount, soil moisture, evapotranspiration, LST, and corresponding yield delta. The yield delta was calculated from actual reported yield and polynomial curve fitting; this was then used as the dependent variable in the models. ML models were trained using the training set whereas the validation dataset was used to evaluate the performance of the models. During the validation process, the predicted yield was calculated by adding model output (i.e., yield delta) with yearly base yield from polynomial curve fitting.



(**a**)

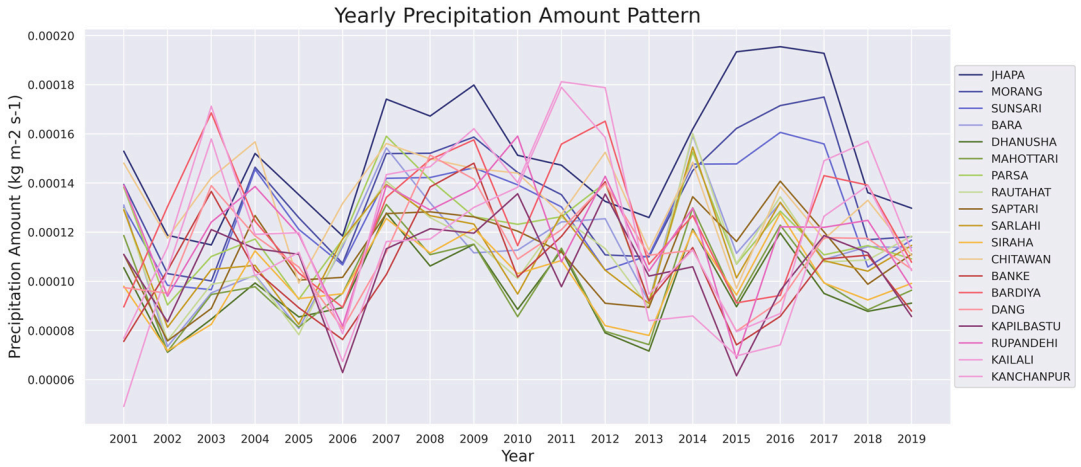


(**b**)

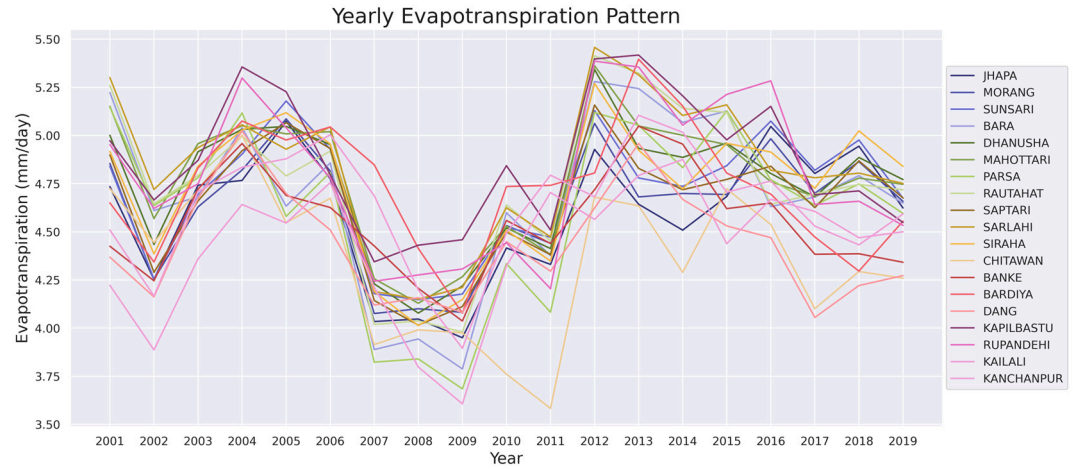


(**c**)

Figure 7. *Cont.*



(**d**)



(**e**)

Figure 7. (**a**) Yearly rainfall (mm/day) pattern for rice-growing districts in Nepal. (**b**) Yearly soil moisture content (kg/m^2) pattern for rice-growing districts in Nepal. (**c**) Yearly LST (K) pattern for rice-growing districts in Nepal. (**d**) Yearly total precipitation amount (kg m^{-2} s^{-1}) pattern for rice-growing districts in Nepal. (**e**) Yearly evapotranspiration (mm/day) pattern for rice-growing districts in Nepal.

2.3. Stack-Ensemble Model

Model stacking is a way to improve model predictions by combining the outputs of multiple models and running them through another machine-learning model called a meta-learner. In the customized model, we only used four tree-based regressions. The reason for using tree-based regression models is that multicollinearity does not affect the estimation. The only limitation was overfitting as the dataset contains only 19 cases (from 2001 to 2019) for each district. To observe whether the model is overfitting or underfitting and the actual prediction performance, the dataset was split into two sets: the aforementioned training and testing sets. After training the model on the training set, we used the validation dataset to check the actual performance of the model. The model accuracy was measured based on the validation dataset in terms of accuracy assessment metrics, which are described in the following section. Figure 8 displays the architecture of the customized stack-ensemble model.

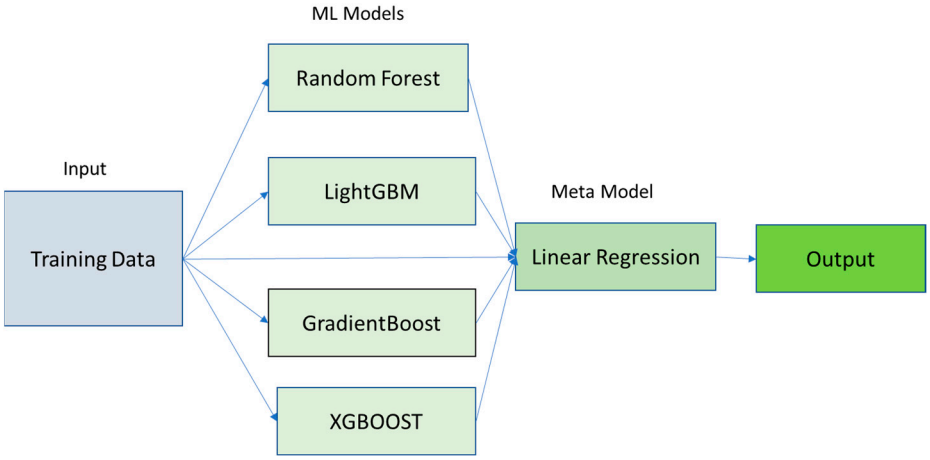


Figure 8. Customized stack-ensemble model.

2.4. Accuracy Assessment Metrics

This study primarily used traditional statistical accuracy assessment metrics such as root mean squared error (*RMSE*) and mean absolute error (*MAE*) for model evaluation. *RMSE* is the standard deviation of the residuals (i.e., the difference between actual and predicted values, which is also called the error). *RMSE* measures how spread out these residuals are or how concentrated the data are around the line of best fit.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} ||y(i) - \hat{y}(i)||^2}{N}} \quad (1)$$

where N is the number of data points, $y(i)$ is the i-th measurement, and $\hat{y}(i)$ is its corresponding prediction.

MAE measures the absolute mean difference of the predicted and actual values.

$$MAE = \frac{\sum_{i=1}^{N} |y_i - x_i|}{n} \quad (2)$$

where y_i is the prediction, x_i is the true value/ground truth value, and n is the total number of data points.

3. Result

Several models were trained and tested and compared regarding the potentiality in accurate rice yield estimation, and we found that the proposed stack-ensemble model achieved overall less error than the benchmark linear regression and other ML models. While evaluating the model's performance on the validation dataset, the model's output was added with yearly-based yield from polynomial curve fitting to calculate final prediction of yield. The predicted yield values were compared with the actual reported yield with accuracy assessment metrics, which are presented in Table 1.

The experiments were carried out in two phases—(initially, only NDVI was used to model yield. We found that the benchmark LR model can achieve an average 685.17 RMSE and 633.83 MAE for all districts whereas other ML models performed slightly better than the benchmark model. The proposed model with the same predictor achieved an overall 451.05 RMSE and 425.35 MAE. The results indicate that the proposed stack-ensemble model combining four different tree-based regression model is more effective and provide more accurate yield estimation. (2) Later, we added five different auxiliary variables (Rainfall, Total precipitation amount, Soil Moisture, Evapotranspiration, and LST) with NDVI and found that the auxiliary variables significantly improve the prediction performance of all models. With new variables added to the stack-ensemble model, its accuracy improved dramatically. The average RMSE was reduced to 328.06 and MAE was reduced to 317.21.

Based on the validation accuracy, this study used only the stack-ensemble model for further analysis.

Table 1. Accuracy assessment metrics of each model based on the validation dataset.

Models	Average RMSE	Average MAE
NDVI vs. Yield		
LR	685.17	636.83
RF	556.19	507.45
Gradient Boost	575.03	562.04
LightGBM	545.85	502.08
XGBoost	552.27	507.02
Stack Ensemble	**451.05**	**425.35**
NDVI + Auxiliary Variables vs. Yield		
LR	550.94	514.50
RF	361.52	337.19
Gradient Boost	372.03	359.04
LightGBM	356.85	334.55
XGBoost	355.90	333.12
Stack Ensemble	**328.06**	**317.21**

A district-wise comparison among predicted and actual yield (kg/ha) with validation accuracy from the stack-ensemble model was presented in Table 2. In some districts, the predicted yields were very close to the actual yield values. It can be seen that the actual reported and predicted yield for Dang district was 4180 kg/ha and 4120 kg/ha, respectively, in the year 2018 where the yield delta was 60 kg/ha, which means the yield estimation was 98.6% accurate. However, the model's accuracy was lowest in Sarlahi district where the actual and predicted yield was 3520 kg/ha and 3068 kg/ha, respectively, and the difference was 451 kg/ha, which means the estimation accuracy was 87.2%. In this district, the model gained the worst accuracy with 455 kg/ha RMSE and 455 kg/ha MAE. We can see variation of yield estimation accuracy from district to district due to their unique geographical characteristics. However, it can be seen that the overall error difference in most of the districts was minimal and the predicted values were close to the actual yield.

Table 2. District-wise prediction and original yield (kg/ha) and validation metrics. Based on the validation dataset.

District	Yield 2018	Predicted Yield 2018	Percentage of Error 2018 (%)	Yield 2019	Predicted Yield 2019	Percentage of Error 2019 (%)	RMSE	MAE
JHAPA	4260	4144	2.7	4403	4203	4.5	163.62	158.08
MORANG	4140	3931	5.05	4237	3953	6.7	248.96	246.08
SUNSARI	3980	3574	10.20	3987	3656	8.3	370.41	368.55
BARA	4060	3773	7.0	4137	3928	5.1	251.22	248.20
DHANUSHA	3980	3654	8.2	3983	3501	12.1	411.30	403.83
MAHOTTARI	3460	3179	8.1	3474	2937	15.4	428.44	408.79
PARSA	4110	3994	2.8	4134	3997	3.3	126.61	126.17
RAUTAHAT	3410	3191	6.4	3443	3024	12.17	334.25	318.90
SAPTARI	3520	3073	12.7	3564	3119	12.5	445.77	445.77
SARLAHI	3520	3068	12.8	3564	3104	12.9	455.45	455.43
SIRAHA	3430	3078	10.2	3515	2744	21.9	599.29	561.04
CHITAWAN	3970	3676	7.4	4002	3755	6.17	270.75	269.74
BANKE	3150	2959	6.1	3476	3134	9.83	276.56	266.07
BARDIYA	4310	4204	2.5	4300	3989	7.23	232.49	208.37
DANG	4180	4120	1.4	4188	4281	2.2	78.40	76.45
KAPILBASTU	3540	3517	0.6	3511	3188	9.2	228.34	172.50
RUPANDEHI	4150	4474	7.8	4145	4542	9.5	362.59	360.79
KAILALI	4160	3803	8.5	4270	3671	14.0	492.81	477.64
KANCHANPUR	3880	3388	12.68	3893	3476	10.71	456.02	454.50

4. Discussion

To understand the importance of factors on crop yield, the feature importance graphs (Figure 9) was depicted using the XGBoost model. The feature importance graphs present the contribution of each factor while developing a tree-based regression model. It can be seen that the importance of each factor varies from district to district. The reason for the varying importance of each factor on rice yield estimation is because of the unique diverse physio-geographic characteristics of each district. In eighteen districts, NDVI was the most contributing factor, followed by rainfall in one district. In addition, LST was the least important predictor for eight districts. Based on the F score, the importance of variables were ranked accordingly as NDVI > rainfall > evapotranspiration > soil moisture > LST.

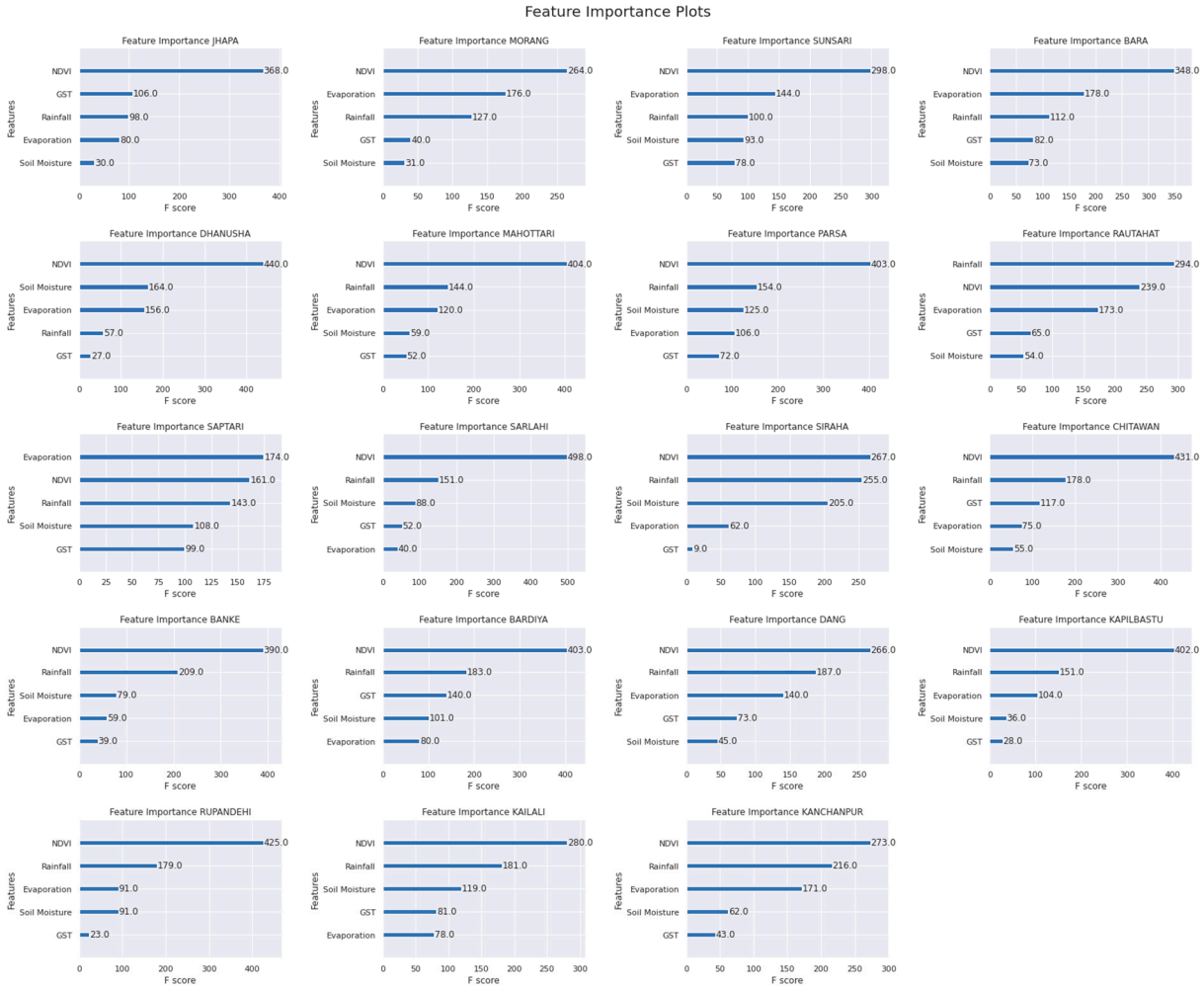


Figure 9. Feature importance plot from XGBoost model for each district.

The actual reported vs. predicted yield graph (Figure 10) shows the time-series trend of rice yield along with a linear curve, as well as estimation of yield during the training and validation period for all districts. It is evident from the time-series graphs that yearly rice yield (blue curve) fluctuates highly and apparently there is no linear pattern. It also shows that linear models such as linear regression are not suitable for crop yield estimation for Nepal as there are no linear trends. Instead, we can see that the non-linear stack-ensemble model exhibits an almost similar pattern to the yield trend curve. However, it can be seen as significant differences sometimes but, even still, the fitted curves look similar to the non-linear trend of yearly rice yield.

Figure 10. Time-series graph of actual yield, linear trend, and predicted yield for each district.

Finally, we plotted the error difference between the actual yield and predicted yield for the years 2018 (Figure 11a) and 2019 (Figure 11b) for all Terai belt districts. It shows the error variation of each district for the years 2018 and 2019. We can see that average errors in all districts are moderate to low. The largest and lowest differences in 2018 were found in Sarlahi, where 452 kg/ha of rice yield was underestimated, and Kapilbastu, where 23 kg/ha of rice yield was underestimated. Additionally, in 2019, the biggest discrepancy was seen in Siraha, where 771 kg/ha of rice yield was underestimated, while the smallest discrepancy was found again in Dang district where the 93 kg/ha of rice yield was overestimated. Overall, the average error difference for all districts was 8.33%.

Based on the analysis and results, we can conclude that the proposed stack-ensemble based rice yield model is an effective and alternative approach for predicting crop yield. This conclusion is supported by the promising results obtained in comparison to both the benchmark and other ML models. The workflow presented in this study can serve as a guide for estimating crop yield, including steps such as detrending time-series yield data, processing remote sensing data (e.g., applying the SAVGOL filter to eliminate noise from NDVI trends), combining remote sensing and meteorological data, and selecting a suitable ML model that can learn from limited observations and provide accurate estimates. This entire workflow can be replicated for rapid rice yield estimation in other locations.

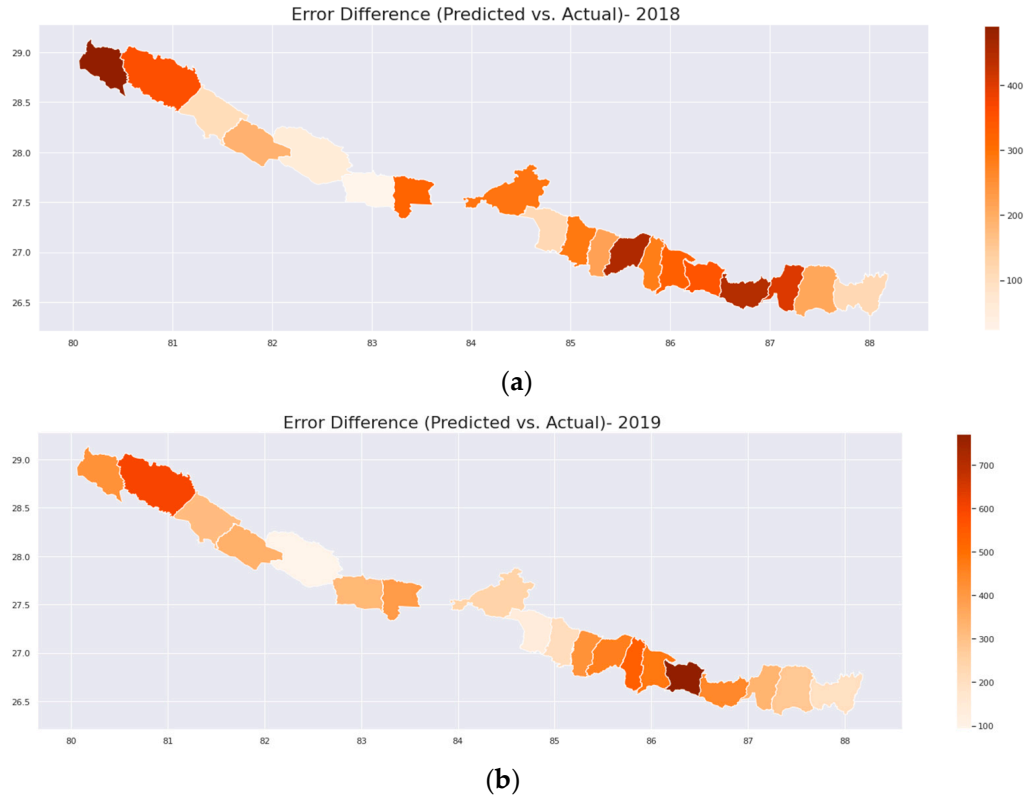


Figure 11. (**a**) Error differences of rice yield (kg/ha) for all district in year 2018 for the non-linear stack-ensemble model approach. (**b**) Error differences of rice yield (kg/ha) for all districts in the year 2019 for the non-linear stack-ensemble model approach.

5. Conclusions

This study investigated the relationship between MODIS NDVI products along with several meteorological factors and yearly rice yield. It also presented appropriate data preprocessing techniques and the performance of customized ML models in order to improve the rice yield estimation accuracy. We found that the MODIS NDVI product alone was not a strong predictor. Therefore, incorporating other climatic variables improved the model's prediction performance and provided a more accurate estimation of rice yield. Along with NDVI, we found that rainfall, soil moisture, and evapotranspiration also showed a strong relationship with rice yield. Finally, we found that tree-based regression models, especially hybrid stack ensembles of multiple models, can perform better with few data points and achieved 288 kg/ha RMSE, which was a significant improvement from the initial 685 RMSE of the benchmark LM model. Due to the unique and distinct physio-geographical setting of each district, we can see a variation of estimation accuracy. There are a few limitations in this study, including data availability and lacking sufficient ground truth information. We found the presence of a significant amount of cloud cover over the Terai belt districts in Nepal during the main crop growing and harvesting season. Therefore, we were not able to derive higher-resolution NDVI products from Landsat/Sentinel 2 imageries. In the future, Sentinel 1 image products could be used to delineate rice fields and calculate a vegetation index that may achieve better estimation accuracy. Additionally, more ground truth data and other meteorological and environmental parameters could be incorporated for statistical analysis and modeling in the future, targeting specific districts that performed poorly, to better understand the crop phenology and improve yield estimation accuracy. However, the overall estimation accuracy for all districts was satisfactory and the workflow can be replicated in other areas that can help ensure food security.

Author Contributions: Methodology, M.D.I., L.D. and L.L.; Validation, S.S. and L.L.; Formal analysis, M.D.I. and L.L.; Investigation, M.D.I. and L.L.; Resources, L.G.; Data curation, F.M.Q. and S.S.; Writing—original draft, M.D.I. and L.G.; Writing—review & editing, L.D., F.M.Q., S.S., L.G., L.L., T.J.M. and A.R.P.; Visualization, M.D.I.; Supervision, T.J.M. and A.R.P.; Project administration, F.M.Q., S.S., L.G., T.J.M. and A.R.P.; Funding acquisition, L.D., F.M.Q., L.G., T.J.M. and A.R.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the NASA SERVIR program (Grant # 80NSSC20K0161, PI: Dr. Liping Di).

Data Availability Statement: All data used in this study are available for other researchers to use. Please contact the corresponding author for the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Godfray, H.C.J.; Beddington, J.R.; Crute, I.R.; Haddad, L.; Lawrence, D.; Muir, J.F.; Pretty, J.; Robinson, S.; Thomas, S.M.; Toulmin, C. Food Security: The Challenge of Feeding 9 Billion People. *Science* **2010**, *327*, 812–818. [CrossRef]
2. Guo, L.; Di, L.; Zhang, C.; Lin, L.; Chen, F.; Molla, A. Evaluating contributions of urbanization and global climate change to urban land surface temperature change: A case study in Lagos, Nigeria. *Sci. Rep.* **2022**, *12*, 14168. [CrossRef]
3. Mosleh, M.K.; Hassan, Q.K.; Chowdhury, E.H. Application of Remote Sensors in Mapping Rice Area and Forecasting Its Production: A Review. *Sensors* **2015**, *15*, 769–791. [CrossRef]
4. Tang, J.; Di, L. Past and Future Trajectories of Farmland Loss Due to Rapid Urbanization Using Landsat Imagery and the Markov-CA Model: A Case Study of Delhi, India. *Remote Sens.* **2019**, *11*, 180. [CrossRef]
5. Fernandez-Beltran, R.; Baidar, T.; Kang, J.; Pla, F. Rice-Yield Prediction with Multi-Temporal Sentinel-2 Data and 3D CNN: A Case Study in Nepal. *Remote Sens.* **2021**, *13*, 1391. [CrossRef]
6. Zhang, C.; Yang, Z.; Zhao, H.; Sun, Z.; Di, L.; Bindlish, R.; Liu, P.-W.; Colliander, A.; Mueller, R.; Crow, W.; et al. Crop-CASMA: A web geoprocessing and map service based architecture and implementation for serving soil moisture and crop vegetation condition data over U.S. Cropland. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *112*, 102902. [CrossRef]
7. Yu, E.; Di, L.; Meyer, D.; Zhao, P.; Lin, L.; Zhang, C.; Cvejovic, S. ICroplandNet: An Open Distributed Training Dataset for Irrigated Cropland Detection. In Proceedings of the 2022 10th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Quebec City, QC, Canada, 11–14 July 2022; p. 6.
8. Chalise, S.; Naranpanawa, A. Climate change adaptation in agriculture: A computable general equilibrium analysis of land-use change in Nepal. *Land Use Policy* **2016**, *59*, 241–250. [CrossRef]
9. Islam, M.M.; Matsushita, S.; Noguchi, R.; Ahamed, T. Development of remote sensing-based yield prediction models at the maturity stage of boro rice using parametric and nonparametric approaches. *Remote Sens. Appl. Soc. Environ.* **2021**, *22*, 100494. [CrossRef]
10. Kratoska, P.H. The Impact of the Second World War on Commercial Rice Production in Mainland South-East Asia. In *Food Supplies and the Japanese Occupation in South-East Asia*; Kratoska, P.H., Ed.; Studies in the Economies of East and South-East Asia; Palgrave Macmillan: London, UK, 1998; pp. 9–31. ISBN 978-1-349-26937-2.
11. AR5 Synthesis Report: Climate Change 2014—IPCC. Available online: https://www.ipcc.ch/report/ar5/syr/ (accessed on 16 January 2023).
12. Rosegrant, M.W.; Cai, X.; Cline, S.A. *World Water and Food to 2025 Dealing with Scarcity*; International Food Policy Research Institute: Washington, DC, USA, 2022.
13. Li, H.; Di, L.; Zhang, C.; Lin, L.; Guo, L. Improvement of In-season Crop Mapping for Illinois Cropland Using Multiple Machine Learning Classifiers. In Proceedings of the 2022 10th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Quebec City, QC, Canada, 11–14 July 2022; pp. 1–6.
14. Zhang, C.; Di, L.; Lin, L.; Li, H.; Guo, L.; Yang, Z.; Yu, E.G.; Di, Y.; Yang, A. Towards automation of in-season crop type mapping using spatiotemporal crop information and remote sensing data. *Agric. Syst.* **2022**, *201*, 103462. [CrossRef]
15. Zhang, C.; Di, L.; Yang, Z.; Lin, L.; Zhao, H.; Yu, E.G. An Overview of Agriculture Cyberinformatics Tools to Support USDA NASS Decision Making. In Proceedings of the 2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Shenzhen, China, 26–29 July 2021; pp. 1–6.
16. Zhao, H.; Di, L.; Sun, Z.; Hao, P.; Yu, E.; Zhang, C.; Lin, L. Impacts of Soil Moisture on Crop Health: A Remote Sensing Perspective. In Proceedings of the 2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Shenzhen, China, 26–29 July 2021; pp. 1–4.
17. Yu, E.G.; Di, L.; Qamer, F.M.; Zhao, H.; Yu, Z.; Lin, L.; Zhang, C.; Cvejovic, S. Rice Modeling Using Long Time Series of High Temporal Resolution Vegetation Indices in Nepal. In Proceedings of the 2022 10th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Quebec City, QC, Canada, 11–14 July 2022; pp. 1–6.
18. Zhao, H.; Di, L.; Sun, Z. WaterSmart-GIS: A Web Application of a Data Assimilation Model to Support Irrigation Research and Decision Making. *ISPRS Int. J. Geo-Inf.* **2022**, *11*, 271. [CrossRef]

19. Estimation of Crop Evapotranspiration from MODIS Data by Combining Random Forest and Trapezoidal Models—ScienceDirect. Available online: https://www.sciencedirect.com/science/article/abs/pii/S0378377421005266?via%3Dihub (accessed on 16 January 2023).
20. Yao, A.; Di, L. Machine Learning-based Pre-season Crop Type Mapping: A Comparative Study. In Proceedings of the 2021 9th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Shenzhen, China, 26–29 July 2021; pp. 1–4.
21. Rahman, A.; Roytman, L.; Krakauer, N.Y.; Nizamuddin, M.; Goldberg, M. Use of Vegetation Health Data for Estimation of Aus Rice Yield in Bangladesh. *Sensors* **2009**, *9*, 2968–2975. [CrossRef]
22. Liang, L.; Di, L.; Zhang, L.; Deng, M.; Qin, Z.; Zhao, S.; Lin, H. Estimation of crop LAI using hyperspectral vegetation indices and a hybrid inversion method. *Remote Sens. Environ.* **2015**, *165*, 123–134. [CrossRef]
23. Di, L.; Rundquist, D.C.; Han, L. Modelling relationships between NDVI and precipitation during vegetative growth cycles. *Int. J. Remote Sens.* **1994**, *15*, 2121–2136. [CrossRef]
24. Son, N.-T.; Chen, C.-F.; Chen, C.-R.; Guo, H.-Y.; Cheng, Y.-S.; Chen, S.-L.; Lin, H.-S.; Chen, S.-H. Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan. *Int. J. Remote Sens.* **2020**, *41*, 7868–7888. [CrossRef]
25. Dubey, S.K.; Gavli, A.S.; Yadav, S.K.; Sehgal, S.; Ray, S.S. Remote Sensing-Based Yield Forecasting for Sugarcane (*Saccharum officinarum* L.) Crop in India. *J. Indian Soc. Remote Sens.* **2018**, *46*, 1823–1833. [CrossRef]
26. Palakuru, M.; Yarrakula, K. Study on paddy phenomics ecosystem and yield estimation using space-borne multi sensor remote sensing data. *J. Agrometeorol.* **2019**, *21*, 171–175. [CrossRef]
27. Rahman, A.; Khan, K.; Krakauer, N.Y.; Roytman, L.; Kogan, F. Use of Remote Sensing Data for Estimation of Aman Rice Yield. *Int. J. Agric. For.* **2012**, *2*, 101–107. [CrossRef]
28. Sonobe, R.; Miura, Y.; Sano, T.; Horie, H. Estimating leaf carotenoid contents of shade-grown tea using hyperspectral indices and PROSPECT–D inversion. *Int. J. Remote Sens.* **2018**, *39*, 1306–1320. [CrossRef]
29. Gandhi, N.; Petkar, O.; Armstrong, L.J. Rice crop yield prediction using artificial neural networks. In Proceedings of the 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), Chennai, India, 15–16 July 2016; pp. 105–110.
30. Chandra, A.; Mitra, P.; Dubey, S.K.; Ray, S.S. Machine Learning Approach for Kharif Rice Yield Prediction Integrating Multi-Temporal Vegetation Indices and Weather and Non-Weather Variables. *ISPRS—Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2019**, *423*, 187–194. [CrossRef]
31. Guruprasad, R.B.; Saurav, K.; Randhawa, S. Machine Learning Methodologies for Paddy Yield Estimation in India: A Case Study. In Proceedings of the IGARSS 2019—2019 IEEE International Geoscience and Remote Sensing Symposium, Yokohama, Japan, 28 July–2 August 2019; pp. 7254–7257.
32. Aghighi, H.; Azadbakht, M.; Ashourloo, D.; Shahrabi, H.S.; Radiom, S. Machine Learning Regression Techniques for the Silage Maize Yield Prediction Using Time-Series Images of Landsat 8 OLI. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2018**, *11*, 4563–4577. [CrossRef]
33. Islam, M.D.; Chakraborty, T.; Alam, M.S.; Islam, K.S. Urban heat island effect analysis using integrated geospatial techniques: A case study on Khulna city, Bangladesh. In Proceedings of the International Conference on Climate Change, Dhaka, Bangladesh, 8–11 January 2019.
34. Islam, M.D.; Li, B.; Islam, K.S.; Ahasan, R.; Mia, M.R.; Haque, M.E. Airbnb rental price modeling based on Latent Dirichlet Allocation and MESF-XGBoost composite model. *Mach. Learn. Appl.* **2022**, *7*, 100208. [CrossRef]
35. Bappa, S.A.; Malaker, T.; Mia, M.R.; Islam, M.D. Spatio-temporal variation of land use and land cover changes and their impact on land surface temperature: A case of Kutupalong Refugee Camp, Bangladesh. *Heliyon* **2022**, *8*, e10449. [CrossRef]
36. Islam, M.D.; Di, L.; Mia, M.R.; Sithi, M.S. Deforestation Mapping of Sundarbans Using Multi-Temporal Sentinel-2 Data & Transfer Learning. In Proceedings of the 2022 10th International Conference on Agro-geoinformatics (Agro-Geoinformatics), Quebec City, QC, Canada, 11–14 July 2022; pp. 1–4.
37. O'Shea, K.; LaRoe, J.; Vorster, A.; Young, N.; Evangelista, P.; Mayer, T.; Carver, D.; Simonson, E.; Martin, V.; Radomski, P.; et al. Improved Remote Sensing Methods to Detect Northern Wild Rice (*Zizania palustris* L.). *Remote Sens.* **2020**, *12*, 3023. [CrossRef]
38. Islam, M.D.; Islam, K.S.; Ahasan, R.; Mia, M.R.; Haque, M.E. A data-driven machine learning-based approach for urban land cover change modeling: A case of Khulna City Corporation area. *Remote Sens. Appl. Soc. Environ.* **2021**, *24*, 100634. [CrossRef]
39. Sun, J.; Di, L.; Sun, Z.; Shen, Y.; Lai, Z. County-Level Soybean Yield Prediction Using Deep CNN-LSTM Model. *Sensors* **2019**, *19*, 4363. [CrossRef]
40. Islam, M.D.; Li, B.; Lee, C.; Wang, X. Incorporating spatial information in machine learning: The Moran eigenvector spatial filter approach. *Trans. GIS* **2022**, *26*, 902–922. [CrossRef]
41. Sun, J.; Lai, Z.; Di, L.; Sun, Z.; Tao, J.; Shen, Y. Multilevel Deep Learning Network for County-Level Corn Yield Estimation in the U.S. Corn Belt. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2020**, *13*, 5048–5060. [CrossRef]
42. Shrestha, R.; Di, L.; Yu, E.G.; Kang, L.; Shao, Y.; Bai, Y. Regression model to estimate flood impact on corn yield using MODIS NDVI and USDA cropland data layer. *J. Integr. Agric.* **2017**, *16*, 398–407. [CrossRef]
43. Jeong, S.; Ko, J.; Shin, T.; Yeom, J. Incorporation of machine learning and deep neural network approaches into a remote sensing-integrated crop model for the simulation of rice growth. *Sci. Rep.* **2022**, *12*, 9030. [CrossRef]
44. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA, 2016; pp. 785–794.

45. Ke, G.; Meng, Q.; Finley, T.; Wang, T.; Chen, W.; Ma, W.; Ye, Q.; Liu, T.-Y. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Dutchess County, NY, USA, 2017; Volume 30.
46. Friedman, J.H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232. [CrossRef]
47. Breiman, L. *Classification and Regression Trees*; Routledge: New York, NY, USA, 2017; ISBN 978-1-315-13947-0.
48. Lu, J.; Carbone, G.J.; Gao, P. Detrending crop yield data for spatial visualization of drought impacts in the United States, 1895–2014. *Agric. For. Meteorol.* **2017**, *237–238*, 196–208. [CrossRef]