Crop production estimation using deep learning technique

Ashapurna Marndi*, K. V. Ramesh and G. K. Patra

CSIR-Fourth Paradigm Institute, Bengaluru 560 037, India and
Academy of Scientific and Innovative Research, Ghaziabad 201 002, India

Reliable estimation of crop requirement and production in advance, help policy makers to adopt timely decision for trade as export–import, which is a basic building block to assure food security of a country. A powerful and robust algorithm is essential to predict the future demand and production of a particular crop for subsequent years. Deep learning methods are used successfully in solving different prediction problems of various applications. This study attempts to design an efficient AI based technique specifically using long short-term memory, a deep learning approach for estimation of crop production using crop production information of neighbouring countries, which are part of the South Asian monsoon system. Detailed sensitivity analysis is conducted to identify the optimal combination of crop production of neighbouring countries that directly and indirectly impact the crop production of India. Here, we designed and developed a predictive model for rice production of India with lead time of one year using deep learning technique. Along with that, as there are significant influences of local climate (i.e. rainfall data) on crop production, that information was also considered along with crop production of neighbouring countries. The results indicated that local and regional scale parameters jointly improve the prediction capability for future years. Capability of the proposed model was validated with export–import data on crop of India and neighbouring countries, and the validation result showed that our proposed technique was efficient and robust in nature.

Keywords: Artificial intelligence, crop production model, deep neural networks, long short-term memory, sensitivity analysis.

CROP production plays a significant role in satisfying basic food requirement of a country and generating foreign exchange. For a developing country, lack of sufficient food throughout the year can spiral up the inflation index and can ruin the financial balance of its people. Amount of crop production in a country significantly influences food security and economic growth. Therefore, it is necessary to improve crop production to maintain the economic standard of a country. Over years, land for crop production has declined in several countries; on the other hand, demand for food consumption has increased due to rise in population. Therefore, it is essential to increase production of major crops to satisfy the requirement of food consumption and corresponding economic growth. Amount of export and import of a particular crop in a country depend mainly on the amount of its crop production, and local food consumption derivable from its population. Prediction of crop production in advance helps policy makers to take timely decision for export–import trade which is the base for assured food security of a country[1].

Few modelling approaches have been attempted for prediction of crop production like remote sensing model[2], statistical model, etc. though they involve various challenges and sometimes do not lead to satisfying results. Crop yield estimation using remote sensing technique sometimes suffers due to lack of high temporal resolution data and cloud coverage[3], and insufficient ground level data available for validation[4]. Availability of sufficient ground level data is required to validate the model, post design and developing phase. Few statistical approaches have been used to design a crop prediction model. However, capability of statistical models is limited in designing nonlinear relationship among data. It is very essential to extract functional relationship between crop production and interrelated factors which may be nonlinear in nature. Therefore, a steady, robust and reliable algorithm is required for extraction of inherent hidden patterns present in the data. Few attempts have been made to solve the yield prediction problems even using deep learning technique. Convolutional neural network (CNN) and recurrent neural network (RNN) were used to predict soya bean yield[5]. In another study, CNN was used for crop yield prediction based on satellite images[6]. Deep neural networks (DNN) have multiple stacked nonlinear layers which transform the raw input data into higher and more abstract representation at each stacked layer.

Crop production of countries with similar atmospheric weather and soil conditions exhibit similar patterns. Generally, neighbouring countries are alike in terms of weather patterns and agricultural practice. They exhibit interdependencies on each other in terms of trading to overcome excess or deficit of crop production. Demand of a particular crop can be estimated by predicting future crop production of other countries. If it is forecasted to have deficit

*For correspondence. (e-mail: asha@csir4pi.in)

of production of a particular crop in several countries, then it usually creates more demand of that particular crop in international market. Therefore, based on the demand of that particular crop, farmers can be provided useful advisory to grow such crops. Supplying such demanded crops to neighbouring countries can be a better profitable business proposition instead of only catering to domestic requirement. This motivated us to draw our focus to design a predictive model for crop production of India in terms of neighbouring countries' crop production.

In this study, a new methodology that consists of three stages has been proposed and designed. In the first stage, an automated model is proposed to identify the best combination of crop production of different neighbouring countries that influence crop production of India, as input for the predictive model. In the next phase, using the above found best selected combination of countries' crop production as input, a predictive model is proposed to estimate crop production of India. Lastly proposed predictive model is validated with export–import data of India for that crop. The validation result showed that our proposed model was better suited for this problem. As rice is one of the leading food crops in the world, and the most consumed staple food in India, prediction of rice production has been considered as test case. The proposed model was evaluated using various statistical performance measures such as root mean square error (RMSE), mean absolute error (MAE) and correlation coefficient (CC).

Crop productions of twenty one Asian countries such as Bangladesh, Nepal, Sri Lanka, Pakistan, Myanmar, China Mainland, Iran, Thailand, Taiwan, Turkey, Malaysia, Bhutan, Indonesia, Iraq, Philippines, Cambodia, Afghanistan, Japan, Timor Leste, Brunei and India were considered for our study. For estimation of crop production of India, crop production of the rest of the twenty countries was considered to identify the optimal combination which was taken as input to the model. Detailed sensitivity analysis was carried out to identify the best combination of crop production of the above Asian countries for designing a suitable crop prediction model. Thus, an automated model was designed and implemented to find the best combination of inputs. Thereafter, the optimal combination of different countries' crop production obtained from the first level was considered as input to the predictive model.

Model data

Rice production data of the above listed Asian countries were collected from 'Food and Agriculture Organization'[7], which is a specialized agency of United Nations having goal of achieving food security for people worldwide. Total duration of data coverage for this study was from 1961 to 2016, which was divided into training and testing set. The data during 1961 to 1990 was considered as training set and the data between 1991 and 2016, as testing set.

Methodology

It is extremely important to design a robust methodology based on artificial neural network (ANN) which is capable of doing reliable forecasting by extracting inherent linear or nonlinear relationships among data.

Selection of model

In a prediction problem, it is important to select an appropriate model as a predictor. Generally, statistical models are efficient to deal with linear relationships of univariate data compared to multivariate data. On the other hand, machine learning algorithms are capable of handling the nonlinear as well as complex relationship[8] of data. Crop productions of all possible combination of countries are fed as input to our proposed model to predict crop production of India.

As the data used in this study was time series in nature, and since long short-term memory (LSTM) is best suited to handle time series data, we have used stacked LSTM, i.e. multiple LSTMs stacked together one after the other. Each layer in these LSTMs consists of multiple hidden neurons. Hidden neurons compute weighted inputs and generate output after passing through appropriate activation function. Epoch decides the number of times the model is trained with training data and plays an important role in training the model properly.

A typical LSTM (Figure 1) composes of two states which are the basic building blocks of a network, i.e. cell state (C_n) and hidden state (h_n). Similarly, it has three gates such as 'forget' gate (f_n), 'input' gate (i_n) and 'output' gate (o_n). These three gates are used to update C_n and h_n in every time stamp and the updated values of these two states are propagated to the next time stamp. These gates are responsible for different functionalities such as f_n is responsible for removing unwanted information, i_n is used for adding new useful information to C_n in every

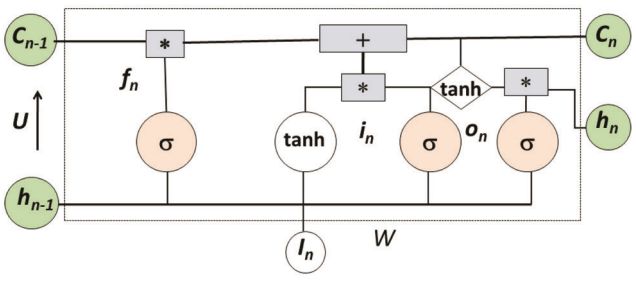

Figure 1. Architecture of a long short-term memory (LSTM) network[9].

time stamp. 'output' gate (o_n) helps in updating h_n in each time stamp by incorporating information from the updated cell. Above discussed three gates – f_n, i_n and o_n of LSTM are presented by the following three equations[9]

$$f_n = \sigma(W_f I_n + U_f h_{n-1} + b_f), \quad (1)$$

$$i_n = \sigma(W_i I_n + U_i h_{n-1} + b_i), \quad (2)$$

$$o_n = \sigma(W_o I_n + U_o h_{n-1} + b_o), \quad (3)$$

where I_n is input to LSTM network.

With help of values for f_n, i_n and o_n obtained from the above equations, C_n and h_n of LSTM were updated at each time stamp by the following equations.

$$C_n = f_n * C_{n-1} + i_n * \tanh(W_c I_n + U_c h_{n-1} + b_c), \quad (4)$$

$$h_n = \tanh(C_n) * o_n. \quad (5)$$

Weight matrices of current and previous time stamps are represented by W_f, W_i, W_o and W_c and U_f, U_i, U_o and U_c respectively. The b_f, b_i, b_o and b_c are the representation of bias vectors for the gates f_n, i_n, o_n and cell state C_n respectively. Also, h_{n-1} is the hidden state of previous state. σ and tan$h$ are the sigmoid and hyperbolic tangent activation functions respectively. Collaborative performance of these gates enables LSTM to work on time series data effectively.

The proposed crop model

The successful use of LSTM in various other domains as a time series predictor, motivated us to explore LSTM for designing a potential crop prediction model. Here, we have designed a predictive model for rice production of India using rice production of neighbouring countries and climate data of India. The proposed algorithm is divided into three components and detailed description of each component is depicted below.

Automated model for selection of optimal combination: Rice production of all the neighbouring countries may not influence significantly on rice production of India. In an artificial intelligence (AI) model, it is sufficient and required to determine optimal set of inputs that influences the output significantly, sufficiently and is non-avoidable. Adding extra inputs leads to decreased performance and thus should be excluded. It is crucial to consider the rice production of countries which influence the most on India's crop production. It is important to consider proper input to design an accurate predictive model in AI. Therefore, an automated model was designed to identify the best possible combination of crop production of neighbouring countries. All possible different combinations of rice production of the twenty countries were computed using mathematical combinations. From Table 1, we can see that different mathematical combinations generate different number of combinations of countries' crop production. The total number of combinations generated was ${}^{20}c_2 + {}^{20}c_3 \ldots {}^{20}c_{18} + {}^{20}c_{19}$, i.e. $2^{20} - ({}^{20}c_0 + {}^{20}c_1 + {}^{20}c_{20}) = 2^{20} - 22$. Computing these many combinations to determine the best combination is complex and resource intensive. Thorough sensitivity analysis which is conducted to identify the best combination of inputs is described below.

The model was trained with each of the combinations from total $2^{20} - 1$ combinations of crop production in neighbouring countries with data during 1961 to 1990, and subsequently tested with testing dataset from 1991 to 2016. Here, we have used LSTM as base model for designing automated model to determine optimal combinations of countries' crop production. Output set was generated corresponding to each input set and different error measures such as MAE, RMSE and CC were computed to verify efficiency of selected combinations. We also computed error ratio which was calculated by considering the ratio of test error and train error. Train errors are generated when model is tested with training data after training of the model. On the other hand, while the model was tested with testing dataset, we considered that error as testing error. Optimal combination was selected based on minimum error, minimum error ratio and maximum correlation coefficient. Block diagram of automated model for selecting the optimal input is presented in Figure 2.

Design of predictive model: Usually, in an AI approach, the forecast method consists of two stages. In the first stage, training data is used to learn the nonlinear relationships between the input and the output. In the second stage, testing data is used to validate the performance of the model. The model of our study was trained with the best combination of crop production of different countries obtained from sensitivity analysis and, model testing was performed with testing dataset between 1991 and 2016. In earlier studies[10–12], it has been described that the simulation of yield is influenced by weather dataset. Rice is mainly cultivated in irrigated and rainfed areas over South Asia. Rainfed rice cultivation depends only on seasonal rainfall. Lack of seasonal rainfall leads to water scarcity and extreme rainfall to damage crops, both lead to significant reduction in rice production. On the other hand in the irrigated region, the rice production depends on water availability through canal system or ground water. Studies have shown that there is direct correlation between annual summer rainfall and kharif rice yield[13,14]. Thus, crop model coupled with weather model can produce better result. Hence, rainfall information was incorporated to train our model further[15,16]. Here, we built three different modes of model by training the model with three different datasets. First one was the model trained with only rainfall dataset, second one was the model trained with data of crop production of best combination of neighbouring countries

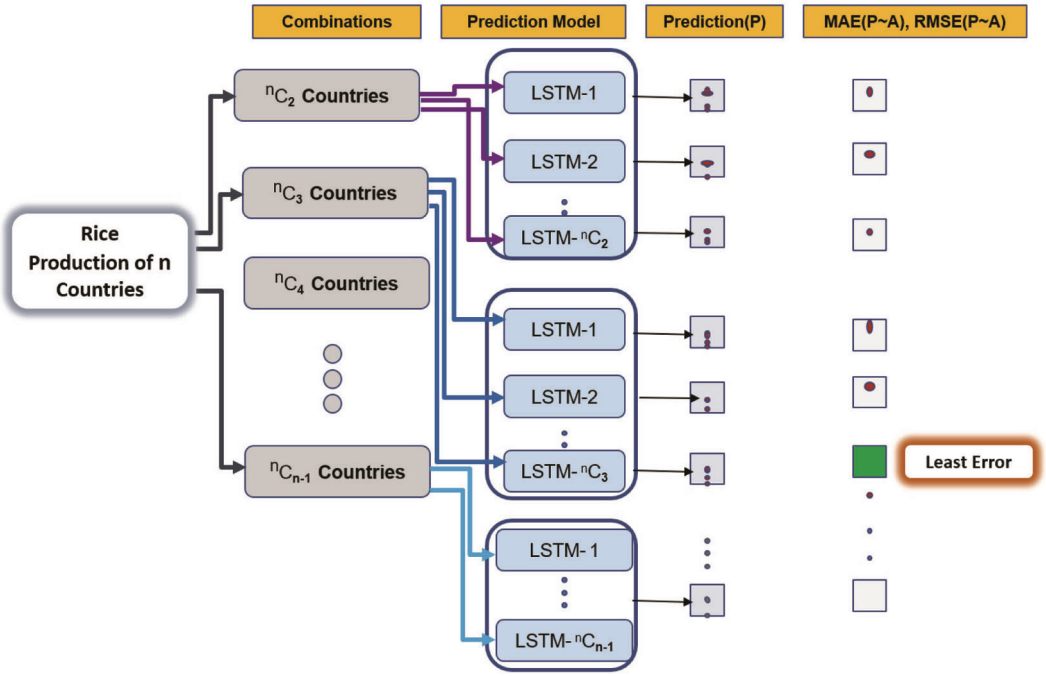

Figure 2. Design of an automated model for identifying best combination of rice production of neighbouring countries where P is predicted value and A is actual value.

Table 1. Different combinations of crop production of different Asian countries

$^{20}C_2 = 190$	$^{20}C_8 = 125970$	$^{20}C_{14} = 38760$
$^{20}C_3 = 1140$	$^{20}C_9 = 167960$	$^{20}C_{15} = 15504$
$^{20}C_4 = 4845$	$^{20}C_{10} = 184756$	$^{20}C_{16} = 4845$
$^{20}C_5 = 15504$	$^{20}C_{11} = 167960$	$^{20}C_{17} = 1140$
$^{20}C_6 = 38760$	$^{20}C_{12} = 125970$	$^{20}C_{18} = 190$
$^{20}C_7 = 77520$	$^{20}C_{13} = 77520$	$^{20}C_{19} = 20$

alone, and in the third scenario, the model was trained with both crop production and rainfall data.

In the final stage, proposed model was validated with trade matrix. Here, net flow of crop was calculated by subtracting the amount of import value from the amount of export value.

Experimental set-up

Number of hidden layers and hidden neurons in each hidden layer are fixed by evaluating training error. Model training starts with one hidden layer and ten hidden neurons on first layer. Number of neurons on each layer and number of layers are added gradually till optimal training loss is achieved. The model for the present study was optimally trained with eight hidden layers and 200 neurons in each layer with 200 epochs based on training validation loss. Other different hyper parameters such as loss function as MSE, optimizer as Adagrad with learning rate 0.0001 and validation split of 0.33 were used for training the model.

Result and analysis

We calculated the percentage of events, i.e. error in this case, which occurred in each error bin for different combinations of crop production. Error bins for both MAE and RMSE were computed by dividing the range of error equally. Figure 3 represents the RMSE error distribution in each of the combinations. It is evident from this figure that RMSE of most of the combination of countries' crop production lies between 0.1 and 0.2. Figure 4 represents MAE distribution for production of each combination of countries. It can be observed from Figure 4 that MAE of most of the combination of countries' crop production lies in the error bin of 0.1.

Sensibility analysis was conducted using mathematical combinations to obtain different combinations of crop production of neighbouring countries. Capability of this model was evaluated by performing MAE, RMSE for training data as well as testing data. Figure 5 *a* shows minimum normalized MAE for training and testing data of all combinations. Similarly, minimum normalized RMSE of training and testing data for all combinations are presented in Figure 5 *b*.

It can be observed from Figure 5 that values of MAE and RMSE are minimum for combination-8. Therefore, combination-8 of crop production was selected as input for the crop estimation model for crop production. The countries that were finally selected as the best combination were Bangladesh, Sri Lanka, Nepal, Myanmar, Pakistan, Thailand, Philippines and Iran.

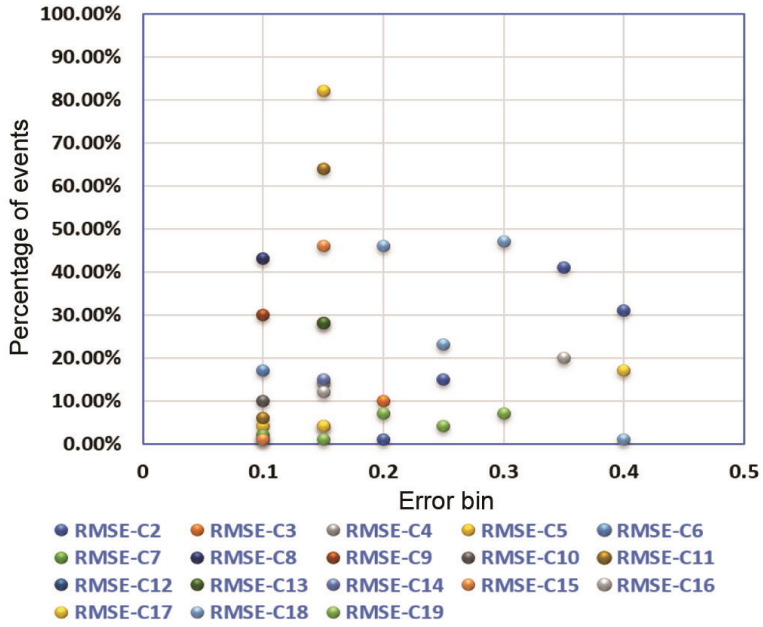

Figure 3. Error distribution of LSTM simulations with different combinations for identifying the minimum error model. *X* axis represents the normalized root mean square error (RMSE) and *Y* axis represents the percentage of simulations that fall within the error range. Combination 2, Combination 3, … Combination 19 are represented by C_2, C_3 … C_{19} respectively.

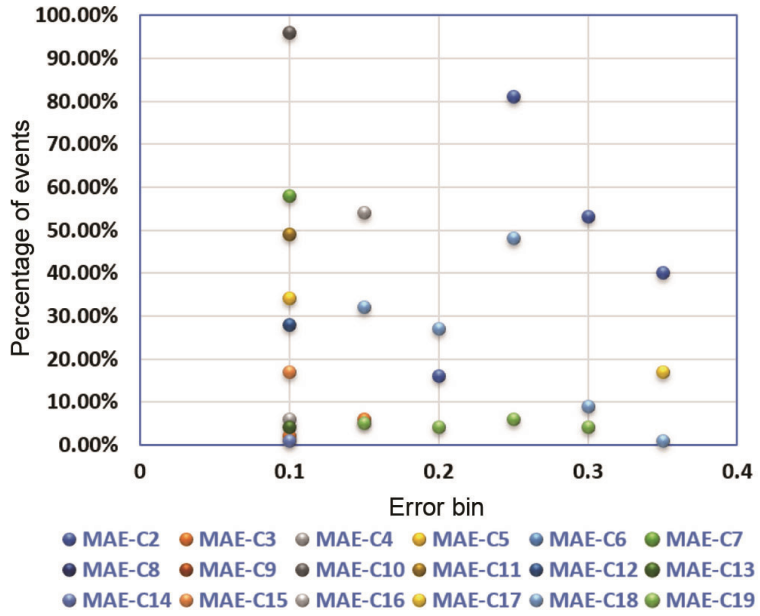

Figure 4. Error distribution of LSTM simulations with different combinations for identifying the minimum error model. *X* axis represents the normalized MAE and *Y* axis represents the percentage of simulations that fall within the error range. Combination 2, Combination 3, … Combination 19 are represented by C_2, C_3 … C_{19} respectively.

Model capability is depicted in Figure 6 by plotting predicted and observed crop production in three different scenarios. Figure 6 *a–c* represents the model capability when model was trained with only rainfall data, only crop production data, combination of rainfall and crop production data respectively. It is evident from Figure 6 that the prediction from the model trained with only rainfall data was not satisfactory. Whereas, model trained with only crop production data of neighbouring countries predicted well compared to the previous model, but it was unable to predict drought years, i.e. 2002 and 2004. On the other hand, model trained with both crop production

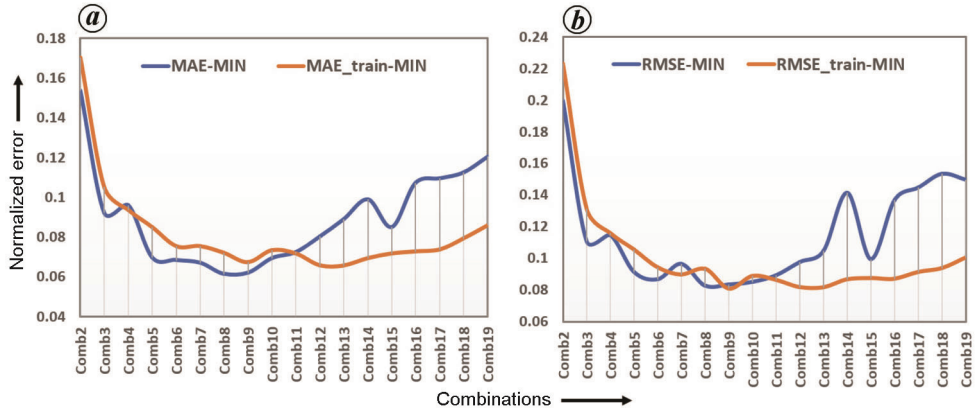

Figure 5. Optimal combinations of neighbouring countries' rice production for identifying the minimum error model. *X* axis represents different combinations of countries' rice production and *Y* axis represents (***a***) minimum normalized MAE and (***b***) minimum normalized RMSE of corresponding combinations of countries' rice production. MAE–MIN = Minimum normalized MAE obtained from each combination when trained model is tested with testing dataset. MAE_train–MIN = Minimum normalized MAE obtained from each combination when trained model is tested with training dataset. RMSE–MIN = Minimum normalized RMSE obtained from each combination when trained model is tested with testing dataset. RMSE_train–MIN = Minimum normalized RMSE obtained from each combination when trained model is tested with training dataset.

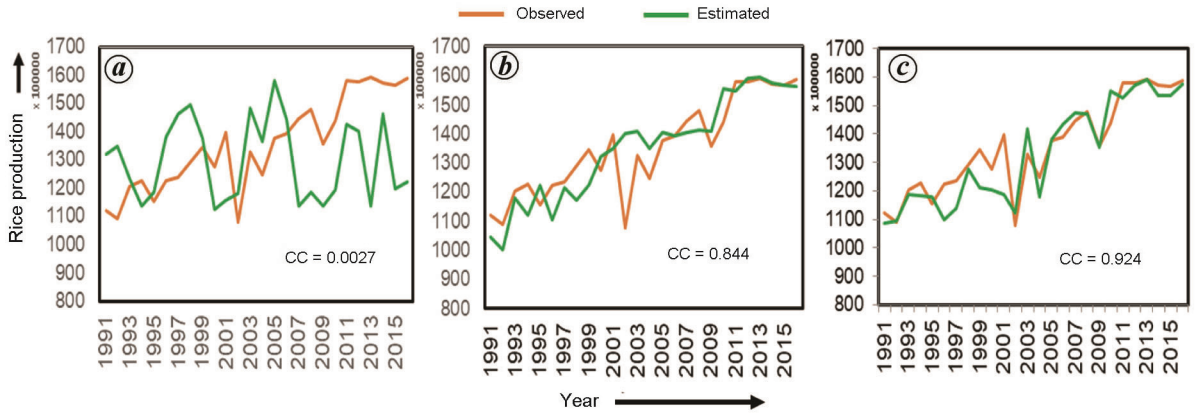

Figure 6. Estimated rice production with (***a***) only rainfall, (***b***) only rice production, (***c***) both rainfall and rice production. *X* axis represents time in year and *Y* axis represents rice production in tonnes.

Table 2. Summary of normalized mean absolute error (MAE), root mean square error (RMSE) and correlation coefficients (CC) between observed and estimated rice production at different scenarios

Training scenarios with different set of input data	MAE	RMSE	CC
With only rainfall data	0.370	0.425	0.0027
With only rice production data	0.157	0.291	0.844
With rice production and rainfall data	0.099	0.138	0.92

and rainfall data performed superior than both the previous models. It could successfully capture lower peak of observed data. Later on, the proposed model was validated with export–import data which is presented in Figure 7 that depicts the relationship among estimated and observed crop production, and net flow. It can be observed from Figure 7 that net flow is following the trend of estimated as well as observed crop production during 1991 to 1993, and 2005 to 2015. It can also be seen that during 1993 to 2004, trend of net flow and crop production are contrasting to each other, which might happen due to some other influencing factors. Model performance was evaluated using different statistical measures such as MAE, RMSE and CC. Summary of normalized MAE, RMSE and CC between the estimated and observed rice production are shown in Table 2. Thus, it is clear from the above analysis that third model provides better predictability, compared to other two models in terms of RMSE, MAE and CC.

Conclusion

In this study, research contributions were made in three aspects: First, detailed rigorous sensitivity analysis was performed to determine optimized input set for the model.

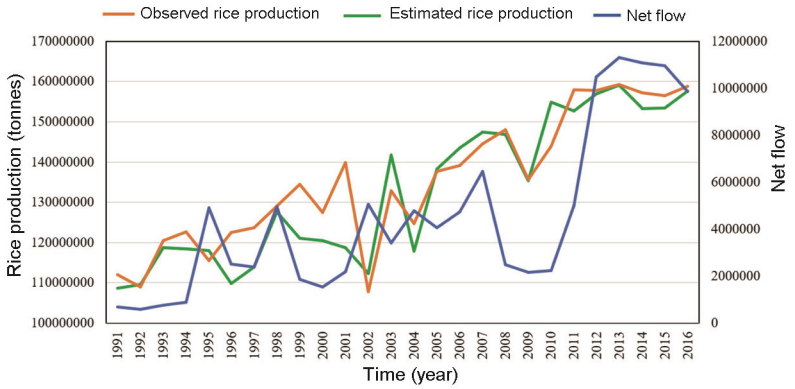

Figure 7. Relationship between rice production (estimated, observed) and net flow (export–import).

The second contribution included designing of the model for estimating crop production of a country and in this case India. Finally, the proposed model was validated with export–import data of India to verify the robustness of the proposed approach. The contribution from this study ensured feeding of correct input set to the model which was a crucial requirement in AI. This compelled us to carry out detailed sensitivity analysis to determine correct input which consisted of combination of crop production of neighbouring countries. As part of second contribution, for designing crop estimation model, along with crop production data, local weather data specifically rainfall data were incorporated. It has been well established in this research contribution that the designed LSTM trained with crop production and rainfall data, performed superior than the other two models described in the previous section. For related future studies, effect of soil can be included for crop growth to improve the capability of the proposed model. The novel method used to determine optimal set of inputs from group of unobvious set of inputs played the most important role in this study's approach, which can be used in other applications across domains. The prediction of crop production is highly useful for a country's food security, and determining this using state-of-art technology such as deep learning and more specifically, the proposed enhanced approach based on LSTM is a key innovation and can be adopted by countries' administrators and policy makers on food security.

1. Horie, T., Yajima, M. and Nakagawa, H., Yield forecasting. *Agric. Syst.*, 1992, **40**, 211–236; doi:10.1016/0308-521X(92)90022-G.
2. Carfagna, E. and Gallego, F. J., Using remote sensing for agricultural statistics. *Int. Stat. Rev.*, 2005, **73**(3), 389–404.
3. Awad, M. M., Toward precision in crop yield estimation using remote sensing and optimization techniques. *Agriculture*, 2019, **9**(3), 54; https://doi.org/10.3390/agriculture9030054.
4. Paliwal, A. and Jain, M., The accuracy of self-reported crop yield estimates and their ability to train remote sensing algorithms. *Front. Sustain. Food Syst.*, 2020; doi:10.3389/fsufs.2020.00025.
5. You, J., Li, X., Low, M., Lobell, D. and Ermon, S., Deep Gaussian process for crop yield prediction based on remote sensing data. In Thirty-First AAAI Conference on Artificial Intelligence, San Francisco, USA, 2017, pp. 4559–4566.
6. Russello, H., Convolutional neural networks for crop yield prediction using satellite images. IBM Center for Advanced Studies, Berelux, 2018.
7. http://www.fao.org/about/en
8. Sarker, I. H., Machine learning: algorithms, real-world applications and research directions. *SN Comput. Sci.*, 2021, **2**, 160; https://doi.org/10.1007/s42979-021-00592-x.
9. Hochreiter, S. and Schmidhuber, J., Long short-term memory. *Neural Comput.*, 1997, **9**(8), 1735–1780; doi:10.1162/neco.1997.9.8.1735.
10. Solow, A. *et al.*, The value of improved ENSO prediction to US agriculture. *Climatic Change*, 1998, **39**, 47–60.
11. Jones, J. W., Hansem, J. W., Royce, F. S. and Messina, C. D., Potential benefits of climate forecasting in agriculture. *Agric. Ecosyst. Environ.*, 2000, **82**, 169–184.
12. Hansen, J. W., Applying seasonal climate prediction to agriculture production. *Agric. Syst.*, 2002, **74**(3), 305–307.
13. Prasanna, V., Impact of monsoon rainfall on the total food grain yield over India. *Proc. Indian Acad. Sci.* (*Earth Planet. Sci.*), 2014, **112**, 529–558.
14. Rahman, M. A. *et al.*, Impacts of temperature and rainfall variation on rice productivity in major ecosystems of Bangladesh. *Agric Food Secur.*, 2017, **6**, 10; https://doi.org/10.1186/s40066-017-0089-5.
15. Baigorria, G. A., Jones, J. W., Shin, D. W., Mishra, A. and O'Brien, J. J., Assessing uncertainties in crop model simulations using daily bias-corrected regional circulation model outputs. *Clim. Res.*, 2007, **34**, 211–222.
16. Cantelaube, P. and Terres, J. M., Seasonal weather forecasts for crop yield modelling in Europe. *Tellus A: Dyn. Meteorol. Oceanogr.*, 2005, **57A**, 476–487.

ACKNOWLEDGEMENT. We thank the Head, CSIR-Fourth Paradigm Institute for encouragement and necessary support. We also thank National Mission on Himalayan studies for funding 'Enhancement of the quality of livelihood opportunities and resilience for the people in the Indian Himalayas, through design of intervention strategies aimed at maximizing resource potential and minimizing risks in urban–rural ecosystem' (NMHS-2017/MG-04/480).

Received 8 January 2021; revised accepted 23 August 2021

doi: 10.18520/cs/v121/i8/1073-1079