2020

Crop Yield Prediction Integrating Genotype and Weather Variables Using Deep Learning

Johnathon Shook
Iowa State University, jmshook@iastate.edu

Tryambak Gangopadhyay
Iowa State University, tryambak@iastate.edu

Linjiang Wu
Iowa State University

Baskar Ganapathysubramanian
Iowa State University, baskarg@iastate.edu

Soumik Sarkar
Iowa State University, soumiks@iastate.edu

See next page for additional authors

Crop Yield Prediction Integrating Genotype and Weather Variables Using Deep Learning

Abstract

Accurate prediction of crop yield supported by scientific and domain-relevant insights, can help improve agricultural breeding, provide monitoring across diverse climatic conditions and thereby protect against climatic challenges to crop production including erratic rainfall and temperature variations. We used historical performance records from Uniform Soybean Tests (UST) in North America spanning 13 years of data to build a Long Short Term Memory - Recurrent Neural Network based model to dissect and predict genotype response in multiple-environments by leveraging pedigree relatedness measures along with weekly weather parameters. Additionally, for providing explainability of the important time-windows in the growing season, we developed a model based on temporal attention mechanism. The combination of these two models outperformed random forest (RF), LASSO regression and the data-driven USDA model for yield prediction. We deployed this deep learning framework as a 'hypotheses generation tool' to unravel GxExM relationships. Attention-based time series models provide a significant advancement in interpretability of yield prediction models. The insights provided by explainable models are applicable in understanding how plant breeding programs can adapt their approaches for global climate change, for example identification of superior varieties for commercial release, intelligent sampling of testing environments in variety development, and integrating weather parameters for a targeted breeding approach. Using DL models as hypothesis generation tools will enable development of varieties with plasticity response in variable climatic conditions. We envision broad applicability of this approach (via conducting sensitivity analysis and "what-if" scenarios) for soybean and other crop species under different climatic conditions.

Keywords

deep learning, explainable, LSTM, attention, crop yield, impact of climate change

Disciplines

Agronomy and Crop Sciences | Artificial Intelligence and Robotics | Climate | Plant Breeding and Genetics

Comments

Authors

Johnathon Shook, Tryambak Gangopadhyay, Linjiang Wu, Baskar Ganapathysubramanian, Soumik Sarkar, and Asheesh K. Singh

Crop Yield Prediction Integrating Genotype and Weather Variables Using Deep Learning

Johnathon Shook[a,1]**, Tryambak Gangopadhyay**[b,1]**, Linjiang Wu**[b]**, Baskar Ganapathysubramanian**[b]**, Soumik Sarkar**[b, 2]**, and Asheesh K. Singh**[a, 2]

[a]Department of Agronomy, Iowa State University, Ames, IA, 50011
[b]Department of Mechanical Engineering, Iowa State University, Ames, IA, 50011
[1]J.S. and T.G. contributed equally to this work.
[2]To whom correspondence may be addressed. Email: singhak@iastate.edu, soumiks@iastate.edu.

ABSTRACT

Accurate prediction of crop yield supported by scientific and domain-relevant insights, can help improve agricultural breeding, provide monitoring across diverse climatic conditions and thereby protect against climatic challenges to crop production including erratic rainfall and temperature variations. We used historical performance records from Uniform Soybean Tests (UST) in North America spanning 13 years of data to build a Long Short Term Memory - Recurrent Neural Network based model to dissect and predict genotype response in multiple-environments by leveraging pedigree relatedness measures along with weekly weather parameters. Additionally, for providing explainability of the important time-windows in the growing season, we developed a model based on temporal attention mechanism. The combination of these two models outperformed random forest (RF), LASSO regression and the data-driven USDA model for yield prediction. We deployed this deep learning framework as a 'hypotheses generation tool' to unravel GxExM relationships. Attention-based time series models provide a significant advancement in interpretability of yield prediction models. The insights provided by explainable models are applicable in understanding how plant breeding programs can adapt their approaches for global climate change, for example identification of superior varieties for commercial release, intelligent sampling of testing environments in variety development, and integrating weather parameters for a targeted breeding approach. Using DL models as hypothesis generation tools will enable development of varieties with plasticity response in variable climatic conditions. We envision broad applicability of this approach (via conducting sensitivity analysis and "what-if" scenarios) for soybean and other crop species under different climatic conditions.

Keywords: deep learning, explainable, LSTM, attention, crop yield, impact of climate change

Introduction

One of the key challenges in plant breeding and crop production is to predict performance (seed yield) in unseen and new environments. This active research area is complicated by the time and expense of generating an extensive dataset to represent a wide range of genotypes and environments. Among different crops, soybean has a long history of cultivation in North America, with the first reported production in Georgia in 1766[1]. Over the years, production in the US and Canada has expanded longitudinally as far west as Kansas-Colorado border and latitudinally from southern Texas to Canada[2,3]. North American annual soybean yield trials (known as Uniform Soybean Tests (UST)) have been coordinated in the United States and Canada through the United States Department of Agriculture (USDA) between public breeders in university and government settings since 1941[4,5]. These trials are used to evaluate current and experimental varieties in multiple environments within their range of adaptation. Therefore, these trials are valuable sources of historical and current data to improve prediction performance with the assimilation of genetic and environmental variables. Management and permanent environmental effects have been examined primarily at small scales due to the labor required for managing large numbers of plots[6,7]. With the addition of each layer of added characterization of the environment, less of the differences need be ascribed to a generic "environmental" component, and can instead be examined individually in combination with plant genetics. The nexus of genetic and non-genetic variables form the cornerstone of plant breeding strategies, irrespective of crop species, for meeting crop production challenges in the future[8,9].

Climatic resiliency in cultivars is an important objective for plant breeders and farmers to get a high seed yield in a myriad of environments[10]. The climatic variability can be associated with changes in temperature and rainfall events (including patterns and magnitude) and other weather variables. In addition to spatial variability, temporal variability of weather variables[11] is equally important but generally less understood or not included in yield prediction studies. It is important to understand how agricultural production is affected by the variability of weather parameters in presence of global climate change, especially with

higher occurrence of extreme weather events. Therefore, prediction of the effects of changing environments on performance can help in making informed plant breeding decisions, marketing decisions, optimizing production and comparing results over multiple years[12].

Traditionally, crop growth models have been proposed to simulate and predict crop production in different scenarios including climate, genotype, soil properties, and management factors[13]. These provide a reasonable explanation on biophysical mechanisms and responses but have deficiencies related to input parameter estimation and prediction in complex and unforeseen circumstances[14]. Previous attempts at yield prediction across environments have relied on crop models generated by quantifying response in a limited number of lines while altering a single environmental variable, limiting the inference scope [15]. To bypass the limitations of crop growth models, linear models have also been used to predict yield with some success [16]. However, these low-capacity models typically rely on a rather small subset of factors, therefore failing to capture the complexity of biological interactions and more site-specific weather variable complexities. Traditional linear methods such as Autoregressive Integrated Moving Average (ARIMA) have been used for time series forecasting problems[17], but these methods are effective in predicting future steps in the same time-series. For time series prediction tasks, deep neural networks show robustness to noisy inputs and also have the capability to approximate arbitrary non-linear functions[18]. Deep learning models can provide solutions in the presence of such complex data comprising of different weather variables, maturity groups and zones, and genotype information.

Long Short Term Memory (LSTM) networks are very useful for time series modeling as they can capture the long-term temporal dependencies in complex multivariate sequences[19]. LSTMs have shown state-of-the-art results in various applications including off-line handwriting recognition[20], natural language processing[21] and engineering systems[22]. LSTMs have also been used effectively for multivariate time series prediction tasks[23–25]. Considering the importance of climate extremes for agricultural predictions, random forest has been utilized to predict grid-cell anomalies-deviations of yields[26]. Previous work[27] using deep learning for yield prediction has utilized multi-spectral images to predict yield (instead of leveraging only multivariate time series as input) without considering model interpretability. Khaki et al.[28] applied deep neural networks for yield prediction of maize hybrids using environmental data, but their model is not capable of explicitly capturing the temporal correlations and also lacks explainability. LSTM based model has been used for corn yield estimation[29], but these models lack interpretability. This study is based on geospatial data without field-scale farming management data and lacks temporal resolution in the absence of daily weather data. Attention based LSTM has been used along with multi-task learning (MTL) output layers[30] for county level corn yield anomaly prediction only based on meteorological data (maximum daily temperature, minimum daily temperature) without field-scale farming data. Other approaches to predict yield rely on the use of sensors to identify the most informative set of variables to predict yield[31,32], which is very useful in multiple applications; however, there is still a need to integrate weather parameters and in a time series approach involving multiple genotypes. Using these motivations, we developed a model that can capture the temporal variability of different weather variables across the growing season in an explainable manner to predict soybean yield from the UST dataset of field trials spanning 13 years across 28 states and provinces.

We propose a framework based on LSTM and temporal attention to predict crop yield with 30 weeks of weather data per year (over 13 years) provided as input, along with a reduced representation of the pedigree to capture differences in the response of varieties to the environment. We vary the number of input time-steps and compare the performance of our proposed Temporal Attention model with the Stacked LSTM model for two variations of each model. We also compared against the results of random forest (RF), LASSO regression and the data-driven state-of-the-art USDA model. The temporal attention mechanism highlights the significant time periods during the growing season leading to high or low yield prediction, concurred with domain knowledge. In this paper, we report improved fidelity interpretation of the prediction outcomes without sacrificing the accuracy for multivariate time-series prediction. Our proposed framework can have widespread applications in plant breeding, crop science research, and agricultural production.

Methods

Preparation of Performance Records

Files from 2003-2015 USTs were downloaded as PDFs[4,5]. Using on-line utility Zamzar (zamzar.com), all 26 PDFs from this period were converted to .xlsx files, with each tab corresponding to a single page in the file. In this way, the vast majority of tables were recovered with no errors or need for human translation. However, random checking for error was manually performed to ensure verity. These tables were manually curated to align all performance records for a given genotype/location combination into a single row. Records that did not have yield data (due to a variety not being planted in a specific location or dying prior to production of seed), were removed from the file.

Following data cleaning, the final dataset comprised of 103,365 performance records over 13 years representing 5839 unique genotypes, along with all available management information. After compilation, we imported performance records in Python for further data analysis.
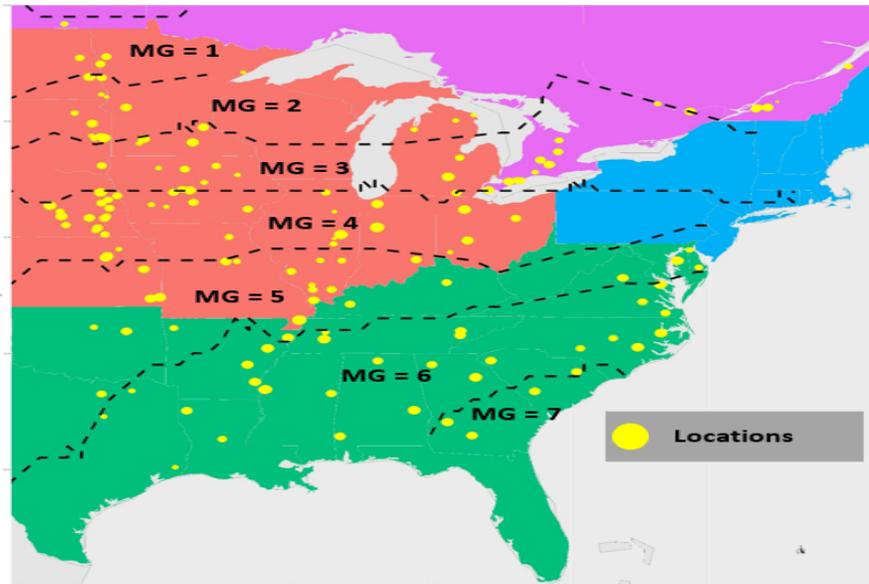
Figure 1. Map showing different locations in the USA and Canada included in our dataset. The dataset comprises of different maturity groups (MGs), some of which are labeled in the figure. The relative size of a yellow dot (representing location) indicates the size of the dataset for that particular location. Dataset included observations from the National Uniform Soybean Tests for years 2003-2015 and is split into North (MG 0 to 4) and South (MG 4 to 8) regions[33,34], consisting of 103,365 performance records over 13 years and 150 locations. These records are matched to weekly weather data for each location throughout the growing season (30 weeks). This generated a dataset with 35,000 plots having phenotype data for all agronomic traits.

Acquisition and Sub-Sampling of Weather Records

Daily weather records for all location/year combinations were compiled based on the nearest available weather station (25km grid) on Weather.com. We downsampled the dataset to include maximum, minimum, and average conditions on different time frames throughout the growing season (defined April 1 through October 31) and this information was appended to performance records.

Genotype Clustering

We included genotype-specific criteria to apply the model for specific genotypes and mean location yield across genotypes. Due to the nature of the UST program, most of the genotypes tested in this period do not have molecular marker data available, preventing the use of a G matrix. To circumvent these restrictions, we developed a completely connected pedigree for all lines with available parentage information, resulting in the formation of a 5839 x 5839 correlation matrix. To improve the model performance, genotypes were clustered based on the organization which developed them, providing additional control over relatedness.

We clustered genotypes in 5 clusters using the K-means Clustering technique based on the correlation matrix to extract information about relatedness. With a specified number of clusters (*n*), the K-means algorithm finds *n* groups of equal variance by choosing centroids of the clusters to minimize a criterion known as *inertia* (also called, within-cluster sum-of-squares). This algorithm is effective for a large number of samples and finds application across different domains. With this hard clustering technique, each genotype belongs to one of the 5 clusters. The clustering is used to represent each line as a function of membership in 5 groups, which is fed into the model to allow differentiation of lines.

Model Development

To leverage the temporal sequence of variables, a modeling approach based on recurrent neural network (RNN) was developed to capture correlation across time. Gradient descent of an error criterion may be inadequate to train RNNs especially for tasks involving long-term dependencies[35]. To overcome these challenges, long short-term memory (LSTM) was used, which is an RNN architecture designed to overcome the error backflow problems[36]. By learning long-range correlations in a sequence, LSTM can accurately model complex multivariate sequences[19].
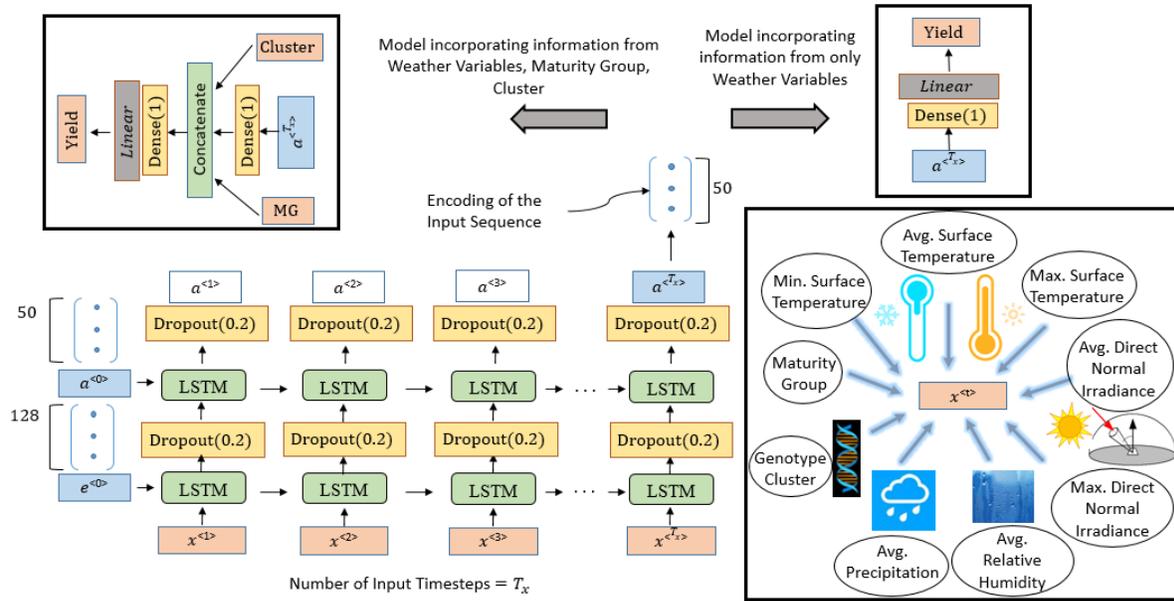
Figure 2. The figure showing the Stacked LSTM Model. The input feature vector is $x^{<t>}$ at time-step 't'. Depending on whether the maturity group and genotype cluster information are incorporated in the model or not, the vector $x^{<t>}$ can be 9-dimensional or 7-dimensional. We included 7 weather variables in our study. The embedding vector $a^{<T_x>}$ encodes the entire input sequence and summarizes the sequential dependencies from the time-step 0 to the time-step T_x. We designed two variants of our proposed model based on input information with the time series encoding part remaining the same for both variants. This model (when including MG, cluster) had 106,511 learnable parameters and the training time/epoch was 60 secs.

Figure 3. The figure showing the Temporal Attention Model. The LSTM encoding part is the same as that of the Stacked LSTM Model where we get the annotations $a^{<t>}$ for each timestep. Instead of only using $a^{<T_x>}$, this model utilizes all annotations which act as inputs for the temporal attention mechanism. Based on the computed context vector, the two variants of this model are designed depending on the input information. This model (when including MG, cluster) had 106,562 learnable parameters and the training time/epoch was 60 secs.

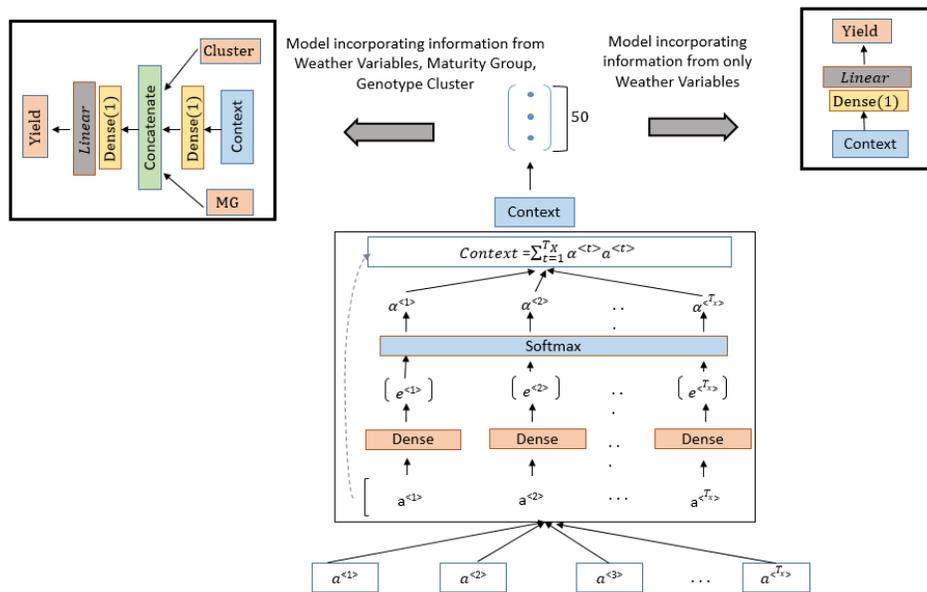We developed two models, based on LSTM: (a) Stacked LSTM Model (without using any attention) (Fig. 2), and (b) Temporal Attention Model (using a temporal attention mechanism) (Fig. 3). The output of both the models is yearly seed yield as this is a many-to-one prediction problem. For each model, we formulated the model variants depending on whether the performance records comprise data of maturity group and genotype cluster. The same modeling approach was used to compute the time-step wise encoding for both models. Two stacked LSTM layers were used to encode the T_x time-steps of the input sequence as shown in Fig. 2. Depending on the variant, for both models, we concatenated MG and genotype cluster values with the compressed time-series information.

In the Stacked LSTM Model, the last hidden state of the encoding part is assumed to be the compressed representation from the entire input sequence. This fixed-dimensional representation was used for predicting the output value of seed yield (Fig. 2). For the Temporal Attention Model, the compressed information (context) is computed after aggregating the information from the sequence of hidden states using the attention mechanism. The concept of soft temporal attention[37] was first proposed in the context of neural machine translation to overcome the bottleneck of the encoder-decoder model[21,38] for long sequences. Compressing all information from the input time-steps into a fixed-length single vector was the major bottleneck for the encoder-decoder model. Temporal attention can be applied for many-to-many time series prediction[24] and many-to-one-prediction[39,40]. The proposed approach (Fig. 3) does not incorporate a decoder LSTM as we are performing a many-to-one prediction problem. Taking in the annotations of all time-steps as input, the attention block aggregates the information and computes the context vector. A greedy search method was utilized to empirically determine the most influential weather variable on seed yield prediction considering data of both the northern and southern U.S. regions. In the first step of the greedy search, the Stacked LSTM model was trained for each of the 7 variables and choose the variable that had the least RMSE. With this variable added, in the second step, the model was trained for each of the other 6 variables. In this way, variables were added. More information is provided in the supplementary materials (Supplementary Tables 5, 6 and 7).

All input features were scaled in the range (-1, 1) with the scaler fitted on the training set. We compute the Root Mean Square Error (RMSE) after inverting the applied scaling to have forecasts and the actual values in the original scale. Data were randomly split into training (80%), validation (10%) and test (10%) sets. Models were evaluated by computing RMSE for the test set. Both models were trained for 200 epochs to get the optimal RMSE scores. For training, Adam optimizer was used[41] (learning rate of 0.001) and the mean squared error loss function was computed. Models were developed using Keras[42] with the TensorFlow backend[43] and the models were trained using NVIDIA GPUs.

Results

To select hyper-parameters (determination of appropriate temporal sampling of weather information to predict yield using our proposed frameworks), the test set RMSE was used to determine optimal (lowest RMSE) number of time points to predict seed yield. Using a step-wise approach building from monthly, bi-weekly, weekly and finally daily data, similar performance was observed in each scenario (approximate test RMSE = 7.206) except for daily data. The intermediate scenario of weekly data was picked for all subsequent analyses, to facilitate faster training of LSTMs and also not to downsample to a higher extent in capturing the long-range temporal dependencies.

Using weekly weather aggregate data in our model, the prediction models were built starting with a heuristic variable importance given to each variable. For example, precipitation was deemed to be most important followed by average surface temperature and so on. However, the largest drop in test RMSE was observed for the maturity group when it was used as a predictor factor in the model and adding the MG classification after the 2nd LSTM as well caused a further improvement in model performance. No perceptible change in performance was observed with variation of the number of clusters (5, 10, 15, 20, 25) using the hard clustering technique (K-means clustering). Therefore, subsequent analyses are done using 5 clusters in the proposed models for prediction and variable search. Adding the genotype cluster information at every time-step and also after the 2nd LSTM, showed better results.

From our greedy search, we observed average relative humidity had the lowest test RMSE. With the inclusion of average relative humidity in the prediction model, average direct normal irradiance was the next most important variable. Sequentially, the remaining weather variables were: maximum direct normal irradiance, maximum surface temperature, minimum surface temperature, average surface temperature, and average precipitation. A second greedy search initiated with the inclusion of maturity group and pedigree-based clustering revealed minimum surface temperature as the most important weather variable (lowest RMSE). The greedy search results revealed the following sequence of weather variable importance obtained from a forward selection approach: average direct normal irradiance, average surface temperature, maximum direct normal irradiance, average precipitation, average relative humidity, and maximum surface temperature. Noticeably, the ranking of the variables was different but the absolute change in RMSE scores was minimal.

Overall, a correlation of 0.894 between predicted and observed yields in the testing and validation sets was attained; largely capturing the differences in performance between environments and years. However, the model remains somewhat limited in its ability to generate genotype-specific yield predictions due to the limited complexity of relationships which can be modeled
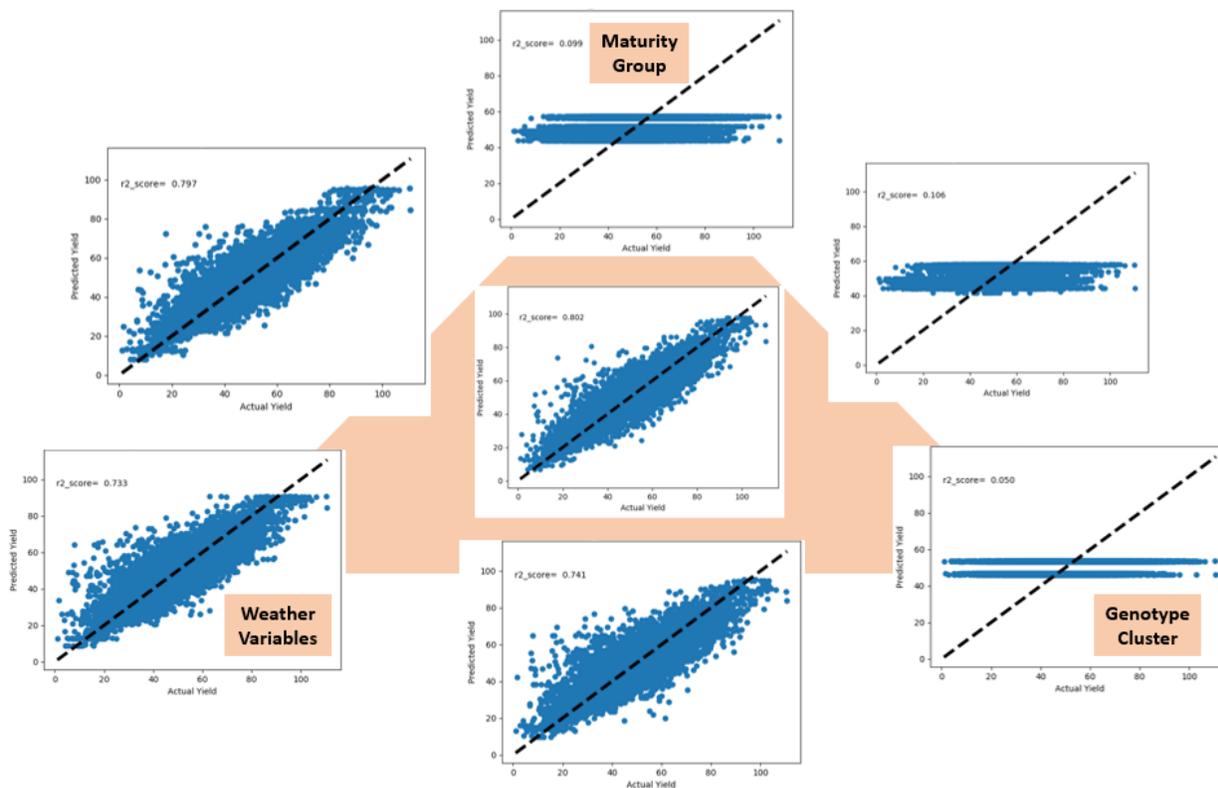
Figure 4. Results for different inputs to the Stacked LSTM model. The vertices of the triangle demonstrate results including only the maturity group, only genotype cluster and only weather variables in the input. The edges show the results with a combination of inputs from the respective vertices. The results showed improvement when the genotype cluster was included with weather variables. The coefficient of determination increased further when the maturity group was included with weather variables. The best result were noticed when information from all sources was incorporated (shown at the center of the triangle). The best performance (RMSE = 7.130) is about 14% of the average seed yield for the test set (50.745) and 44.5% of the standard deviation (16.019).

using LSTM, and a lack of genomic information on each genotype. Since, a lack of molecular marker data for each line precludes us to leverage genomic prediction and its integration with the LSTM model, it is the next step of the approach presented in our paper. As currently implemented, the model's average absolute error is 5.4 bu/acre, which is reasonable given the levels of variability within a given environment/year combination. For example, in Ames, IA, during 2003, yields ranged from 33.3-55.3 bu/acre. In spite of this large range of difference, an average error of only 4.5 bu/ac was observed for this environment. No perceptible trends are observed when we looked at state wide results combined over years. We also looked at originating breeding state as well as private company entries, and no geographical trends were noticeable.

Both proposed models (Stacked LSTM, Temporal Attention) showed similar performance, and results improved when more information were included (Fig. 4). The coefficient of determination was highest (0.802) when information from all the sources (maturity group, genotpye cluster, weather variables) were incorporated. The best model performance (test RMSE = 7.130) was $\tilde{1}$4% of the average yield for the test set (50.745) and 44.5% of the standard deviation (16.019)(Fig. 4). Comparatively, test RMSE of 12.779 was obtained from least absolute shrinkage and selection operator (LASSO) regression, while Random Forest test RMSE was 9.889 with same input features. Therefore, both Stacked LSTM and Temporal Attention models outperform LASSO and RF models.

In comparison with the data-driven state-of-the-art USDA model[16], our deep learning approach performs significantly better demonstrating much lower absolute errors. The USDA approach uses a linear regression approach with coefficients based on historical statewide yields and weather averages. However, the USDA model does not predict performance for individual locations. Due to this limitation, we compare results of our model with the USDA model using year wise average across states for the test set. In comparison with the USDA model, the absolute errors of our model are lower for all 12 years (except in 2011). For 2014 and 2015, the absolute errors of deep learning models were 0.03 and 0.35 (compared to 1.32 and 1.70 for the USDA model), respectively. Detailed comparison results are provided in the supplementary material (Supplementary Table 10).
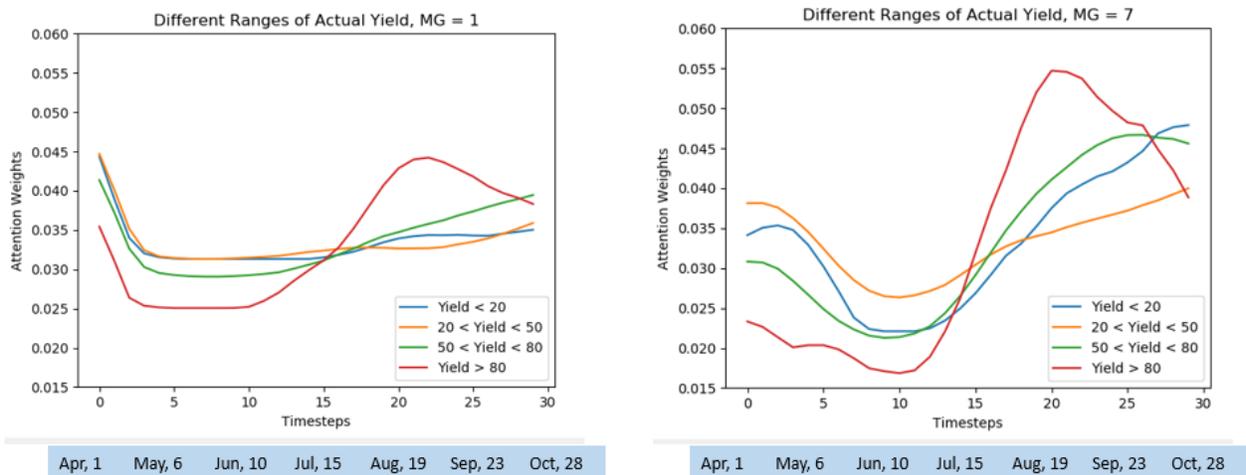

Figure 5. Results showing the distribution of attention weights for the entire input sequence (spanning the growing season). Considering different ranges of actual yield, the results are demonstrated for two different maturity groups (MG = 1, MG = 7) providing stark geo-climatic regions (Fig. 1). Early season variables were observed to be comparatively less important for prediction of the highest yielding genotypes.

In addition to accurate yield prediction, the Temporal Attention Model provided insights (Fig. 5) about how early-season variables were less important for yield prediction in the highest yielding genotypes for two geographically distinct maturity groups: MG1 (Northern US adaptation) and MG7 (Southern US adaptation). We observed mild sigmoid curves for the highest yielding group in the case of both MG1 and MG7. However, we note that while MG1 had a significantly large number of plots ($\approx$ 550) for the highest yielding group, MG7 had only about 30 such plots. It points to the increasing importance of features in the August – September time phases for both North and South US regions. These time phases coincide with crop reproductive phases, emphasizing their importance in the final yield, and need functional validation which is outside of the scope of our research. However, this is an example of hypotheses generation advantage of these models motivating future research.

Discussion

We establish the potential for use of a long short-term memory-based method for yield prediction to allow models to account for temporal differences in the occurrence of weather events. Predictions using this system can be made reasonably accurate due to a large amount of training data made available through the mining of historical records. Our approach (using LSTM and attention) is an efficient modeling scheme to analyze soybean crop growth interaction with the weather, and to identify hypothesis for plasticity, as well as to identify key physio-environmental features that are important to include into any predictive model. For example, differences in the timing of extreme heat events, as well as drought periods, would affect soybean plants in various ways depending on the stage of plant development. For example, heat stress during flowering is particularly damaging while heat in vegetative stages of development may not produce significant reduction to harvested yield [44]. With a larger encompassing dataset, breeders and researchers can be empowered to parse out most informative time periods, weather variables and crop responses. This information sets up the framework for breeding strategies to develop climate resilient and responsive varieties.

Our results – via our hypothesis generation approach – show a potential mismatch in the heuristic/empirical results for the importance of weather variables. The finding of minimum surface temperature as the most significant weather variable suggests that nighttime temperatures play a larger role in yield prediction than previously suggested[45]. Our study is a retrospective design, and cannot conclude definitively that this is the case; however, these findings necessitate further empirical investigations and can be used to formulate the next set of hypotheses. Our findings are significant, as minimum temperatures have been reported to be increasing at a faster rate than maximum temperatures[46]. More studies are needed to ascertain the relative importance of these variables and can motivate morpho-physiological attentive breeding approaches to assemble sturdier varieties for future scenarios.

A large capacity machine learning approach, such as the one presented in this paper using LSTM-RNN will be robust to incorporate weather changes and adjust performance predictions accordingly. Additional information that may improve the results of this approach is the inclusion of any supplemental irrigation provided, soil fertility levels, disease pressure and resistance levels, and direct genetic markers for the tested varieties, all of which would further strengthen predictive ability. Therefore, future implementations may be expanded to include genomic data, additional factors such as preceding crop, row spacing, planting date, soil texture, or additional temporal data in the forms of soil sensor measurements and remote sensing data for morphological and physiological traits. The approach presented in this work will further enhance phenomic assisted breeding that collects in-season data using different sensors and payloads[31,32,47] using machine and deep learning approaches suitable in plant sciences applications[48–50].

Our work shows a unique strategy to assimilate and utilize complex data for seed yield prediction. For comparative purposes, we compared our models with the RF, LASSO and the data-driven USDA model. The USDA model has a limitation on the type of data it can utilize and is limited in its application. For example, as the USDA model computes predictions at the state level, the finer resolution available with our model may help in making regional marketing decisions, as well as in creating yield predictions which can capture intra-state variation due to factors such as differences in rainfall in different areas of the state. Since our results are built on more than a decade of data, it also reflects that early season weather variables are less useful in seed yield prediction and needs empirical evidence to confirm the genetic variability in plasticity of soybean genotypes in earlier stages of growth and development. Importantly, we emphasize that the utilization of the attention module within a LSTM framework allows us to tease out potentially important features for further testing. This alleviates the disadvantage of DL models – which serve as purely blackbox predictive models – by allowing for hypothesis generation that will allow scientific insight via targeted follow up analysis and experiments.

The advantages of LSTM based models have been recently established for maize yield prediction at a county level[29], but the model lacked interpretability. Attention based LSTM along with multi-task learning (MTL) output layers has also been used for maize yield prediction using county level data based on meteorological data (maximum daily temperature, minimum daily temperature, and daily precipitation)[30]. These studies are important for solving the yield prediction challenge; however, models are based on geospatial data without field-scale farming management data and variety information is indiscernible, and based on limited weather variables. In our soybean study, we included seven weather variables and detailed field-scale farming data with multiple maturity groups spanning continental U.S. and full variety representation.

We have shown that an LSTM-based approach can improve seed yield prediction accuracy due to the ability to identify both temporal effects of weather events and the relative importance of various weather variables for crop yield prediction. Advances in developing an explainable yield prediction model using attention mechanism is an attractive development. The basic framework of LSTM for the phenotypic prediction can be applied to any crop with weather-dependent variability in order to better understand the genotype x environment effects found in the course of multi-environment testing. As such, this approach can be immediately useful for researchers in a variety of crops and environments and may prove to be exceptionally powerful when used in collaborative efforts between researchers operating in contrasting climatic zones, and in conjunction with sensor data for prescriptive breeding[32] including for root traits[51]. The insights provided by our model can help in understanding

the impact of weather variability on agricultural production in the presence of climate change, and devise breeding strategies for variety plasticity to circumvent these climatic challenges.

The ability to make accurate predictions of crop performance can lead to optimization across many different levels of organizations. At the federal level, improved crop insurance recommendations can be made based on weather forecasts before planting, and be continually updated throughout the season as more data is recorded and forecasts are updated. Railroads, grain cooperatives, and end-users can streamline the logistics of handling the desired quantities of grain if they are permitted a better understanding of how much grain (and of what quality) will be produced in a given region. Farmers can make better marketing decisions if they have an accurate and high confidence prediction of their production for the year, allowing them to sell their crops at the most opportune time. We envision that similar work on other crops and over a longer time span will generate invaluable insights for cultivar development and plant breeding and production related research in a challenging climate.

Conclusion

Unraveling causality would be a substantial step forward in understanding impact of climate change on variety's plasticity. Viewed through the lens of causality, DL based predictive models vs process based predictive models have distinct pros and cons. Process based models have clear causal relationships (by construction); however causality is limited to the confines of the model parameters, and it is non-trivial to assimilate additional data to extract broader causal trends. On the other hand, incorporating causality into DL based models is an open problem in the AI/ML community, with much activity. No principled approaches exist to accomplish this. However, DL based models (in contrast to process-based models) have the ability to seamlessly assimilate additional data. Our vision is therefore to evaluate if systematically augmenting DL based predictive models with increasing amounts of physio-morphological informative features provides a way towards unraveling causal relationships. We accomplish this by deploying our DL framework as a 'hypotheses generation tool'. We build DL models using a large volume of data and variety of information incorporating domain based knowledge. We then systematically probe the impact of various physio-morphological and environmental parameters on yield (via sensitivity analysis, and "what if" scenario evaluation), and establish a framework to generate hypotheses in different crop species and physio-morphological characteristics under different climatic conditions. Until causality based DL becomes feasible, the hypotheses generation DL models will have the maximum impact in meeting the need of climate change scenarios and to incorporate plasticity response in future varieties.

Acknowledgements

Author contributions statement

A.K.S., J.S., S.S. and B.G. conceived the research; All authors contributed in the design of the analysis and interpretation; J.S. compiled the UST performance and pedigree data; T.G. and J.S. performed statistical analysis, T.G. and L.W. built machine learning models and results were interpreted by T.G. and J.S. with inputs from S.S., A.K.S., and B.G.; J.S. and T.G. wrote the first draft with inputs from A.K.S. and S.S.; All authors contributed to the development of the manuscript.

References

1. Hymowitz, T. & Harlan, J. R. Introduction of soybean to north america by samuel bowen in 1765. *Econ. Bot.* **37**, 371–379 (1983).

2. Soybeans: Planted acreage by county. https://www.nass.usda.gov/Charts_and_Maps/Crops_County/sb-pl.php. Accessed: 2020-05-06.

3. Canadian soybean seeded acres (1980 to current). http://soycanada.ca/statistics/seeded-area-acres/. Accessed: 2020-05-06.

4. Uniform soybean tests, northern region. https://www.ars.usda.gov/midwest-area/west-lafayette-in/crop-production-and-pest-control-research/docs/uniform-soybean-tests-northern-region/. Accessed: 2020-05-06.

5. Uniform soybean tests. https://www.ars.usda.gov/southeast-area/stoneville-ms/crop-genetics-research/docs/uniform-soybean-tests/. Accessed: 2018-01-10.

6. Zhang, L., Zhu, L., Yu, M. & Zhong, M. Warming decreases photosynthates and yield of soybean [glycine max (l.) merrill] in the north china plain. *The Crop. J.* **4**, 139–146 (2016).

7. Puteh, A. B. *et al.* Soybean [glycine max (l.) merrill] seed yield response to high temperature stress during reproductive growth stages. *Aust. J. Crop. Sci.* **7**, 1472 (2013).

8. Lenaerts, B., Collard, B. C. & Demont, M. Improving global food security through accelerated plant breeding. *Plant Sci.* **287**, 110207 (2019).

9. Hickey, L. T. *et al.* Breeding crops to feed 10 billion. *Nat. Biotechnol.* **37**, 744–754 (2019).

10. Durrell, J. Pathways to impact for building thriving and resilient communities in dry areas. Tech. Rep., International Center for Agricultural Research in the Dry Areas, Beirut, Lebanon (2017).

11. Melillo, T. T. R., Jerry M. & Gary W. Yohe, E. Climate change impacts in the united states: The third national climate assessment. https://nca2014.globalchange.gov/report/sectors/agriculture (2014).

12. Jagtap, S. S. & Jones, J. W. Adaptation and evaluation of the cropgro-soybean model to predict regional yield and production. *Agric. ecosystems & environment* **93**, 73–85 (2002).

13. Blanc, É. Statistical emulators of maize, rice, soybean and wheat yields from global gridded crop models. *Agric. For. Meteorol.* **236**, 145–161 (2017).

14. Roberts, M. J., Braun, N. O., Sinclair, T. R., Lobell, D. B. & Schlenker, W. Comparing and combining process-based crop models and statistical models with some implications for climate change. *Environ. Res. Lett.* **12**, 095010 (2017).

15. Bishop, K. A., Leakey, A. D. & Ainsworth, E. A. How seasonal temperature or water inputs affect the relative response of c3 crops to elevated [co2]: a global analysis of open top chamber and free air co2 enrichment studies. *Food Energy Secur.* **3**, 33–45 (2014).

16. Westcott, P. C. & Jewison, M. Weather effects on expected corn and soybean yields. *USDA ERS* **FDS-13g-01** (2013).

17. Petrică, A.-C., Stancu, S. & Tindeche, A. Limitation of arima models in financial and monetary economics. *Theor. & Appl. Econ.* **23** (2016).

18. Dorffner, G. Neural networks for time series processing. In *Neural network world*, vol. 6, 447–468 (1996).

19. Malhotra, P., Vig, L., Shroff, G. & Agarwal, P. Long short term memory networks for anomaly detection in time series. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium)*, 89 (Presses universitaires de Louvain, 2015 April 22-24).

20. Doetsch, P., Kozielski, M. & Ney, H. Fast and robust training of recurrent neural networks for offline handwriting recognition. In *Frontiers in Handwriting Recognition (ICFHR), 2014 14th International Conference on*, 279–284 (IEEE, 2014).

21. Sutskever, I., Vinyals, O. & Le, Q. V. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, 3104–3112 (2014).

22. Gangopadhyay, T., Locurto, A., Michael, J. B. & Sarkar, S. Deep learning algorithms for detecting combustion instabilities. In *Dynamics and Control of Energy Systems*, 283–300 (Springer, 2020).

23. Jiang, Z. *et al.* Predicting county level corn yields using deep long short term memory models. *arXiv preprint arXiv:1805.12044* (2018).

24. Gangopadhyay, T., Tan, S. Y., Huang, G. & Sarkar, S. Temporal attention and stacked lstms for multivariate time series prediction. In *NeurIPS 2018 Workshop on Modeling and Decision-Making in the Spatiotemporal Domain* (NeurIPS, 2018).

25. Shook, J. M. *et al.* Integrating genotype and weather variables for soybean yield prediction using deep learning. *bioRxiv* 331561 (2018).

26. Vogel, E. *et al.* The effects of climate extremes on global agricultural yields. *Environ. Res. Lett.* **14**, 054010 (2019).

27. You, J., Li, X., Low, M., Lobell, D. & Ermon, S. Deep gaussian process for crop yield prediction based on remote sensing data. In *Thirty-First AAAI Conference on Artificial Intelligence* (2017).

28. Khaki, S. & Wang, L. Crop yield prediction using deep neural networks. In *INFORMS International Conference on Service Science*, 139–147 (Springer, 2019).

29. Jiang, H. *et al.* A deep learning approach to conflating heterogeneous geospatial data for corn yield estimation: A case study of the us corn belt at the county level. *Glob. change biology* (2019).

30. Lin, T. *et al.* Deepcropnet: a deep spatial-temporal learning framework for county-level corn yield estimation. *Environ. Res. Lett.* **15**, 034016 (2020).

31. Parmley, K. *et al.* Development of optimized phenomic predictors for efficient plant breeding decisions using phenomic-assisted selection in soybean. *Plant Phenomics* **2019**, 5809404 (2019).

32. Parmley, K. A., Higgins, R. H., Ganapathysubramanian, B., Sarkar, S. & Singh, A. K. Machine learning approach for prescriptive plant breeding. *Sci. Reports* **9**, 1–12 (2019).

33. Gillen, A. M. & Shelton, G. W. Uniform soybean tests southern states 2018. *USDA-ARS* (2018).

34. Nowling, G. & Cai, G. Uniform soybean tests northern region 2018. *USDA-ARS* (2018).

35. Bengio, Y., Simard, P., Frasconi, P. *et al.* Learning long-term dependencies with gradient descent is difficult. *IEEE transactions on neural networks* **5**, 157–166 (1994).

36. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural computation* **9**, 1735–1780 (1997).

37. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

38. Cho, K. *et al.* Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

39. Gangopadhyay, T., Tan, S. Y., Locurto, A., Michael, J. B. & Sarkar, S. An explainable framework using deep attention models for sequential data in combustion systems. In *NeurIPS 2019 Workshop on Machine Learning and the Physical Sciences* (NeurIPS, 2019).

40. Gangopadhyay, T. *Deep learning for monitoring cyber-physical systems*. Master's thesis, Iowa State University (2019).

41. Kingma, D. P. & Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

42. Chollet, F. *et al.* Keras (2015).

43. Abadi, M. *et al.* Tensorflow: A system for large-scale machine learning. In *OSDI*, vol. 16, 265–283 (2016).

44. Westgate, M. & Peterson, C. Flower and pod development in water-deficient soybeans (glycine max l. merr.). *J. Exp. Bot.* **44**, 109–117 (1993).

45. Gibson, L. & Mullen, R. Influence of day and night temperature on soybean seed yield. *Crop. Sci.* **36**, 98–104 (1996).

46. Karl, T. R. *et al.* Asymmetric trends of daily maximum and minimum temperature. *Pap. Nat. Resour.* 185 (1993).

47. Gao, T. *et al.* A novel multirobot system for plant phenotyping. *Robotics* **7**, 61 (2018).

48. Singh, A., Ganapathysubramanian, B., Singh, A. K. & Sarkar, S. Machine learning for high-throughput stress phenotyping in plants. *Trends plant science* **21**, 110–124 (2016).

49. Singh, A. K., Ganapathysubramanian, B., Sarkar, S. & Singh, A. Deep learning for plant stress phenotyping: trends and future perspectives. *Trends plant science* **23**, 883–898 (2018).

50. Ghosal, S. *et al.* An explainable deep machine vision framework for plant stress phenotyping. *Proc. Natl. Acad. Sci.* **115**, 4613–4618 (2018).

51. Falk, K. G. *et al.* Computer vision and machine learning enabled soybean root phenotyping pipeline. *Plant methods* **16**, 5 (2020).

52. Gers, F. A., Schmidhuber, J. & Cummins, F. Learning to forget: Continual prediction with lstm. In *Ninth International Conference on Artificial Neural Networks, ICANN 1999*, vol. 2, 850–855 (IET, 1999).

53. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R. & Schmidhuber, J. Lstm: A search space odyssey. *IEEE transactions on neural networks learning systems* **28**, 2222–2232 (2017).

Supplementary Materials

Clustering

Clustering is an unsupervised machine learning technique used to group unlabeled examples. A metric (similarity measure) is used to estimate the similarity between examples by combining the examples' feature data. With the increase in the number of features, the similarity measure computation can become more complex. By assigning a number to each cluster, each complex example is represented by a cluster-ID. This makes clustering a simple yet powerful technique that finds applications in domains including image segmentation, anomaly detection, social network analysis, and medical imaging. The output of the clustering technique (Cluster ID) can be then used as input instead of a high-dimensional feature for machine learning algorithms.

The choice of a clustering algorithm depends on whether it can scale efficiently to the available dataset. The clustering algorithms that compute the similarity between all pairs of examples, are not practical to be used for a large number of examples (n) as the runtime for this type of algorithm is proportional to the square of n. K-means clustering algorithm scales linearly with n and thus can be used for large-scale data. While the centroid-based clustering algorithm organizes data into non-hierarchical clusters, density-based clustering works by connecting highly dense areas into clusters. Density-based clustering is not suitable for data with high-dimensions and distribution-based clustering is not applicable when the data-distribution type is not known. Hierarchical clustering, which computes a tree of clusters is mostly meant for hierarchical data. This leads to the choice of a centroid-based algorithm for clustering the genotypes represented by a correlation matrix. K-means clustering is simple, efficient and the most commonly used centroid-based clustering algorithm. We implemented the K-means clustering algorithm in this work for these reasons.

Modeling Approach

Long Short Term Memory Networks (LSTMs)

Recurrent Neural Networks (RNNs) can explicitly capture temporal correlations in time series data, and efficient learning of the temporal dependencies leads to highly accurate prediction and forecasting, often outperforming static networks. Deep RNNs are trained using the error backpropagation algorithm; however, the propagation of error gradients through the latent layers and unrolled temporal layers suffer from the vanishing gradient problem. Therefore, gradient descent of an error criterion may be inadequate to train RNNs especially for tasks involving long-term dependencies[35]. Moreover, standard RNNs fail to learn in the presence of time lags greater than 5-10 discrete time-steps between relevant input events and target signals [52]. To overcome these challenges, Long short-term memory (LSTM) was used, which is an RNN architecture designed to overcome the error backflow problems[36]. By using input, output and forget gates to prevent the memory contents being perturbed by irrelevant inputs and outputs, LSTM networks have the ability in learning long-range correlations in a sequence and can accurately model complex multivariate sequences[19].

The cell state in an LSTM block can allow the information to just flow along with it unchanged and information can be added to or removed from the cell state. In an LSTM block, there are input, output and forget gates that prevent the perturbation of the memory contents with irrelevant information. These gates regulate the augmentation of any information to the cell state. An overview of the LSTM block is demonstrated in Fig. 6.

Forget gate layer is the first step of a LSTM block. A sigmoid layer (σ_1) decides the information to be removed the cell state of the previous time-step $C^{<t-1>}$. Forget gate naturally permits LSTM to learn local self-resets of memory contents that have become irrelevant [52]. This is performed as:

$$e_t = \sigma_1(W_e \cdot [a^{<t-1>}, x^{<t>}] + b_e) \tag{1}$$

The next step is to augment the cell state with new information. A sigmoid layer(σ_2, the input gate layer) followed by a tanh layer generates potential new information $\tilde{C}^{<t>}$ for augmentation.

$$r_t = \sigma_2(W_r \cdot [a^{<t-1>}, x^{<t>}] + b_r) \qquad \tilde{C}^{<t>} = tanh(W_C \cdot [a^{<t-1>}, x^{<t>}] + b_C) \tag{2}$$

Thereafter, the new cell state $C^{<t>}$ is obtained as follows:

$$C^{<t>} = e_t \times C^{<t-1>} + r_t \times \tilde{C}^{<t>} \tag{3}$$

Thereafter, the hidden state of the previous time-step $a^{<t-1>}$ is passed through the third sigmoid layer (σ_3) for information selection. After combining with the new cell state $C^{<t>}$ (filtered with tanh layer), the new hidden state $a^{<t>}$ is computed as:

$$o_t = \sigma_3(W_o \cdot [a^{<t-1>}, x^{<t>}] + b_o) \qquad a^{<t>} = o_t \times tanh(C^{<t>}) \tag{4}$$
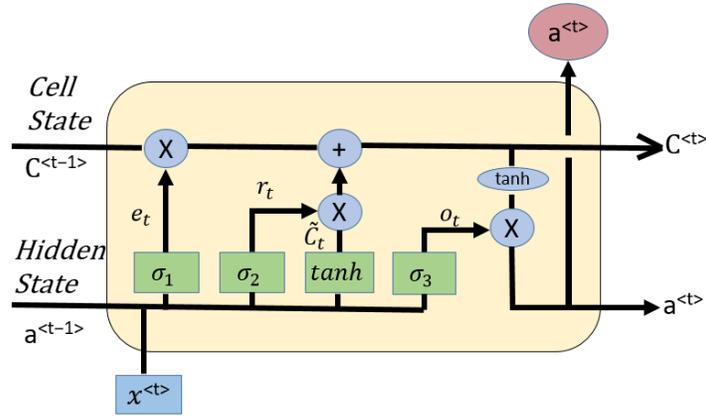
Figure 6. An overview of the LSTM block. The input, output and forget gates regulate whether information can be augmented or removed from the cell state.
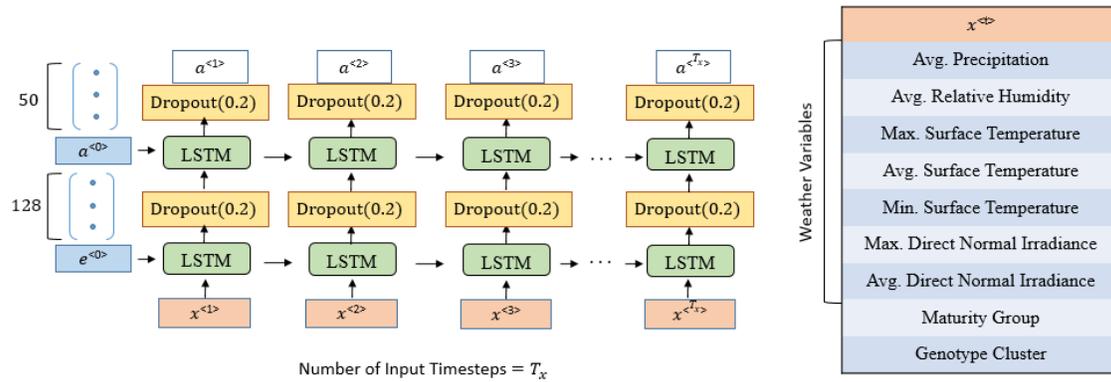

Figure 7. LSTM is used for encoding the input sequence which is of length T_x and the output from the first LSTM layer is a batch of sequences that are propagated through another layer of LSTM. We used dropout regularization after each LSTM layer to prevent overfitting.

After updating, the values of $a^{<t>}$ and $C^{<t>}$ are passed to the LSTM block of the next time-step. The forget gate and output activation function is the most critical components of the LSTM block and removing any of them can significantly impair performance [53].

Encoding Using Stacked LSTM

To encode the information of the input time-steps, we had two LSTM layers stacked on top of each other to get the T_x annotations as shown in Fig. 7. We finalized this model after an extensive hyper-parameter and architecture search. The Stacked LSTM can capture the long-range dependencies and temporal correlations for nonlinear data, therefore, they were ideal for our research problem. An LSTM layer consists of a sequence of directed nodes where each node corresponds to a single time-step.

The first layer of LSTM takes in input the information from all timesteps sequentially. For input at each time-step, we concatenated the information from different variables and this concatenated information acts as the input to the LSTM node. The LSTM node computes the hidden state as a function of the previous hidden state and the input vector for the current timestep. Each LSTM node updates the hidden state and cell state. The next LSTM node receives as input these updated states and the concatenated information of that time-step. After performing computations at each timestep of the time series, the first LSTM layer generates a sequence of hidden states for input to the next LSTM layer. The encodings returned by the first LSTM layer act as inputs for the next LSTM layer.
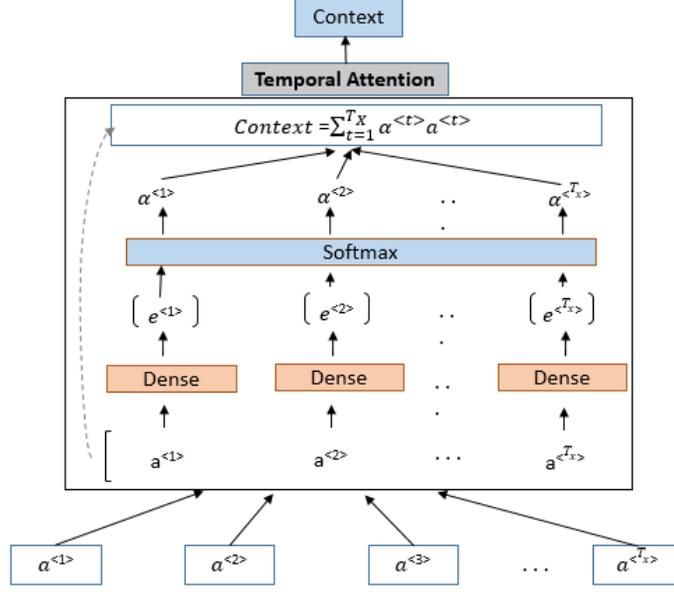
Figure 8. The temporal attention method used to compute the context vector and learn the attention weights simultaneously.

Temporal Attention

The temporal attention method takes in input a sequence of vectors and the aim is to compute aggregated information from these vectors. The vectors are annotations corresponding to the input time-steps. We computed the context vector from the weighted sum of annotations (hidden states) as shown in Fig. 8. Annotation $a^{<t>}$ focuses on the information surrounding the time-step t in the sequence. The attention weight $\alpha^{<t>}$ signifies the contribution of the information at a time-step t for prediction.

The context vector is computed like this:

$$context = \sum_{t=1}^{T_x} \alpha^{<t>} a^{<t>} \tag{5}$$

The attention weight for each annotation $a^{<t>}$ was computed using the softmax function.

$$\alpha^{<t>} = \frac{exp(e^{<t>})}{\sum_{t=1}^{T_x} exp(e^{<t>})} \tag{6}$$

The alignment model (dense layer, d) scores how well the input around time-step t is aligned with the prediction. It is parameterized as a feedforward neural network model as shown in Fig. 8. It is jointly trained with the entire network.

$$e^{<t>} = d(a^{<t>}) \tag{7}$$

Experiments

Downsampling the dataset

The original multivariate time-series data set comprises of 214 days (thus having 214 time-steps). Each day is represented by 7 weather variables (Table 1). We initially perform experiments using our Stacked LSTM model keeping the number of time-steps (T_x) as 214. We vary the number of input time-steps to compute the effect T_x has on the accuracy of the prediction. We consider the first 210 time-steps while downsampling the daily data to weekly ($T_x = 30$), biweekly ($T_x = 15$) and monthly ($T_x = 7$) values. We kept the sense of the variables the same while downsampling and thus instead of considering mean for all the variables, we compute an average of average (ADNI, ARH, AvgSur), maximum of maximum (MDNI, MaxSur), minimum of minimum (MinSur). For average precipitation, we consider both the total precipitation and the average precipitation for the considered time interval (7 days, 14 days, 30 days). The model performs better when the downsampling is performed using average precipitation and therefore we implement this in our experiments. [1]

---

[1] The abbreviations mentioned in Table 1 are used in Table 5, Table 6 and Table 7.

Table 1. Abbreviations used for different weather variables. At each time-step of the LSTM architecture, we used multivariate input.

Weather Variable	Explanation	Unit
ADNI	Average Direct Normal Irradiance	Wm^{-2}
AP	Average Precipitation Previous Hour	*inches*
ARH	Average Relative Humidity	*Percentage*
MDNI	Maximum Direct Normal Irradiance	Wm^{-2}
MaxSur	Maximum Surface Temperature	$°C$
MinSur	Minimum Surface Temperature	$°C$
AvgSur	Average Surface Temperature	$°C$

Number of Input time-steps

We kept the model architecture the same while performing experiments with different input sequence lengths. The RMSE values corresponding to different T_x were almost same as observed in Table 2 except for $T_x = 7$. For almost all of the experiments, we chose an intermediate value of $T_x = 30$ to facilitate faster training of LSTMs, to capture long-range temporal dependencies and also, not downsampling to a higher extent.

Table 2. Test Set RMSE values for different input sequence lengths.

Input Sequence Length (T_x)	Test RMSE
7	7.216
15	7.205
30	7.207
214	7.206

Adding Maturity Group, Genotype Cluster Informations

We performed three different experiments to optimize the augmentation of the information of Maturity Group (MG). To do this, we developed three different architectures. In the first approach, we concatenated MG to every time-step of the input. In the second approach, we concatenated MG to every time-step and also after the 2nd LSTM layer just before prediction. For the third architecture, we added MG only after the 2nd LSTM. The results are given in Table 3. The second approach which gave the least RMSE became our selected approach for the paper. Interestingly, we observe that even though MG is a static value for all time-steps, approach 1 gave better results than approach 3.

Table 3. Different approaches to augment MG information.

Approach	Test RMSE
MG added to every TS	7.246
MG added to every TS & after 2nd LSTM	7.223
MG added only after 2nd LSTM	8.138

With MG information already augmented, we performed a similar search to select a suitable architecture for adding the genotype cluster information. We observed that approach 2 also gives the least RMSE in this case (Table 4).

Table 4. Different approaches to augment genotype cluster information.

Approach	Test RMSE
Cluster added to every TS	7.230
Cluster added to every TS & after 2nd LSTM	7.201
Cluster added only after 2nd LSTM	7.239

Greedy Search - Weather Variables

With MG and cluster information augmented, we perform a greedy search to get insights about which weather variable is most important for prediction. In the first step of the greedy search, we trained our Stacked LSTM model for each of the 7 variables

listed in Table 1 separately. We chose the variable which gave the least RMSE value. With MinSur as the chosen first variable, in the second step, we trained 6 models with each of the other 6 variables added. Thereafter, we selected the variable which gave the lowest RMSE and thus, ADNI becomes the second most important variable. The third step involved training 5 models and we repeated this greedy search process till we got our 7th variable. The variables in decreasing order of importance (as obtained in the greedy search) are listed in Table 5. We observed that the RMSE values remain in the range (7.19 - 7.20) and adding extra weather variables after the 1st variable didn't show a considerable impact on the RMSE values.

Table 5. Greedy Search for the weather variables (Entire dataset).

Variable Added	Test RMSE
MinSur	7.204
ADNI	7.202
AvgSur	7.207
MDNI	7.190
AP	7.189
ARH	7.211
MaxSur	7.203

We performed a similar greedy search to get insights about the significance of variables separately for the northern and the southern parts. To do this, we divided our dataset based on the locations into two parts - North (MG 0 to 4) and South (MG 4 to 8). Our greedy search results for north data are shown in Table 6, while for the south data are shown in Table 7 in decreasing order of importance for the variables.

Table 6. Greedy Search for the weather variables (northern locations).

Variable Added	Test RMSE
ADNI	7.206
MDNI	7.202
MinSur	7.206
ARH	7.208
MaxSur	7.224
AvgSur	7.210
AP	7.228

Table 7. Greedy Search for the weather variables (southern locations).

Variable Added	Test RMSE
AvgSur	7.233
MDNI	7.236
MaxSur	7.222
ARH	7.234
ADNI	7.223
MinSur	7.249
AP	7.265

Performance Comparison of two models with RF, LASSO

We compared the performance of the two models (Stacked LSTM, Temporal Attention) with Random Forest (RF) and LASSO regression on same input features. RF and LASSO models were developed using scikit-learn Python library. We vary the input sequence length T_x to demonstrate the comparative results for these models. Each model has two variants based on the input information. For one variant, the input included 7 weather variables and for the other variant, the input included 7 weather variables, MG and cluster. The two models demonstrated comparable accuracy as shown in Table 8. The best RMSE value (7.130) obtained for $T_x = 30$ is about 14% of the average yield for the test set (50.745) and 44.5% of the standard deviation (16.019). Both the Stacked LSTM and Temporal Attention models outperformed RF and LASSO.

Table 8. Comparison of the performance of the two models (2 variants of each model based on the input information) with Random Forest and LASSO by varying the input sequence length (T_x) using test RMSE values.

T_x	Model	RMSE(Weather Variables)	RMSE(Including all)
7	Temporal Attention	8.272	7.137
	Stacked LSTM	8.246	7.132
	Random Forest	11.146	11.052
	LASSO	14.466	14.444
15	Temporal Attention	8.254	7.129
	Stacked LSTM	8.251	7.131
	Random Forest	11.018	10.666
	LASSO	13.732	13.711
30	Temporal Attention	8.236	7.148
	Stacked LSTM	8.289	7.130
	Random Forest	10.131	9.889
	LASSO	12.813	12.779

Results with different inputs to the Stacked LSTM model

For the Stacked LSTM model, we vary the inputs to the model in order to notice the change in performance. As both of our proposed models showed comparable accuracy, we chose only one model to do these experiments. The results are shown in Table 9 in terms of RMSE and R^2 (coefficient of determination) regression score function for the test set. All the information from MG, cluster and weather variables were required to get the best performance.

Table 9. Comparative performance of the Stacked LSTM for different inputs.

Model Input	Test RMSE	Test Set R^2
Only MG	15.201	0.099
Only Cluster	15.612	0.050
Only Weather Variables	8.274	0.733
MG, Cluster	15.149	0.106
MG, Weather Variables	7.211	0.797
Cluster, Weather Variables	8.159	0.741
MG, Cluster & Weather Variables	7.130	0.802

Comparison with USDA Model

We compare the performance of our deep learning model (Stacked LSTM) with the USDA's weather-based soybean yield prediction model[16]. The USDA model which uses a linear regression approach doesn't predict performance for individual locations. It predicts yield state-wise. Due to this limitation of the USDA model, we compare the models using year wise average across states for the test set (Table 10). We computed the absolute error (between predicted and actual yield) for both the models. The deep learning model showed much-improved performance compared to the domain knowledge-based USDA model.

Data Availability - Insights

We try to gain insights based on data availability for performance on the test set as shown in Fig 9. We plotted the test RMSE values using a heat map for all the (MG, Genotype Cluster) combinations. We also plotted the ratio (number of samples in training set/number of unique locations) to get an estimation of the data availability and data distribution for all the (MG, Genotype Cluster) combinations. From the figure, we observed that for the highest RMSE (MG = 7, Cluster = 1) the corresponding data availability ratio is low. This holds true for most of the highest RMSE combinations, although not for all. Therefore, while not conclusive, it seems bad performance can be attributed to less data availability/location in the training set leading to high RMSE in the predicted yield. The framework developed in this paper allows similar investigations to motivate other insights, and future hypotheses driven research.

Table 10. Comparative performance of Deep Learning (Stacked LSTM) model with the USDA model.

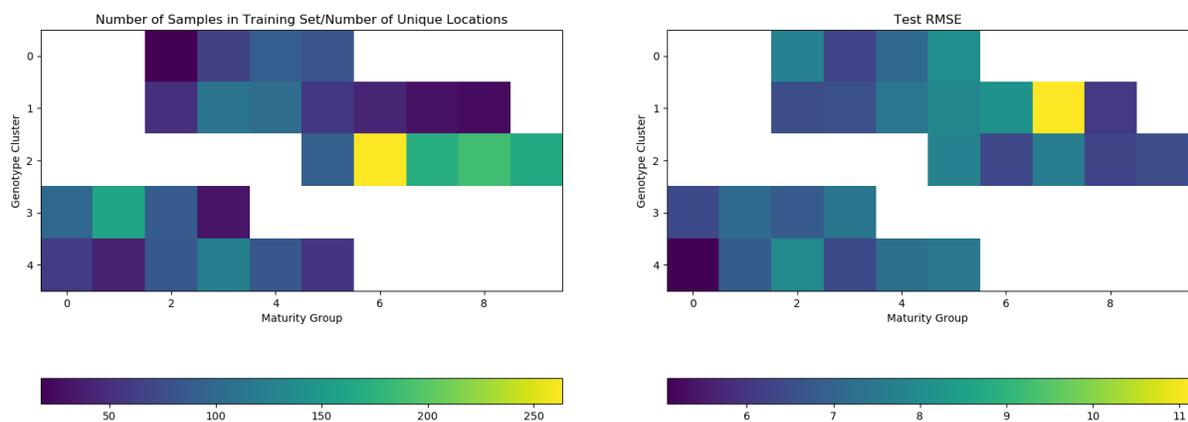| Year	Absolute Error (Deep Learning)	Absolute Error (USDA)
2003	0.63	5.49
2004	0.12	0.33
2005	0.19	2.75
2006	0.44	2.18
2007	0.21	0.37
2008	0.30	2.98
2009	0.04	0.84
2010	0.40	1.37
2011	0.34	0.06
2012	0.29	1.89
2013	0.04	0.27
2014	0.03	1.32
2015	0.35	1.70

Figure 9. Attempt to gain insights behind performance on the test set based on data availability in the training set.