# Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt

Environmental Research Letters

LETTER

# Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt

**Aleksandra Wolanin**[1,6], **Gonzalo Mateo-García**[2], **Gustau Camps-Valls**[2], **Luis Gómez-Chova**[2], **Michele Meroni**[3], **Gregory Duveiller**[3], **You Liangzhi**[4] and **Luis Guanter**[5]

[1] Remote Sensing and Geoinformatics Section, GFZ German Research Centre for Geosciences, Helmholtz-Centre, Potsdam, Germany
[2] Image Processing Laboratory, Universitat de València, València, Spain
[3] European Commission, Joint Research Centre (JRC), Ispra, Italy
[4] Environment and Production Technology Division, The International Food Policy Research Institute (IFPRI), Washington, D.C., United States of America
[5] Centro de Tecnologías Físicas, Universitat Politècnica de València, València, Spain
[6] Author to whom any correspondence should be addressed.

**E-mail:** ola@gfz-potsdam.de, gonzalo.mateo-garcia@uv.es, Gustau.Camps@uv.es, luis.gomez-chova@uv.es, Michele.meroni@ec.europa.eu, Gregory.duveiller@ec.europa.eu, l.you@cgiar.org and lguanter@fis.upv.es

**OPEN ACCESS**

**RECEIVED**
21 October 2019

**REVISED**
17 December 2019

**ACCEPTED FOR PUBLICATION**
7 January 2020

**PUBLISHED**
11 February 2020



**Keywords:** wheat yield, Indian Wheat Belt, food security, remote sensing, explainable artificial intelligence (XAI), deep learning (DL), regression activation map (RAM)

Supplementary material for this article is available online

## Abstract

Forecasting crop yields is becoming increasingly important under the current context in which food security needs to be ensured despite the challenges brought by climate change, an expanding world population accompanied by rising incomes, increasing soil erosion, and decreasing water resources. Temperature, radiation, water availability and other environmental conditions influence crop growth, development, and final grain yield in a complex nonlinear manner. Machine learning (ML) techniques, and deep learning (DL) methods in particular, can account for such nonlinear relations between yield and its covariates. However, they typically lack transparency and interpretability, since the way the predictions are derived is not directly evident. Yet, in the context of yield forecasting, understanding which are the underlying factors behind both a predicted loss or gain is of great relevance. Here, we explore how to benefit from the increased predictive performance of DL methods while maintaining the ability to interpret how the models achieve their results. To do so, we applied a deep neural network to multivariate time series of vegetation and meteorological data to estimate the wheat yield in the Indian Wheat Belt. Then, we visualized and analyzed the features and yield drivers learned by the model with the use of regression activation maps. The DL model outperformed other tested models (ridge regression and random forest) and facilitated the interpretation of variables and processes that lead to yield variability. The learned features were mostly related to the length of the growing season, and temperature and light conditions during this time. For example, our results showed that high yields in 2012 were associated with low temperatures accompanied by sunny conditions during the growing period. The proposed methodology can be used for other crops and regions in order to facilitate application of DL models in agriculture.

## 1. Introduction

The Food and Agriculture Organization (FAO) of the United Nations estimates that 50% more food needs to be produced by 2050 in order to feed the increasing world population (FAO 2017). However, the ongoing efforts to increase food production are curbed by climate change, which has already impacted global mean yields of major crops in the last decades, and is further projected to negatively affect food security (FAO, IFAD, UNICEF, WFP and WHO 2018, Mbow *et al* 2019). It is therefore crucial to not only accurately

predict crop yield, but also to model and characterize the processes involved by understanding the meteorological drivers of crop yield variability.

Meteorological variables influence crop growth, development, and final grain yield in a nonlinear manner and often with complex interactions (Siebert *et al* 2017, Akter and Rafiqul Islam 2017). These variables are accounted for in both process-based as well as statistical models to estimate crop yield (e.g. Lobell *et al* 2011, Iizumi *et al* 2018). While process-based models require detailed (and not always available) information on the farmers' practices, recent increase in the availability of global satellite observations and advancements in statistical methods have fueled the application of machine learning (ML) models at various scales (e.g. Lobell *et al* 2011, Guan *et al* 2017, Cai *et al* 2019). In particular, such models may have the capability of accounting for additional factors reducing growth and yield (e.g. pests, diseases, weeds and other perils).

In order to better exploit the wealth of information in the meteorological and satellite-derived vegetation data, we propose to apply convolutional neural networks (CNNs), a class of deep learning (DL) models (Goodfellow *et al* 2016, LeCun *et al* 2015). DL methods promise great advances in Earth observation and geosciences (Reichstein *et al* 2019), as they can account for the large size of the input datasets, nonlinear relations, and interconnections among multiple variables. Since CNNs can learn the features directly from the data, this approach does not depend on the manual selection of specific parameters.

A major shortcoming of ML methods in general, and of DL methods in particular, is that the learned relations are hidden under very complicated prediction functions. However, recent years have seen the emergence of a whole field of ML called 'Explainable Artificial Intelligence' to face this issue (Miller 2019), and techniques and methodologies have been introduced to study what the ML/DL models are learning (Montavon *et al* 2018). In this work, we focus on an efficient method to explain the model predictions called regression activation mapping (RAM) (Zhou *et al* 2016, Wang *et al* 2017, Wang and Yang 2017). RAM contains the immediate information for the final prediction, but also maintains the correspondence to the input data in the time (or spatial) dimension and shows the contribution of each time (or space) instant to the final output. As a result, we not only benefit from DL in terms of improvements in the model performance, but we specifically focus on evaluating the underlying drivers, thereby placing confidence in the model as well as gaining insight into the relevant conditions leading to crop yield variability.

Here, we focus on the wheat yield estimation in the Indian Wheat Belt, where concerns of yield stagnation have risen due to increasingly inconsistent growth trend of wheat yields in recent years (Ray *et al* 2012, Tripathi and Mishra 2017). High temperature has been identified as the most detrimental factor, affecting both crop growth and grain formation (Lobell *et al* 2012, Jain *et al* 2017), but also heavy and untimely rainfall and hailstorm events have caused large-scale damages to the crops (Singh *et al* 2017). Finally, cloudy weather with high humidity and low temperatures increases chances of wheat diseases (e.g. wheat rusts and spot blotch) (Duveiller *et al* 2007, Kaur *et al* 2015), which have been spreading in India (Hodson 2011). In addition, in the specific case of the double cropping system in India (typically rice-wheat), the timing of wheat sowing may be suboptimal because it is conditioned by local farming practices and by the timing of the rice harvest (Global Information and Early Warning System on Food and Agriculture (GIEWS) 2019). As the Indian Wheat Belt poses many challenges regarding complex processes and nonlinearities, it constitutes a good scenario to evaluate our approach that could be eventually used in other similar settings.

Therefore, the aims of the paper are two-fold: (1) to develop a DL model for wheat yield estimation and evaluate its application for within-season yield forecast, and (2) to scrutinize what this model learned by visualizing and interpreting the drivers of yield estimation. This knowledge extraction process may have implications in further crop management actions, as well as interactions with stakeholders and farmers.

## 2. Materials and methods

Our analysis was performed in the Indian Wheat Belt region, which supplies around 70% of India's total wheat production (figure 1). We estimated district-level crop yield using as input a set of time series of meteorological and satellite-derived vegetation variables at a daily resolution that are summarized in table 1 and described in S1. In our analysis, we used three vegetation indices (VIs): the normalized difference vegetation index (NDVI) (Tucker *et al* 1985), the normalized difference water index (NDWI) (Gao 1996), and near-infrared reflectance of vegetation (NIRv) (Badgley *et al* 2017); and seven environmental variables: minimum, mean and maximum air temperatures ($T_{min}$, $T_{mean}$, $T_{max}$), downward shortwave radiation flux ($SW_{down}$), water vapor pressure deficit (VPD), precipitation, and day-length (table 1). Pixel level values of input variables were spatially aggregated to the district level as the weighted average according to fractional area occupied by wheat in each pixel (S1 in the supplementary information, available online at stacks.iop.org/ERL/15/024019/mmedia).

The DL models (CNN) and the baseline models (ridge regression, RR, and random forest, RF) were trained and tested on the dataset of 143 districts over 13 years (2001–2013) (S2). The time range of the analysis was determined by the availability of both MODIS imagery and wheat yield data for all states. In the standard setup, all districts were pooled together

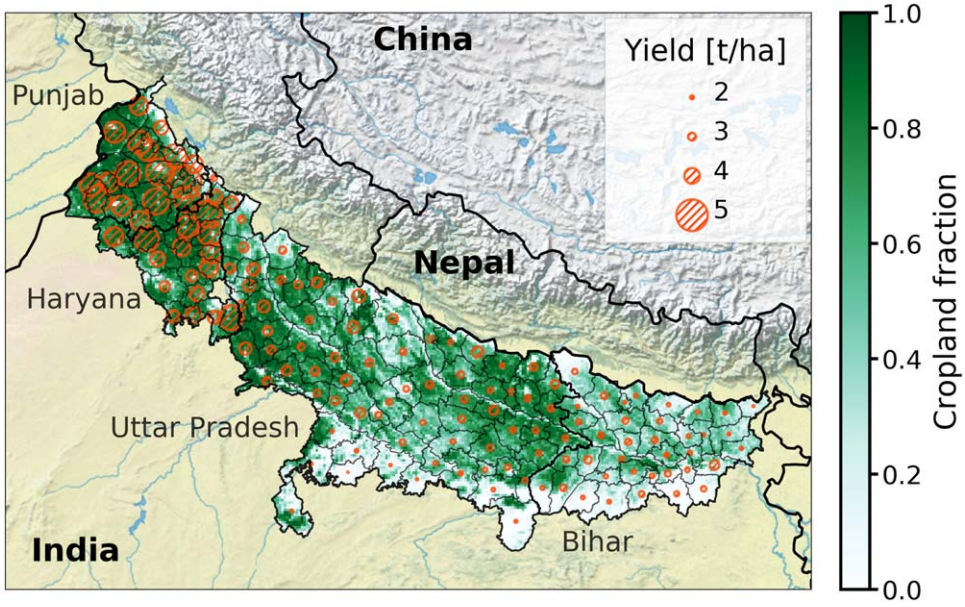


**Figure 1.** The study area of the Indian Wheat Belt region. The *rabi* season cropland fraction for season 2007–08 is shown for the four wheat producing states used in this study (Punjab, Haryana, Uttar Pradesh, Bihar) in green. The circles represent the wheat yield for the districts (delineated in gray) for the same season.

and a separate model for each year was calculated, where data from this year were used for testing and data from the rest of the years were used for training.

The CNN model stacked one-dimensional convolutional layers (Conv1D) along time dimension and max pooling layers (MaxPool1D) with a window size of five (figure 2(a)). After the second Conv1D, the data were fed into a global average pooling layer (GAP), which computed the mean in the time dimension for each variable. Finally, a simple linear layer was applied to obtain the final yield prediction.

We propose to use RAMs as a tool to visualize and interpret how the CNN models achieve their results. Activation mapping has been previously applied for image analysis in the classification (Zhou *et al* 2016) and regression problems (Wang and Yang 2017). Here, we show their application on time series data, as inspired by the application of class activation map to interpret the temporal data in Wang *et al* (2017). For a time series input, the RAM is another time series such that its average over time (plus bias) corresponds to the predicted output value. As a result, RAM contains the immediate information for the final prediction, but also maintains the correspondence between the last convolutional feature maps and the input data in the time dimension. RAM is calculated by combining the convoluted data that is fed into GAP layer with the weights of the final linear layer (figure 2(b)). In particular, if $\{z_{1,t}, \ldots, z_{d,t}\}$ are the input time series to the GAP layer, the output of the GAP layer is the mean over the time dimension $\frac{1}{T}\sum_t z_{i,t}$ for each time series $i \in \{1,..,d\}$. The final yield prediction ($\hat{y}$) is then the linear combination of those averaged time series:

$$\hat{y} = \sum_i^d \left( \frac{1}{T} \sum_t^T z_{i,t} \right) w_{z,i} + b_z \ , \tag{1}$$

where $w_{z,i}$ and $b_z$ are respectively the weights and the bias of the output linear combination. In this setting, we define the RAM ($r_t$) as the following time series:

$$r_t = \sum_i^d z_{i,t} w_{z,i} \ . \tag{2}$$

RAM satisfies that $\hat{y} = \frac{1}{T}\sum_t r_t + b_z$. Hence the RAM value at a given time step $r_t$ can be interpreted as an estimation of the derivative of the output w.r.t. time, $r_t \approx d\hat{y}/dt$, and the final estimation as the integral of RAM, $\hat{y} = \frac{1}{T}\int r(t)dt$.

The models were applied directly to the multivariate time series of input variables (VIs and meteorological data). The time period analyzed corresponded to 245 days, starting from October 1st, that cover the *rabi* growing season when wheat is grown. As a result, each input for the DL model was a tensor of size (245, $n_{vars}$), where $n_{vars}$ is a number of input variables that varied depending on the experiment. In case of the baseline models, the daily data was reshaped into vectors (of length 245 $\times$ $n_{vars}$). Afterwards the baseline models were trained using those vectors as input and the final yield as output. Application of RR and RF to monthly averages slightly increased model performance (S3).

The model performances were compared among CNNs, as well as the baseline models for three different combinations of the input data, with two, five or ten input variables using the time series data only ('No district'), as well as using additionally the district information as input ('Incl. district') as in a mixed-effects model (Wu 2009). Specifically, in the 'Incl. district' setting, we added a district-dependent bias in the last layer of the

**Table 1.** Main characteristics of the data used in this study. Most of the data was downloaded using the Google Earth Engine (GEE) cloud computing platform (Gorelick *et al* 2017).

| Category | Variables | Spatial resolution | Temporal resolution | Source |
|---|---|---|---|---|
| Satellite observations of vegetation | NDVI, NDWI, NIRv calculated from MODIS Bands 1, 2 and 7 | 500 m | 16 day resolution, 1 day sampling | MODIS MCD43A4 V6, exported from GEE |
| Meteorological data | $T_{min}$, $T_{mean}$, $T_{max}$, $SW_{down}$, VPD | 0.25° | 3 h, aggregated to daily values | GLDAS-2.1, exported from GEE |
| | Precipitation | 0.05° | Daily | CHIRPS, exported from GEE |
| | Day-length | District | Daily | Calculated between civil twilight before the sunrise and the end of civil twilight after the sunset |
| Crop fraction | *Rabi* season wheat crop fraction | 5000 m | Yearly | Annual Cropland Datasets of National Remote Sensing Centre in India |
| Wheat yield data | *Rabi* season wheat yield | District | Yearly | Indiastat |

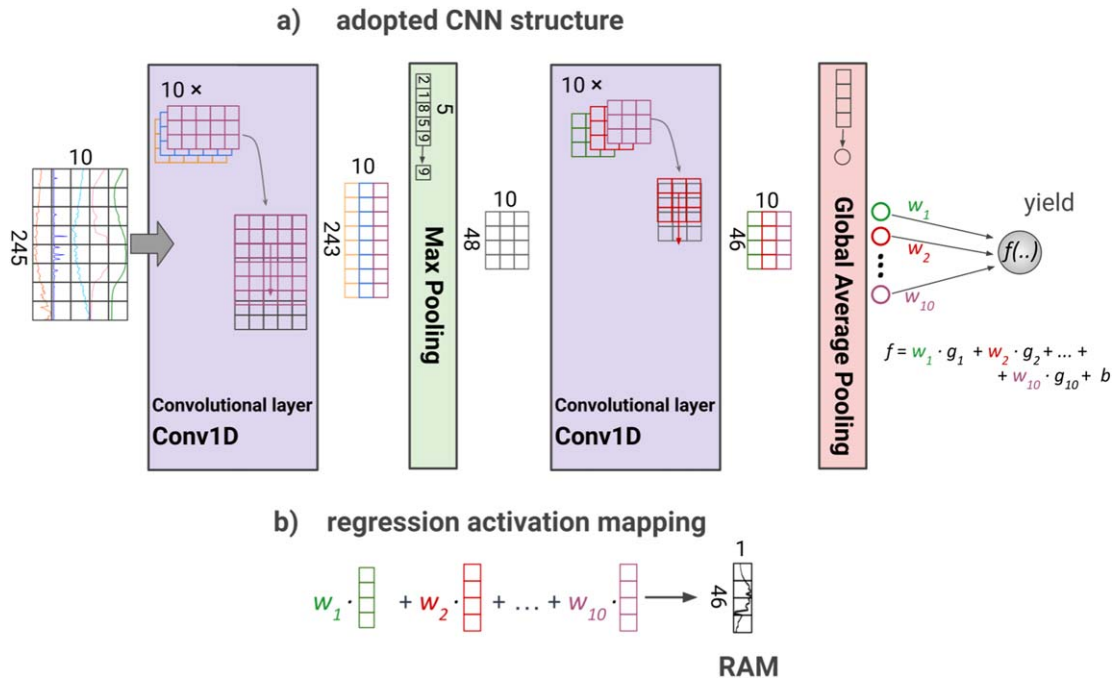


**Figure 2.** Convolutional neural network (CNN) architecture used in this study (a) and regression activation mapping (b). The numbers reflect the input data size, size of features during the model processing, and number of the filters used in the CNN model (with 10 input variables and district information, as in table S2), where the input for the DL model was a tensor of size (245, 10), where 245 corresponds to the length of the analyzed time period and 10 is the number of input variables.

CNN model that was fitted during the training process. The district information was represented in RR as an additional binary class matrix, and in RF—as a categorical variable. The best model architectures and hyperparameters were determined as described in S3. In addition, we compared the model performance to the null model, in which the yield for each district for every year was calculated as the average of yields from the other available years in the input dataset for this district (e.g. yield for the district Patiala in 2006 was calculated as the mean of yields in Patiala for years 2003–2005 and 2007–2013). As a next step, to evaluate what input parameters were the most important for the wheat yield estimation, we run ensembles of five CNN models in two simple sets of varying combinations of input variables: (1) with one variable only (for all ten variables); and (2) with two variables: VI + one meteorological parameter (21 combinations).

Finally, two models where selected for RAM calculations and visualizations: CNN with two input variables, NDWI and $T_{min}$ ($CNN_2$), as well as all ten input variables ($CNN_{10}$), both including district information. These models were run in model ensembles of ten members each. In addition to creating a separate model for each year (as previously), we also re-trained these models on all the input data ($CNN_{2,all}$ and $CNN_{10,all}$) in order to compare RAMs across different years created with the same model weights.

Because our study compares many different models, we chose to use a single evaluation metric to simplify the process of model selection and the presentation of the results. The selected metric is the Nash–Sutcliffe efficiency (NSE) (Nash and Sutcliffe 1970), that provides an indication of the goodness of fit and can range from $-\infty$ to 1, with the best possible score (1.0) meaning that modeled crop yields are equal to reported ones.

## 3. Results

### 3.1. Performance of the models

The performances measured as NSE for different models and input datasets are compared for all years averaged and for one year separately (2012) in table 2. The results for 2012 were shown to demonstrate that the overall good performance among all years does not necessarily guarantee accurate predictions for abnormal years. 2012 was a peculiar year characterized by very high yields compared to average (figure S8), as indicated by poor performance of the null model (table 2). CNN models provided the best results overall (best NSE of 0.868 among all years in the 'Incl. district' setting, compared to the best NSE of 0.757 for RR and 0.836 for RF) that were stable among model runs (S4), which shows applicability of such models to dense time series for yield estimation. The null model was already a good predictor for the yield among all years (NSE of 0.812) and it performed better than any model that excluded district information, as the majority of the yield variation came from the spatial variation (S5). However, when considering the abnormal year 2012, the null model performed much worse than the other models (NSE of 0.288), which demonstrates the

**Table 2.** Yield estimation performance for the test sub-sets using CNNs, ridge regression and random forest at their best performing architectures for different input variables combinations: 2 input vars. (NDVI + $T_{min}$), 5 input vars. (NDVI, NDWI, $T_{min}$, $SW_{down}$, VPD), 10 input vars. (NDVI, NDWI, NIRv, $T_{min}$, $T_{mean}$, $T_{max}$, $SW_{down}$, VPD, precipitation, day-length). Excluding or including district information (+ district) refers to 'No district' or 'Incl. district' settings, respectively.

| Model | input vars. | NSE | |
|---|---|---|---|
| | | All years | 2012 |
| CNN | 2 | 0.663 | 0.251 |
| CNN | 5 | 0.740 | 0.494 |
| CNN | 10 | 0.788 | 0.625 |
| CNN | 2 + district | 0.830 | 0.532 |
| CNN | 5 + district | 0.862 | 0.741 |
| CNN | 10 + district | **0.868** | **0.743** |
| RR | 2 | 0.734 | 0.418 |
| RR | 5 | 0.744 | 0.436 |
| RR | 10 | 0.756 | 0.432 |
| RR | 2 + district | 0.737 | 0.421 |
| RR | 5 + district | 0.746 | **0.438** |
| RR | 10 + district | **0.757** | 0.434 |
| RF | 2 | 0.704 | 0.338 |
| RF | 5 | 0.744 | 0.443 |
| RF | 10 | 0.754 | 0.481 |
| RF | 2 + district | 0.827 | **0.493** |
| RF | 5 + district | 0.831 | 0.480 |
| RF | 10 + district | **0.836** | 0.453 |
| Null model[a] | | 0.812 | 0.288 |

[a] Null model is calculated for each district for every year as the average of yields from the remaining available years in this district.

importance of using satellite observations to estimate yields for atypical years.

### 3.2. Impact of input variables on the model performance

Performance of CNNs improved with the increase in the number of input parameters, both in terms of the best performing models, as well as among all tested model settings (table 2). When testing models with one or two variables to evaluate the parameters importance, the model performance and variable ranking were not consistent among the years, with 2012 showing the worst performance (figures 3 and 4). The models performed much better and had a smaller performance variation if the information on the district was included. However, in 2012, some model combinations gave better results without district information, which is related to the fact that the null model performed poorly for this year.

In case of models with one variable only and no district information ('No district' in figure 3), VIs performed better among all the parameters (and NDWI the best). Models with day-length as the only variable had a very good performance, even though this parameter does not carry any information about the crop condition. It suggests that the model was actually trying to learn the specific day-length patterns related to districts, as yields have a clear spatial pattern in this region (figure 1) that is somewhat similar to the latitudinal distribution of day-length (figure S9).

When the district information was fed into the model ('Incl. district' in figure 3), the best performing model across all years was the one using NDWI, followed closely by $SW_{down}$ and other VIs. The good performance of NDWI and $SW_{down}$ was reproduced for models with two variables (figure 4). However, this behavior was not consistent among all the years for neither one- nor two-variable models. Good performance of NDWI and $SW_{down}$ in 2012 (figures 3 and 4) could be a leading factor responsible for their high overall ranking, as the differences among the models in this year were much larger than for other years. For example, in case of 2002, the ranking of pairs including NDWI and $SW_{down}$ were among the worst, but the general variability was very small. These results suggest that accounting for various parameters is important, as the conditions that limit or boost the crop yields vary among the years.

### 3.3. RAMs and their sensitivity to input variables

Overall, the main features of RAMs were the same for all the ensemble members (figure 5), as well as for the models re-trained on all the input data ($CNN_{2,all}$ and $CNN_{10,all}$). We compare in detail RAMs of the district Patiala for 2006, for selected CNN models using two ($CNN_{2,all}$) or ten ($CNN_{10,all}$) input variables in figures 6(a)–(b). As mentioned before, the average of RAM plus bias (bias both of the whole model and the district-specific, see S6), equals the estimated yield. Even though RAMs are products of the complex nonlinear interactions of the input data, some basic patterns can be directly inferred. For example, the overall shape of RAMs is closely related to the growing cycle as shown by NDWI. In case of the 2-variable model (figure 6(a)), all observed temporal patterns of RAM can be related to NDWI or $T_{min}$, as these were the only input parameters. As a result, a small dip in RAM around day 100 can be associated with the increase in $T_{min}$, as NDWI during this time was steadily increasing, which suggests the negative impact of higher $T_{min}$ on the crop yield. To support this claim, we modified $T_{min}$ and calculated the resulting changes in RAMs as shown in figure 6(c). Two small changes were performed on the $T_{min}$ data—in the first case we removed the $T_{min}$ peak around day 100 (shown in blue in figure 6(c)) and in the second case, we increased $T_{min}$ around day 150 (shown in yellow in figure 6(c)). The changes in the resulting RAMs—which directly relate to changes in the estimated yield—are highlighted in blue and yellow in figure 6(c). The decrease in $T_{min}$ led to the increase in RAM and removed the previously observed dip. On the other hand, increase in $T_{min}$ led to a similar decrease in RAM,

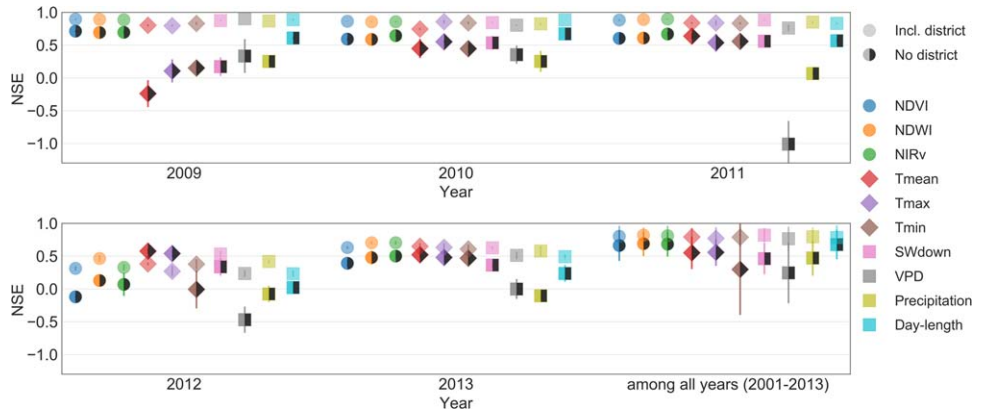


**Figure 3.** Model performance (NSE) for ensembles of five CNN models with only one variable as the input for selected years and averaged among all years. The results for two model settings are tested: 'No district' is using only one bias in the whole model in the final layers, while in case of 'Incl. district' a bias per district (time-independent) is additionally fitted during the training process.

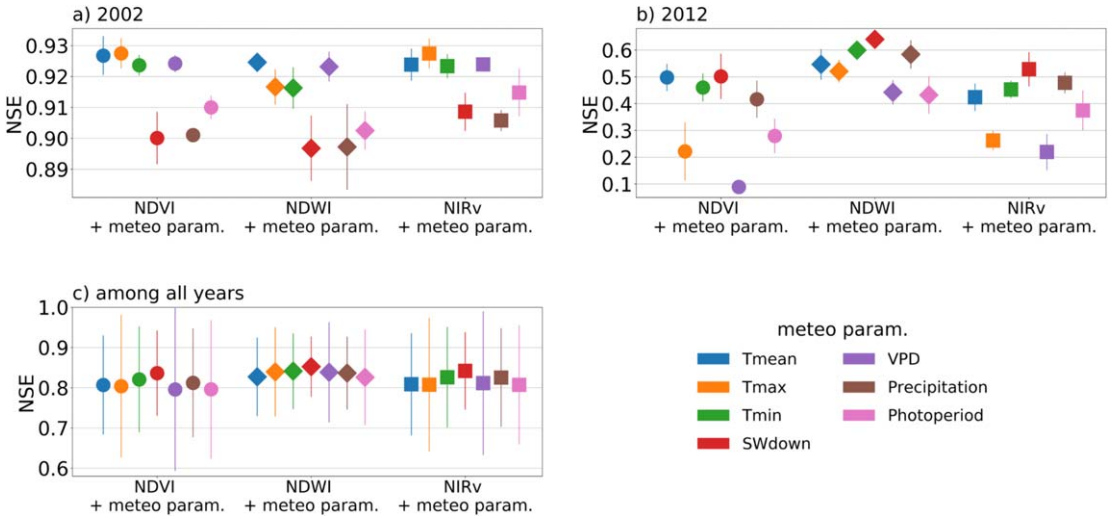


**Figure 4.** Model performance (NSE) for ensembles of five CNN models with only two input variables (VI + meteorological variable) for selected years and averaged among all years. Only results for models including district information ('Incl. district') are shown.

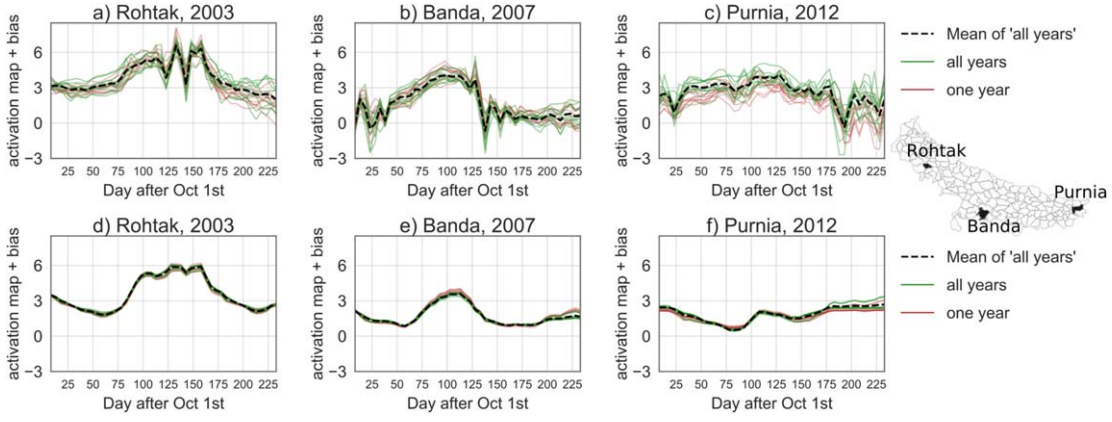


**Figure 5.** Variability of regression activation maps calculated for different districts and years for different members in a model ensemble of ten with ten input variables (a)–(c) or in a model ensemble of ten with two input variables (d)–(f). The thin green lines ('one year') correspond to models calculated for one year using other years as training, red lines ('all years') correspond to models trained on the whole dataset, while the thick dashed lines (Mean of 'all years') correspond to the average of the 'all years' models.

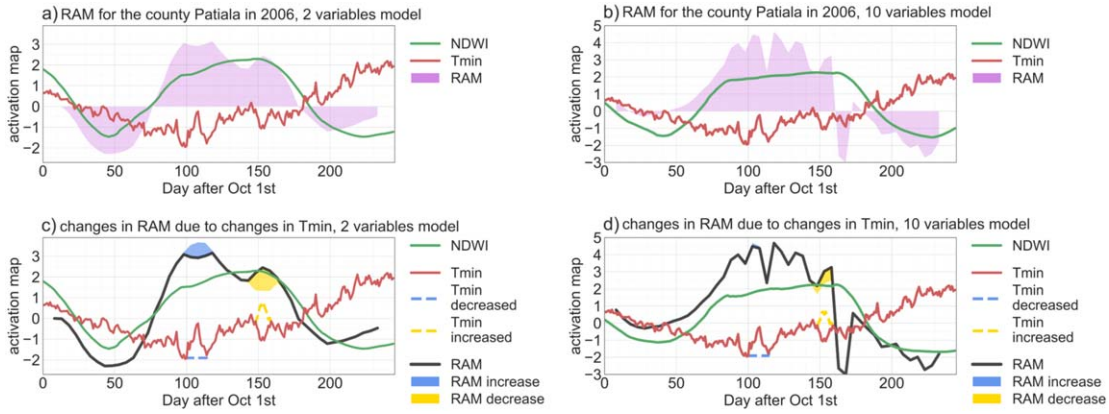


**Figure 6.** Regression activation maps for Patiala (2006) for models with input of two ($CNN_{2,all}$) and ten ($CNN_{10,all}$) variables. Original RAMs ((a) and (b)), and changes in RAMs due to modification in $T_{min}$ ((c) and (d)) are shown. Colorful lines reflect the changes in $T_{min}$, while the highlighted areas correspond to changes in RAMs (that are directly transferred into changes in the estimated yield). $T_{min}$ and NDWI are scaled (as the input to CNNs).

and therefore a decrease in the yield. However, analogous $T_{min}$ modifications for $CNN_{10,all}$ resulted in different magnitudes of changes in RAMs due to decrease or increase in $T_{min}$. The increase in $T_{min}$ towards the end of the growing season had a much bigger impact on RAM than the decrease in $T_{min}$ earlier in the growing season, which led to a positive but very small change in RAM.

To better identify the parameters and features that led to certain responses in RAMs, we compared selected input variables and RAMs for Patiala for two very different years: the year 2006, when the yield was quite poor (4.233 t/ha), and 2012, when the yield was very good (5.473 t/ha) in figures 7(a) and (b). In 2006, RAM demonstrated overall lower values and many strong drops, while in 2012, RAM showed consistently high values despite not much higher VIs. To analyze the drivers of this variability, we exchanged, one in turn, four input variables ($T_{min}$, $T_{max}$, $SW_{down}$, precipitation) among 2006 and 2012 to check how this would affect changes in both RAMs and estimated yields (figures 7(c)–(j)). For example, $SW_{down}$ was mostly responsible for creating the dips in RAMs in 2006 for Patiala—as many of them were filled when $SW_{down}$ from 2012 was used, especially the big drop around day 160 (figure 7(g)). Overall, the yield in 2006 increased using any of the input parameters from 2012, but the highest increase was observed for $T_{max}$, followed by $SW_{down}$, $T_{min}$ and only slightly for precipitation. The same rank of variables in terms of the magnitude of the impact was observed in the case of modifying the input variables in 2012.

Similar features were consistent among all districts in 2006 and 2012, as shown by anomalies (positive in red, negative in blue) in RAM, NDWI, $T_{max}$, $T_{min}$ and $SW_{down}$ in figure 8. In general, these years were relatively sunny, but in 2012, both $T_{max}$ and $T_{min}$ had strong negative anomalies during the second half of the growing season. The rapid and short negative anomalies in RAMs were usually related to negative anomalies in $SW_{down}$ (highlighted with green boxes), which sometimes were also accompanied by positive anomalies in $T_{min}$ (as cloudy skies capture the long-wave radiation emitted by Earth at night). It is noted that this link may lead to a wrong interpretation of the effect of $T_{min}$, when considered alone (e.g. a model with $T_{min}$ as the only weather input). Positive anomalies in RAM in 2012 were connected to strong negative anomalies in $T_{max}$ and $T_{min}$ (highlighted with magenta boxes).

### 3.4. Predictive skill of CNNs

We examined if our CNN model could be potentially useful for predicting the yield during the growing season by analyzing the impact of shortening the input time series in 25-day steps on the model performance. As compared to the null model, $CNN_{10}$ had better prediction from day 145, which corresponds to the last week of February (figure 9), In general, the month of February is the time of the anthesis stage, conditions during which are crucial for the final crop yield. Therefore, covering this time period is essential for a good yield estimation. From day 195 (middle of April), which widely corresponds to the harvest time in this region, the model performance did not increase. This suggests that after harvest the additional data were not useful anymore for the yield estimation. Although this may seem trivial, it is nevertheless important that model performance appears insensitive to post-harvest data, as adding such data could occur whilst in an operational context.

## 4. Discussion

### 4.1. Application of DL models for yield estimation

Even though our DL modeling approach was constrained by the requirement for network explainability (global pooling layer after the convolutional layers to facilitate RAMs), it outperformed RF and RR in the

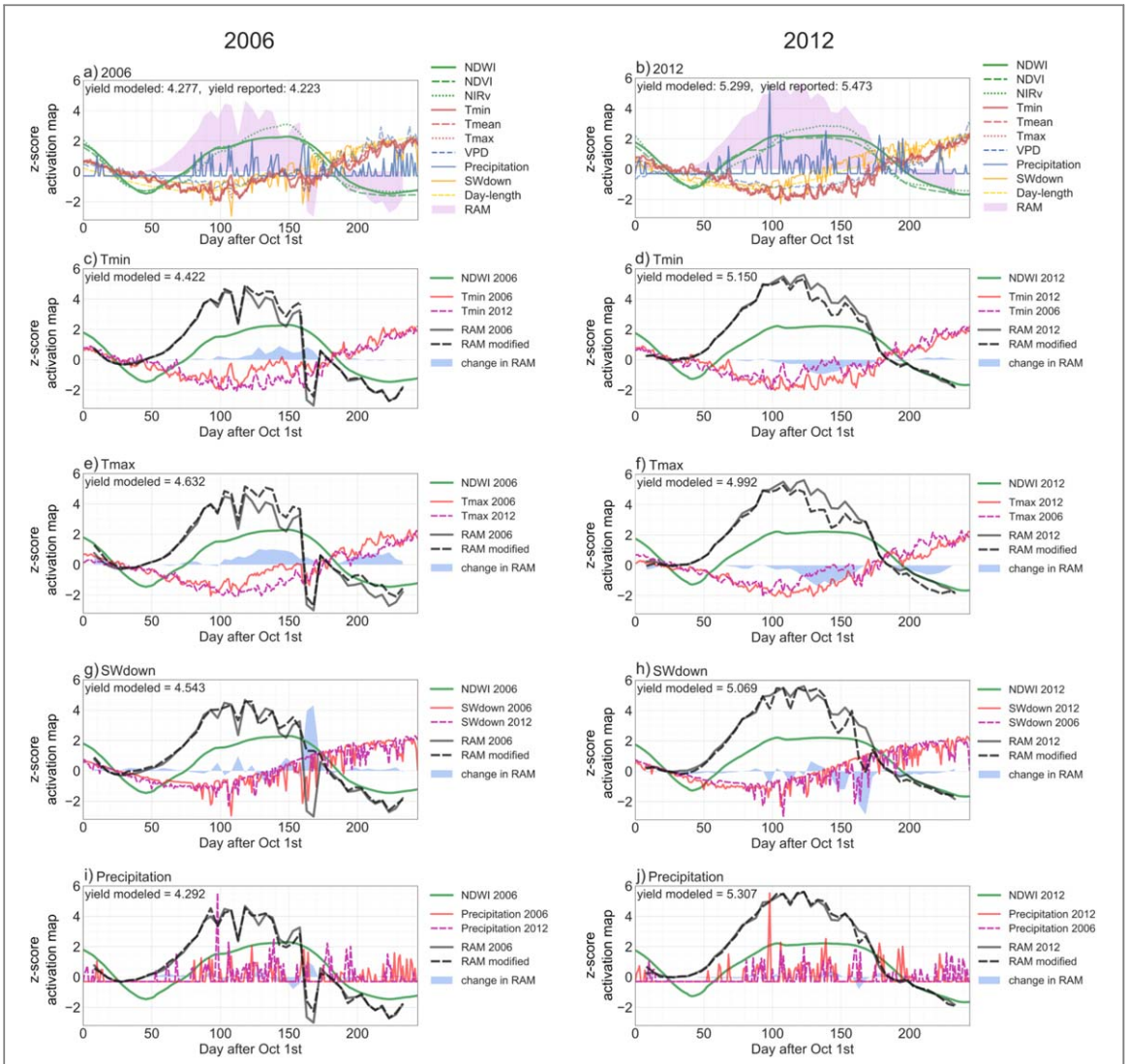


**Figure 7.** Regression activation maps for Patiala for the model with input of ten variables ($CNN_{10,all}$). Original RAMs and scaled input data for year 2006 (a) and year 2012 (b). Modified input data and RAMs for year 2006 (c), (e), (g), (i) and 2012 (d), (f), (h), (j). Modified input parameters (imported from the other year) are shown in dashed red lines (c)–(j). Input parameters shown are scaled (z-score, therefore unitless), while RAMs represent the intermediate step of the data processed in the model (in general unitless, but practically related to the yield in the yield units).

yield estimation task, though RF provided a strong performance improvement as compared to RR. Although DL models are already widely used for a variety of problems in image and signal processing domain (LeCun *et al* 2015), DL applications for crop yield estimation are still rare (e.g. You *et al* 2017, Crane-Droesch 2018), as the transfer of available tools from the field of ML into the environmental applications requires adaptations that account for specificities of the data. Our results demonstrate that CNNs can be a valid tool for capturing complex processes in agricultural systems, and their application can be therefore further explored for other crops and regions.

### 4.2. Impact of environmental conditions on the crop yield as captured by DL

To generalize well in the yield estimation problem, it is important to train models over several years and simultaneously account for multiple vegetation and environmental variables, as the main yield drivers vary among seasons. Since many of the input variables in the crop yield models are correlated and inter-related, multiple parameters should be considered when trying to understand the impact of variables on the crop yield in the model. For example, the impact of $T_{min}$ varied depending on whether it was the only input meteorological variable or one of many (figure 6).

RAM's general shape reflected VIs, which emphasizes the importance of the length of the growing period and agrees with previous studies (Lobell *et al* 2012, Jain *et al* 2017). Meteorological variables used together with VIs are expected to have only a marginal contribution to yield estimation, as their impact is already reflected by vegetation growth that is captured by VIs. Overall, $SW_{down}$ turned out to be the most important meteorological variable for yield estimation, even though the ranking varied significantly among the years. Comparison of performances among models

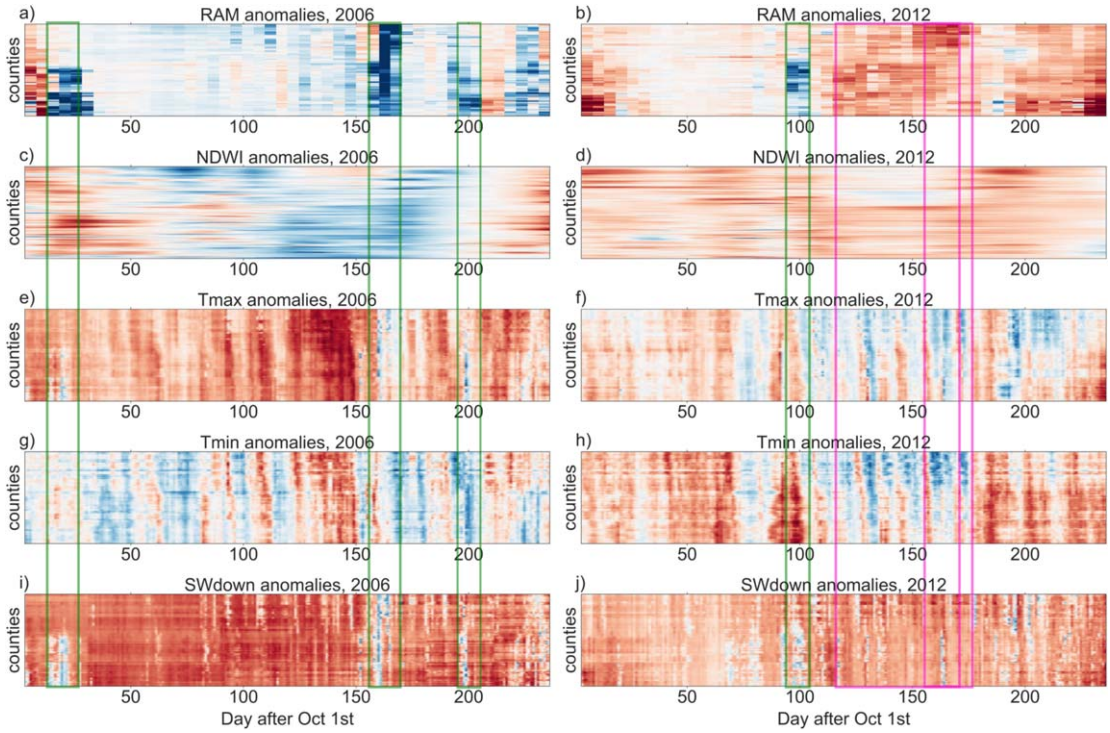


**Figure 8.** Anomalies of RAMs and selected input variables for all districts for the years 2006 and 2012. Time series for each district are shown as rows, and columns correspond to time steps. Starting from the top, districts are located in states of Punjab, Haryana, Uttar Pradesh and Bihar. The anomalies were calculated by subtracting district-specific multi-year averages from the scaled data for the given year. Positive anomalies are shown in red, negative in blue. The strong and rapid negative anomalies in RAMs that could be related to negative anomalies in $SW_{down}$ are highlighted in green rectangles, while positive anomalies in 2012 that could be related to strong negative anomalies in $T_{max}$ and $T_{min}$ are highlighted in magenta rectangles.

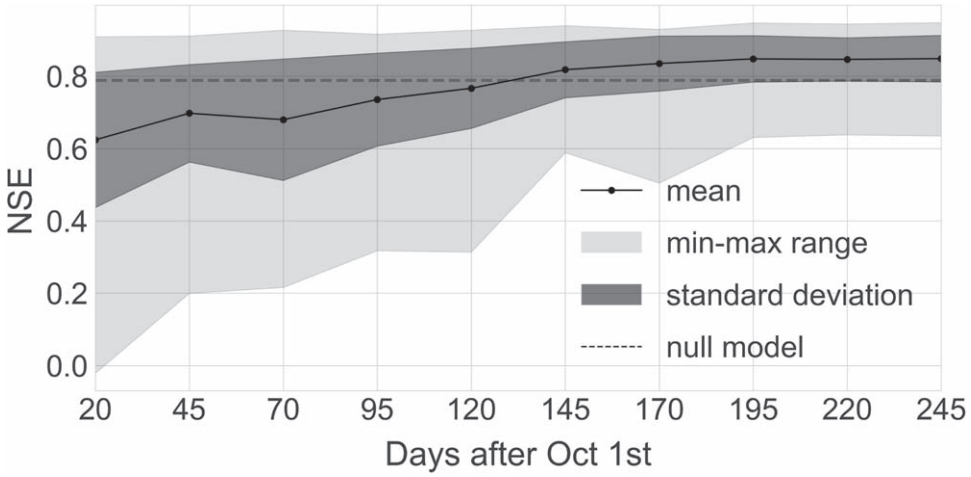


**Figure 9.** Model performance (NSE) as a function of changes in the end limit of the input data among all years. One $CNN_{10}$ model was retrained for each year separately for the input data shortened in 25-day steps. The shaded range of values corresponds to NSE variability among different years.

with one or two input variables, as well as the RAM analysis, showed that high yields in 2012 were associated with low temperatures accompanied by sunny conditions during the growing period, though most recent studies focus primarily on the impact of increasing temperatures in the Indian Wheat Belt (e.g. Duncan *et al* 2014, Jain *et al* 2017, Song *et al* 2018) and in other regions (Lobell *et al* 2005). As an important driver of photosynthesis, $SW_{down}$ affects the resources that crops can build during the growing season, which is apparently not fully reflected in the VIs. High radiation levels are also especially important during the critical period for grain number determination (20–30 days before anthesis to ten days afterwards) (Fischer 1985). Analysis of RAMs showed that the model was trained to recognize events of decreased light as having negative impact on the crop yield, which suggests that the decline of photosynthesis due

to decreased radiation was larger than the benefit due to increased diffused light. This agrees with the negative effects associated with a reduction in sunlight shown on a global scale (Proctor *et al* 2018). Decreased penetration of sunlight for the crop has already spurred demand for breeding for high photosynthetic or radiation-use efficiency (Joshi *et al* 2007). These observations also suggest that sun-induced fluorescence could be a good direct proxy for the crop yield, as it carries information on light conditions, and was previously shown to perform better than VIs for yield estimation (Song *et al* 2018). Although NIRv was found superior to other VIs for estimating GPP (Badgley *et al* 2017), we did not detect any benefit of using it for the yield estimation.

### 4.3. Physical interpretability of RAMs

The novelty of our approach extends beyond obtaining a satisfactory model performance, as we adapted the activation mapping to the regression problem for the temporal data in order to localize important patterns for the yield estimation. The analysis of RAMs applied in the time dimension facilitates DL network interpretability and provides the needed transparency on what DL models learn and how they accomplish their prediction. Analysis of the different patterns in RAMs can provide a consistent description of how the network captures the weather effects in relation to the temporal progression of the crop, and can ultimately lead to a valid physical interpretation. The natural way to interpret RAMs is to regard it as a kind of derivative of the yield function in time. In the global pooling layer this function is then integrated out by the network to estimate the crop yield, which as a result reflects the accumulated effects of crop growth and meteorological conditions. Thus, the application of RAM allows for an instinctive analysis of how certain events in the time domain impact the estimated crop yield and enables a comparison with known drivers. In our case, RAMs were roughly reflecting the development of the crop and at the same time were sensitive to the variability of meteorological conditions, which somehow reflects the basic approach of relating yield to the accumulated biomass. Such an approach can increase the confidence in the model (as it captures the processes that are expected to be relevant), but might also draw attention to so far neglected features or model weaknesses. For example, although high temperatures leading to low yields are often considered in the Indian Wheat Belt, the importance of sunny conditions for good yields is rather neglected. On the other hand, RAMs varying towards the end of the analyzed time period might suggest that the model has not completely learned to ignore the time after the harvest. In general, the proposed methodology facilitates the application of DL in agriculture, not only to improve yield estimation and prediction but also to gain insight into the key drivers of crop yield.

## Acknowledgments

The work by AW has been funded by the joint project of International Cooperation and Exchange Programs between NSFC and DFG (41761134082). The work by GMG and LGC has been funded by the Spanish Ministry of Economy and Competitiveness (TEC2016-77741-R, ERDF). The MODIS and GLDAS data used in this study were acquired as part of the mission of NASA's Earth Science Division and archived and distributed by the Goddard Earth Sciences (GES) Data and Information Services Center (DISC).

## Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## ORCID iDs

Aleksandra Wolanin https://orcid.org/0000-0002-9029-6911
Gregory Duveiller https://orcid.org/0000-0002-6471-8404

## References

Akter N and Rafiqul Islam M 2017 Heat stress effects and management in wheat. A review *Agron. Sustain. Dev.* **37** 37
Badgley G, Field C B and Berry J A 2017 Canopy near-infrared reflectance and terrestrial photosynthesis *Sci. Adv.* **3** e1602244
Cai Y *et al* 2019 Integrating satellite and climate data to predict wheat yield in australia using machine learning approaches *Agric. For. Meteorol.* **274** 144–59
Crane-Droesch A 2018 Machine learning methods for crop yield prediction and climate change impact assessment in agriculture *Environ. Res. Lett.* **13** 114003
Duncan J M A, Dash J and Atkinson P M 2014 Elucidating the impact of temperature variability and extremes on cereal croplands through remote sensing *Glob. Change Biol.* **21** 1541–51
Duveiller E, Singh R P and Nicol J M 2007 The challenges of maintaining wheat productivity: pests, diseases, and potential epidemics *Euphytica* **157** 417–30
FAO 2017 The Future of Food and Agriculture—Trends and Challenges Rome, FAO
FAO, IFAD, UNICEF, WFP and WHO 2018 The State of Food Security and Nutrition in the World 2018. Building climate resilience for food security and nutrition *Technical Report* Rome, FAO
Fischer R A 1985 Number of kernels in wheat crops and the influence of solar radiation and temperature *J. Agric. Sci.* **105** 447–61
Gao B 1996 NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space *Remote Sens. Environ.* **58** 257–66
Global Information and Early Warning System on Food and Agriculture (GIEWS) 2019 Giews crop prospects and food situation quarterly global report #1 march 2019 *Technical Report* Food and Agriculture Organization of the United Nations (http://www.fao.org/3/i9553en/i9553en.pdf)
Goodfellow I, Bengio Y and Courville A 2016 *Deep Learning* (Cambridge, MA: MIT Press)

Gorelick N, Hancher M, Dixon M, Ilyushchenko S, Thau D and Moore R 2017 Google Earth Engine: planetary-scale geospatial analysis for everyone *Remote Sens. Environ.* **202** 18–27

Guan K, Wu J, Kimball J S, Anderson M C, Frolking S, Li B, Hain C R and Lobell D B 2017 The shared and unique values of optical, fluorescence, thermal and microwave satellite data for estimating large-scale crop yields *Remote Sens. Environ.* **199** 333–49

Hodson D P 2011 Shifting boundaries: challenges for rust monitoring *Euphytica* **179** 93–104

Iizumi T, Shiogama H, Imada Y, Hanasaki N, Takikawa H and Nishimori M 2018 Crop production losses associated with anthropogenic climate change for 1981-2010 compared with preindustrial levels *Int. J. Climatol.* **38** 5405–17

Jain M, Singh B, Srivastava A A K, Malik R K, McDonald A J and Lobell D B 2017 Using satellite data to identify the causes of and potential solutions for yield gaps in India's Wheat Belt *Environ. Res. Lett.* **12** 094011

Joshi A K, Mishra B, Chatrath R, Ferrara G O and Singh R P 2007 Wheat improvement in India: present status, emerging challenges and future prospects *Euphytica* **157** 431–46

Kaur P, Singh H, Rao V U M, Hundal S S, Sandhu S S, Nayyar S, Rao B B and Kaur A 2015 Agrometeorology of wheat in Punjab state of India (https://doi.org/10.13140/RG.2.1.5105.6721)

LeCun Y, Bengio Y and Hinton G 2015 Deep learning *Nature* **521** 436–44

Lobell D B, Ortiz-Monasterio J I, Asner G P, Matson P A, Naylor R L and Falcon W P 2005 Analysis of wheat yield and climatic trends in Mexico *Field Crops Res.* **94** 250–6

Lobell D B, Schlenker W and Costa-Roberts J 2011 Climate trends and global crop production since 1980 *Science* **333** 616–20

Lobell D B, Sibley A and Ortiz-Monasterio J I 2012 Extreme heat effects on wheat senescence in India *Nat. Clim. Change* **2** 186–9

Mbow C *et al* 2019 Food security *Climate Change and Land: an IPCC special report on climate change, desertification, land degradation, sustainable land management, food security, and greenhouse gas fluxes in terrestrial ecosystems* ed P R Shukta *et al* (Geneva: Intergovernmental Panel on Climate Change (IPCC)) (https://www.ipcc.ch/site/assets/uploads/sites/4/2019/11/08_Chapter-5.pdf) in press

Miller T 2019 Explanation in artificial intelligence: insights from the social sciences *Artif. Intell.* **267** 1–38

Montavon G, Samek W and Müller K-R 2018 Methods for interpreting and understanding deep neural networks *Digit. Signal Process.* **73** 1–15

Nash J and Sutcliffe J 1970 River flow forecasting through conceptual models: I. A discussion of principles *J. Hydrol.* **10** 282–90

Proctor J, Hsiang S, Burney J, Burke M and Schlenker W 2018 Estimating global agricultural effects of geoengineering using volcanic eruptions *Nature* **560** 480–3

Ray D K, Ramankutty N, Mueller N D, West P C and Foley J A 2012 Recent patterns of crop yield growth and stagnation *Nat. Commun.* **3** 1293

Reichstein M, Camps-Valls G, Stevens B, Jung M, Denzler J, Carvalhais N and Prabhat 2019 Deep learning and process understanding for data-driven Earth system science *Nature* **566** 195–204

Siebert S, Webber H and Rezaei E E 2017 Weather impacts on crop yields—searching for simple answers to a complex problem *Environ. Res. Lett.* **12** 081001

Singh S K, Saxena R, Porwal A, Neetu and Ray S S 2017 Assessment of hailstorm damage in wheat crop using remote sensing *Curr. Sci.* **112** 2095

Song L, Guanter L, Guan K, You L, Huete A, Ju W and Zhang Y 2018 Satellite sun-induced chlorophyll fluorescence detects early response of winter wheat to heat stress in the Indian Indo-Gangetic Plains *Glob. Change Biol.* **24** 4023–37

Tripathi A and Mishra A K 2017 The wheat sector in India: production, policies and food security *The Eurasian Wheat Belt and Food Security* ed Y Gomez *et al* (Cham: Springer) pp 275–96

Tucker C J, Townshend J R and Goff T E 1985 African land-cover classification using satellite data *Science* **227** 369–75

Wang Z, Yan W and Oates T 2017 Time series classification from scratch with deep neural networks: a strong baseline *Proc. of the IEEE IJCNN (Anchorage, AK, 14–19 May 2017)* pp 1578–85 (https://ieeexplore.ieee.org/document/7966039)

Wang Z and Yang J 2017 Diabetic retinopathy detection via deep convolutional networks for discriminative localization and visual explanation arXiv:1703.10757

Wu L 2009 *Mixed Effects Models for Complex Data* (Boca Raton, FL: CRC Press)

You J, Li X, Low M, Lobell D and Ermon S 2017 Deep Gaussian process for crop yield prediction based on remote sensing data *Proc. 31st AAAI Conf. on Artificial Intelligence (AAAI-17)* (https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14435)

Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A 2016 Learning deep features for discriminative localization *The IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* (https://ieeexplore.ieee.org/document/7780688)