Interpretable Long Short-Term Memory Networks for Crop Yield Estimation

Anna Mateo-Sanchis, Jose E. Adsuara, Maria Piles, *Senior Member, IEEE*, Jordi Muñoz-Marí, *Member, IEEE*, Adrián Pérez-Suay, and Gustau Camps-Valls, *Fellow, IEEE*

***Abstract*— Food security is at stake, with climate change heavily impacting agriculture and food production. In the present context of extreme events and changing conditions, developing advanced crop yield models can learn from all available information, and providing interpretable predictions for decision-making is of paramount relevance. This work explores the potential and limitations of developing interpretable crop yield models using long short-term memory (LSTM) neural networks, which typically excel at extracting information from time series. LSTMs were designed and trained with multisource satellite and meteorological time series over Continental US (CONUS) and corn, soybean, and wheat yield data from the US Department of Agriculture. Two recent attribution methods are used to interpret and extract knowledge from the developed models: integrated gradients (IG), based on back-propagation, and Shapley (SHAP) values, based on perturbations. Our results show that: 1) LSTM models achieved high accuracy ($R^2 > 0.56$); 2) multisource combinations outperformed single-variable models in all crop models; 3) both attribution methods were equivalent in detecting essential drivers and their contribution; 4) satellite estimates of enhanced vegetation index (EVI) and vegetation optical depth (VOD) together with meteorological estimates of maximum temperature (TMX) were the most relevant input features for crop yield estimations; and finally and 5) we discovered critical periods of the crop growth cycle for the corn, soybean, and wheat models. The suggested strategy constitutes an important step toward modeling and understanding crop production systems and advancing in automatic data-driven and accountable field management.**

***Index Terms*— Crop yield, integrated gradients (IG), interpretability, long short-term memory, remote sensing, Shapley (SHAP) values.**

I. Introduction

IN THE face of the global population increase and the current climate change emergency, better management of agricultural regions and crop production is critical. Future climate conditions and the frequency and intensification of extreme events are likely to reduce the production and availability of crops, worsening the food security situation [1]. In this context, robust crop yield estimations are essential to improve agricultural practices and policies. Yet, modeling and forecasting are not enough. Understanding the model predictions is further needed to support agricultural management decisions and to avoid catastrophic food safety conditions [2].

Earth observation (EO) provides vast, reliable, and up-to-date information for monitoring and describing crop status and health in near real-time. Several approaches based on remotely sensed and meteorological data for monitoring croplands and estimating yield exist now at various regional and continental scales [3], [4], [5], [6], [7]. Previous studies have relied on kernel machines and recently neural networks and have applied common architectures such as long short-term memory (LSTM) networks or convolutional neural networks (CNNs) for developing crop yield models with promising results [8], [9], [10].

Model interpretability is especially important when data-driven deep-learning DL networks are used to support decision-making [11]. It is crucial to analyze what a DL model is learning to comprehend how it works and, consequently, the rationale behind the predictions it suggests. There is a wide variety of approaches for extracting knowledge from a DL model [12], like *feature visualization*, to describe the network structure, *feature attribution*, to figure out how each feature contributed to a detailed forecast, and *model distillation*, to simplify the model. Despite this, only a few studies focus on interpretable crop yield models by using DL: the authors extracted knowledge from agricultural system dynamics during the crop season, focusing on the relevance of input variables and their influence on crop production using regression activation maps in a CNN architecture [8] and a permutation analysis in an LSTM [13].

In the present work, we investigate several strategies to create an NN-based learning system able to exploit satellite and meteorological data and understand model predictions in the US CONUS. Our goal is to shed light on what the trained LSTM learned from meteorological and satellite time-series and crop production data, including methods to evaluate the critical periods of the input variables affecting crop yields. We focus on two recent attribution methods: integrated gradients (IG) and Shapley (SHAP) values. Experimental results on the US CONUS using a wide diversity of meteorological and satellite variables give empirical evidence of the performance and viability of the proposal.

II. Data Collection

We collected crop yield statistics provided by the U.S. Department of Agriculture (USDA) at the county administra-

Manuscript received 14 November 2022; revised 3 February 2023; accepted 7 February 2023. Date of publication 10 February 2023; date of current version 1 March 2023. This work was supported in part by the Leaves Project under Grant RTI2018-096765-A-100 and Grant MCIU/AEI/FEDER and in part by the UE and AI for Complex Systems: Brain, Earth, Climate, Society Project under Grant CIPROM/2021/056. *(Corresponding author: Anna Mateo-Sanchis.)*

The authors are with the Image Processing Laboratory (IPL), Parc Científic, Universitat de València, Paterna, 46980 València, Spain (e-mail: anna.mateo@uv.es).

Digital Object Identifier 10.1109/LGRS.2023.3244064

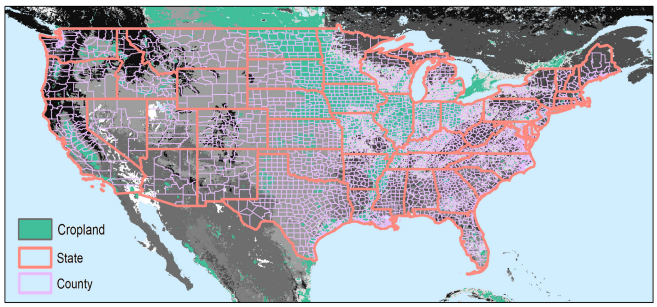

Fig. 1. Study area including state and county frontiers.

TABLE I
CHARACTERISTICS OF EO DATA USED FOR CROP YIELD ESTIMATION EXPERIMENTS

Product	Source	Pixel size	Revisit	Purpose
EVI	MOD13C1 v6	0.05°	16 days	Greenness
SM	SMAP	9 km	3 days	Soil moisture
VOD	SMAP	9 km	3 days	Crop water content
TMX	Daymet v3	1 km	monthly	Max. Temperature
PRE	Daymet v3	1 km	monthly	Precipitation
IGBP	MCD12C1 v6	0.05°	-	Land cover

tive level from 2015 to 2018 in the US CONUS (Fig. 1). In particular, we focused on yield data of the three most important kinds of cereal: corn, soybean, and wheat. The survey dataset has a total of samples of $N = 1744$, $N = 2060$, and $N = 1036$ for corn, soybean, and wheat, respectively. This yield data was converted from its original units of bushels per acre to tons per hectare (t/ha).

Furthermore, our dataset contains potential multisource remote sensing and meteorological products to characterize vegetation, soil, and atmospheric conditions. Particularly, the variables selected were important drivers in previous studies [6]: enhanced vegetation index (EVI) from optical satellites, soil moisture (SM), and vegetation optical depth (VOD) from microwave satellites, and a maximum temperature (TMX) and precipitation (PRE) from meteorological stations (additional data detailed in Table I) (data available at https://zenodo.org/record/7602711#.Y9z303bMJPY).

All variables selected were projected to a common grid of 9-km spatial resolution. Then, the IGBP land cover was used to identify the cropland regions in the study area, screening out the mixed and nonagricultural pixels. Cropland pixels were then collected and spatially county-averaged according to their geographic frontiers obtaining a mean time series. Finally, we matched the temporal frequency of all products to that of EVI; a 16-day composite was applied to SMAP satellite data, and a cubic interpolation to meteorological data. We then fixed a temporal window from April to October covering the crop season (13 temporal steps).

III. METHODS

LSTM architecture is commonly used for both regression and classification purposes involving some form of sequential data, including time series. This section summarizes its characteristics, functioning, and configurations tested for crop yield experiments. Additionally, we describe both interpretability methods used to extract knowledge from models trained.

A. LSTM Networks

LSTM is a special recurrent neural network (RNN) capable of learning long-term dependencies in sequential data. This kind of NN allows for exploiting the data temporal dimension efficiently, remembering past information for long periods. The LSTM model is based on gates for providing the ability to add or discard information to the cell state. The LSTM cell has three gate types: 1) the forget gate determines the information to be removed from the cell state; 2) the input gate decides the information that should be included in the cell state; and 3) the output gate determines the output of the LSTM cell.

In this work, we trained LSTMs with the following architecture: an LSTM layer and a fully connected layer. We established a fixed learning rate (0.001), 20 epochs, and online batch (i.e., batch size equals 1) size for training. Finally, we evaluated a different number of hidden units (HUs) (varying in size between 2 and 20) through the best cross-validation error using the stochastic gradient descent (SGB) optimizer. A larger number of HUs was explored but results did not improve noticeably.

B. Interpretability Methods

Interpretability has been identified as a potential weakness of NNs, particularly for geosciences [14], [15]. We employ two attribution methods to evaluate the input variable importance in the LSTM models: 1) the back-propagation-based method IG and 2) the perturbation-based method based on SHAP values.

1) Integrated Gradients: [16] IG allows attributing the latent representations of $\phi_d(\mathbf{x})$ at the output of the LSTM ϕ to each input point location x_n by integrating the gradient along the straight-line path in the input space from the baseline $\tilde{x}_n$ to the input x_n and then defines feature-wise scores

$$\mathrm{IG}_d(x_n) = (x_n - \tilde{x}_n) \cdot \int_{\alpha=0}^{1} [\nabla \phi_d(\tilde{\mathbf{x}} + \alpha \cdot (\mathbf{x} - \tilde{\mathbf{x}}))]_n d\alpha. \quad (1)$$

Note that if $\phi_d(\tilde{\mathbf{x}}) \approx 0$, the scores fulfill $\phi_d(\mathbf{x}) = \sum_n \mathrm{IG}_d(x_n)$, so the attribution values deploy a complete explanation.

2) SHAP Value Sampling: Given a specific data point, the goal is to decompose the model prediction and assign SHAP values [17] to individual features of the instance. Let $\mathcal{N} = \{1, \ldots, n\}$ be the finite set of players (or covariates in our case), each nonempty subset $\mathcal{S} \subseteq \mathcal{N}$ is called a *coalition* (optimal set of covariates) and $\mathcal{N}$ itself the *grand coalition*. A transferable utility game is defined by the pair $(\mathcal{N}, v)$, where $v : 2^{\mathcal{N}} \to \mathbb{R}$ is a mapping called the *characteristic function* or the *coalition function* of the game assigning a real number to each coalition and satisfying $v(\emptyset) = 0$. The SHAP value is a single-valued solution concept for cooperative games. The ith component of the single solution vector satisfying this solution concept for any cooperative game $(\mathcal{N}, v)$ is given by the following equation:

$$\phi_i^{\mathrm{Sh}} = \frac{1}{|\Pi(\mathcal{N})|} \sum_{\pi \in \Pi(\mathcal{N})} \left[v\left(\mathcal{P}_i^{\pi} \cup \{i\}\right) - v\left(\mathcal{P}_i^{\pi}\right) \right] \quad (2)$$

that is the average of all player's marginal contributions in permutation π. Therefore, the SHAP value of a player is the

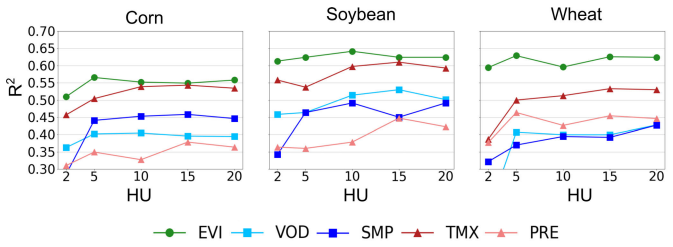

Fig. 2. Coefficient of determination (R^2) for single-variable models testing the number of HUs in LSTM architectures.

average marginal contribution of the player to the value of the predecessor set over every possible permutation of the player set.

The SHAP values of feature values are explanatory attributions to the input features. The problem is that the computation of SHAP values requires an exponential number of characteristic function evaluations (i.e., the number of machine-learning models to train), resulting in exponential time complexity being prohibitive in a machine-learning context. For this reason, various SHAP value approximations have been proposed. In this work, we applied the efficient implementation developed by [18] whose approximation of the SHAP value is possible in polynomial time.

C. Experimental Setup

First, we analyzed each remote sensing and climate product (i.e., variable) individually to define a ranking of variables per crop and source of data for crop yield estimation. We evaluated various configurations of LSTM to find an adequate model per variable depending on their statistical results in terms of error (RMSE) and accuracy (R and R^2). Second, each crop's products with the best score based on their data source (optical satellite sensor, microwave satellite sensor, and meteorological station) were selected for a synergistic multisource approach to modeling crop yield. We established a final combination of three variables per crop and evaluated different LSTM configurations as in the previous experiments using single variables. Finally, we employed IG and SHAP methods on the test sets of best models and compare the attribution scores for each input variable in each temporal step for understanding the LSTM models.

In all experiments, we separated three independent datasets to determine the best-performing configuration of the models. Data were randomly split into 70% for training and 30% for testing. Furthermore, a 20% into the training dataset was used for validation to optimize the network hyperparameters.

IV. Results

A. Crop Yield Estimation: Single-Source Approach

We evaluated the statistical results of a model with a single variable as input for crop yield estimation. The coefficient of determination (R^2) obtained for each model for a range of HUs (from 2 to 20) is given in Fig. 2. RMSE values (not shown) were coherent with the R^2 values. The best estimates were obtained in all crops and LSTM configurations by the EVI variable, reaching $R^2 > 0.5$ in all cases. These results emphasize the important relationship between crop greenness and yield.

TABLE II

Statistical Results Obtained for the LSTM Architectures for Crop Yield Estimation for Single and Multivariable Approaches [RMSE Units (t/ha)]

Crop	N	Var. Comb	HU	R	R^2	RMSE
Corn	1744	EVI	2	0.71	0.50	1.47
			5	**0.75**	**0.56**	**1.38**
			10	0.74	0.55	1.42
			15	0.74	0.55	1.43
		EVI+SM+TMX	2	0.72	0.52	1.48
			5	0.77	0.59	1.34
			10	**0.78**	**0.61**	**1.32**
			15	0.77	0.59	1.33
Soy	2060	EVI	2	0.78	0.61	1.42
			5	0.79	0.62	1.40
			10	**0.80**	**0.64**	**1.38**
			15	0.79	0.62	1.40
		EVI+VOD+TMX	2	0.76	0.58	1.52
			5	0.81	0.66	1.34
			10	0.82	0.67	1.33
			15	**0.83**	**0.69**	**1.28**
Wheat	1036	EVI	2	0.77	0.59	1.33
			5	**0.79**	**0.62**	**1.30**
			10	0.77	0.59	1.31
			15	0.79	0.62	1.32
		EVI+VOD+TMX	2	0.56	0.31	1.74
			5	0.74	0.55	1.42
			10	0.80	0.64	1.26
			15	**0.81**	**0.66**	**1.25**

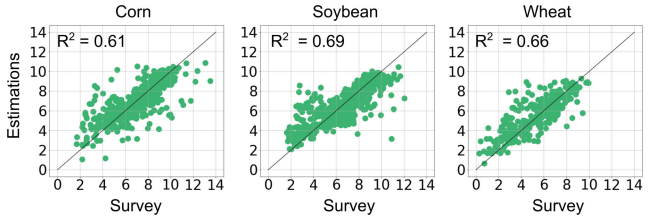

Fig. 3. Estimates versus survey data for the best LSTM models.

On the other hand, the high R^2 obtained for the maximum temperature (red lines) also stands out in the three crops studied. Even in some LSTM configurations, statistical results were comparable to EVI scores (e.g., 15 HUs for corn and soybean). It thus becomes the variable of meteorological source with the greatest relevance. It should also be noted that PRE came last in the variable's ranking in both corn and soybean crops. Nevertheless, models that used PRE were able to outperform R^2 scores of 0.36 in a reasonable number of configurations, especially for wheat models.

Interestingly, results revealed different behaviors for the two microwave products, depending upon the crop under study. On the one hand, soil moisture (SM) outperforms the VOD results in the majority of LSTM configurations for corn yield models. This result is probably due to the importance of water supply during the corn growth [19]. On the contrary, for both soybean and wheat models, VOD achieved better statistical results, although they were not noticeably superior in some configurations compared to SM.

B. Crop Yield Estimation: Multisource Approach

For the multisource approach, the most informative variable (greater R^2 values and lower RMSE values) according to its source was considered for further LSTM crop yield models. This led to the following variable combinations: 1) EVI +

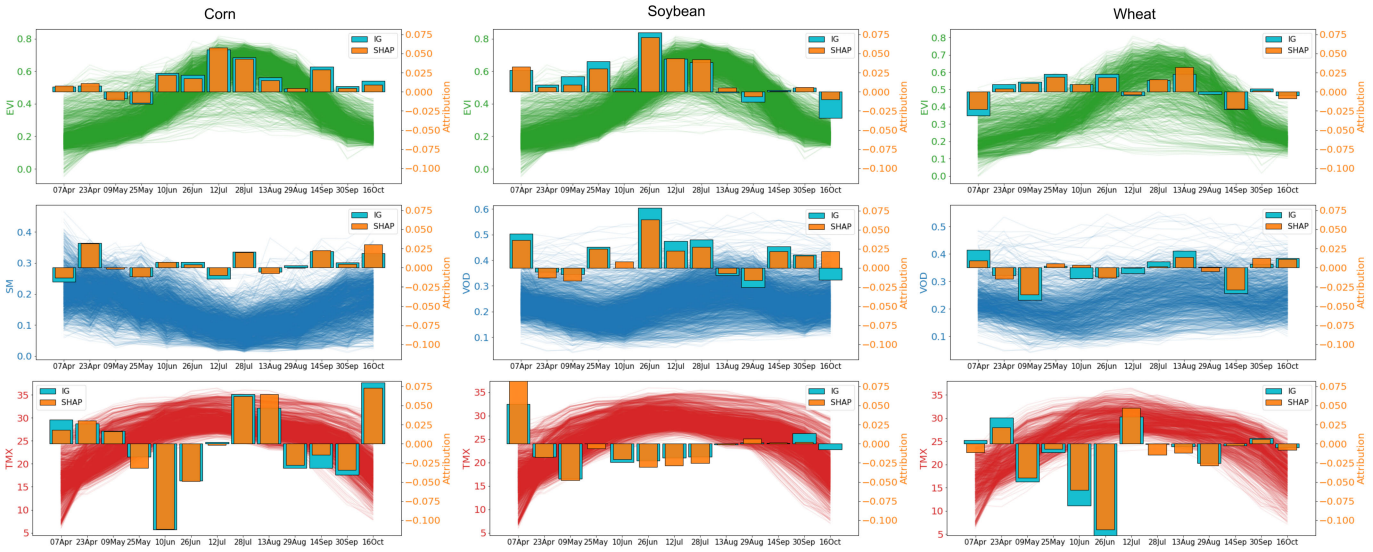

Fig. 4. Mean attribution scores provided for IG (cyan bars) and SHAP (orange bars) interpretability methods for multisource LSTM models (from left to right: corn, soybean, and wheat). The longer the bar length, the greater the contribution of this variable in the model. Furthermore, plots show the time series of different sources combined. Colors: green for EVI data, blue for microwave data, and red for TMX data.

SM + TMX for corn and 2) EVI + VOD + TMX for soybean and wheat.

Statistical results in terms of accuracy and error for both single and multisource approaches are displayed in Table II. Results indicate that multisource models lead to a better fit and more accurate models than single-variable models in all crops studied (higher R^2 and lower RMSE). These results align with previous studies using ML for crop yield estimation on large areas and reporting the benefit of integrating multisensor time information [6], [7]. The synergy of complementary information provided by different sources thus helped to represent better or mimic the agricultural environment, improving the model estimates. Finally, the results did not improve substantially when incorporating a larger number of HUs (accurate estimates were obtained from 5 HUs). Multisource models needed a greater number of HUs. In Fig. 3, yield estimations of the multisource models are shown. These estimations and their statistics suggest a good adjustment and low overfitting, the necessary conditions to motivate and continue the interpretability study.

C. Interpretability

We are interested in developing accurate crop yield models and, more importantly, interpreting and understanding how our LSTM models learn from the input data. In this regard, we carried out an interpretability study focused on test datasets from the multisource models achieved in the previous experiments (see Table II). First, we compared both attribution methods (IG and SHAP) with different foundations to check if they agree when assigning the importance of each variable in each temporal step. Then, we deeply analyzed the most important attributions established by the methods to figure out the most critical periods of the crop cycle and variables in our models.

We can observe the feature relevance in the LSTM multisource models for both attribution methods tested in Fig. 4. These graphs show the variable importance of each time step across the growing and senescence stages of crop development for crop yield estimation. The length of the bars indicates the contribution of the variable at a specific date which varies throughout the different phases of the crop season. Furthermore, it reveals that IG (propagation-based) and SHAP (perturbation-based) attributions are comparable in the same model. We can see how both blue and orange columns agree in most characteristics.

Generally, positive attributions were found, especially during the growing period (before autumn) for EVI data in all crops considered (first row in Fig. 4). It is reasonable that the EVI obtained important attributions, as crop greenness can be related to health and hence to high crop yield. However, the TMX variable achieved mostly negative attribution values during the crop season (third row). The sign of the attribution could be explained by noting that high-temperature values during the crop season are likely to reduce the performance of the fields [20]. About the microwave-based parameters (second row), water supply (SM) obtained positive attributions in more than half of the features throughout the crop season for the corn model but without an evident relevance. VOD, in turn, was important for soybean right before their maximum and at early and late phases. For wheat, however, VOD obtained very low attributions, and there is no clear behavior pattern.

For a more exhaustive analysis, we further examined the ten features of the highest relevance for the corn, soy, and wheat multisource models. Fig. 5 shows the results for the SHAP method; results obtained for IG are comparable (not shown). In this figure, we can observe which features (variables and a specific temporal step) are most relevant for the developed models, providing insights into key crop stages and associated plant/environmental drivers that could improve crop management approaches. We can observe the predominant importance of the TMX variable for the corn model, especially from June to the end of the season (EOS) and also the EVI relevance during its peak in July. SM was significant in April, during the emergence phase of the crop. For soybean, the three variables of the multisource models (EVI, VOD, and TMX) contribute to

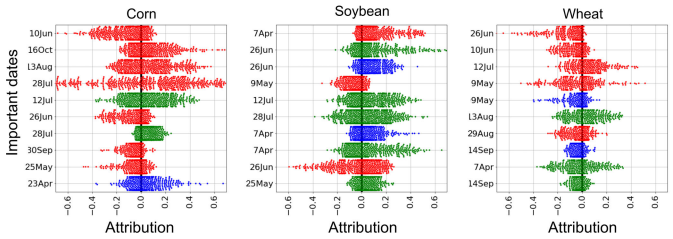

Fig. 5. Top 10 features sorted depending on their attributions provided by SHAP method in the multisource LSTM models. Each point represents the particular attribution of each input sample. The temporal steps are listed in the y-axis. Colors: green for EVI, blue for microwave data, and red for TEMP.

the early phases of the crop (April) and until midseason (July). EVI and VOD have their maximum relevance before attaining their maximum values. TMX is the most relevant feature for the wheat model, as in the corn model but with mostly negative attributions. Main TMX features are from May to July. VOD is negatively attributed in May and September when the VOD values decrease. Surprisingly, the EVI characteristics with the highest relevance were mostly negatively attributed.

V. Conclusion

Several LSTM architectures have been implemented in this work to achieve reliable and interpretable crop yield estimation models using multisensor satellite and meteorological data. Our experiments confirmed that a synergistic approach could potentially enhance yield prediction capabilities. Furthermore, our study sheds light on the current challenge of interpreting the crop yield models developed. Thus, we applied two attribution methods with different foundations to tackle the difficult task of interpretability: IG and SHAP. The results reveal the agreement of both techniques when they award the feature importance to each variable in each time step during the crop season. The proposed attribution techniques were able to learn about crop phenological progress and yield in the US CONUS from the input time series and output survey data. Hence, we can determine the critical time instants throughout the crop season. This interpretability analysis revealed the importance of EVI and VOD before their peak (about midseason), SM in early phases, and TMX with positive and negative attributions across the season, depending on the crop.

This study confirmed the feasibility of using deep-learning methods to predict crop yield and explain decisions, suggesting that models captured complex processes in agricultural systems. Our future work is tied to investigating the predictive power and explainability of convolutional architectures. Also, the IG and SHAP individual attributions and their intersections could be analyzed further for a more in-depth understanding of the attributions.

References

[1] D. Chaparro, M. Piles, M. Vall-Llossera, A. Camps, A. G. Konings, and D. Entekhabi, "L-band vegetation optical depth seasonal metrics for crop yield assessment," *Remote Sens. Environ.*, vol. 212, pp. 249–259, Jun. 2018.

[2] Z. Mehrabi et al., "Research priorities for global food security under extreme events," *One Earth*, vol. 5, no. 7, pp. 756–766, 2022.

[3] D. K. Bolton and M. A. Friedl, "Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics," *Agricult. Forest Meteorol.*, vol. 173, pp. 74–84, May 2013.

[4] R. López-Lozano et al., "Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: Strengths and limitations at pan-European level," *Agricult. Forest Meteorol.*, vol. 206, pp. 12–32, Jun. 2015.

[5] J. E. Adsuara, A. Perez-Suay, J. Munoz-Mari, A. Mateo-Sanchis, M. Piles, and G. Camps-Valls, "Nonlinear distribution regression for remote sensing applications," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 12, pp. 10025–10035, Dec. 2019.

[6] L. Martinez-Ferrer, M. Piles, and G. Camps-Valls, "Crop yield estimation and interpretability with Gaussian processes," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 12, pp. 2043–2047, Dec. 2021.

[7] A. Mateo-Sanchis et al., "Learning main drivers of crop progress and failure in Europe with interpretable machine learning," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 104, Dec. 2021, Art. no. 102574. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0303243421002816

[8] A. Wolanin et al., "Estimating and understanding crop yields with explainable deep learning in the Indian Wheat Belt," *Environ. Res. Lett.*, vol. 15, no. 2, 2020, Art. no. 024019.

[9] K. Kuwata and R. Shibasaki, "Estimating crop yields with deep learning and remotely sensed data," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2015, pp. 858–861.

[10] P. Nevavuori, N. Narra, and T. Lipping, "Crop yield prediction with deep convolutional neural networks," *Comput. Electron. Agricult.*, vol. 163, Aug. 2019, Art. no. 104859. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169919306842

[11] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, "Explainable machine learning for scientific insights and discoveries," *IEEE Access*, vol. 8, pp. 42200–42216, 2020.

[12] W. Samek, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, vol. 11700. Springer, 2019.

[13] A. Pérez-Suay et al., "Interpretability of recurrent neural networks in remote sensing," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Sep./Oct. 2020, pp. 3991–3994.

[14] D. J. Lary, A. H. Alavi, A. H. Gandomi, and A. L. Walker, "Machine learning in geosciences and remote sensing," *Geosci. Frontiers*, vol. 7, no. 1, pp. 3–10, 2016.

[15] M. Reichstein et al., "Deep learning and process understanding for data-driven Earth system science," *Nature*, vol. 566, pp. 195–204, Feb. 2019.

[16] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 3319–3328.

[17] L. Shapley, "A value for N-person games," in *Contributions to Theory Games*. 1953, pp. 307–317.

[18] J. Castro, D. Gömez, and J. Tejada, "Polynomial calculation of the Shapley value based on sampling," *J. Comput. Oper. Res.*, vol. 36, no. 5, pp. 1726–1730, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0305054808000804

[19] J. Ko and G. Piccinni, "Corn yield responses under crop evapotranspiration-based irrigation management," *Agricult. Water Manage.*, vol. 96, no. 5, pp. 799–808, May 2009.

[20] C. Zhao et al., "Temperature increase reduces global yields of major crops in four independent estimates," *Proc. Nat. Acad. Sci. USA*, vol. 114, no. 35, pp. 9326–9331, 2017.