Crop Yield Estimation and Interpretability With Gaussian Processes

Laura Martínez-Ferrer, *Graduate Student Member, IEEE*, María Piles, *Senior Member, IEEE*, and Gustau Camps-Valls, *Fellow, IEEE*

***Abstract*— This work introduces the use of Gaussian processes (GPs) for the estimation and understanding of crop development and yield using multisensor satellite observations and meteorological data. The proposed methodology combines synergistic information on canopy greenness, biomass, soil, and plant water content from optical and microwave sensors with the atmospheric variables typically measured at meteorological stations. A composite covariance is used in the GP model to account for varying scales, nonstationary, and nonlinear processes. The GP model reports noticeable gains in terms of accuracy with respect to other machine learning approaches for the estimation of corn, wheat, and soybean yields consistently for four years of data across continental U.S. (CONUS). Sparse GPs allow obtaining fast and compact solutions up to a limit, where heavy sparsity compromises the credibility of confidence intervals. We further study the GP interpretability by sensitivity analysis, which reveals that remote sensing parameters accounting for soil moisture and greenness mainly drive the model predictions. GPs finally allow us to identify climate extremes and anomalies impacting crop productivity and their associated drivers.**

***Index Terms*— CONUS, crop yield estimation, Gaussian processes (GPs), interpretability, modeling, moderate-resolution imaging spectroradiometer (MODIS), soil moisture active passive (SMAP).**

I. Introduction

RESEARCH and technological advances in the field of remote sensing have greatly improved the ability to detect and quantify the physical and biological stress that affects the productivity of agricultural crops as well as their status and evolution. Due to the exponential increase of the population in the last 50 years, the demand in crop production has increased, thus tripling the production of major cereals such as wheat and rice [1]. In the same vein, the Food and Agriculture Organization (FAO) of the United Nations estimates that 50% more food needs to be produced by 2050 [2].

In this context, the availability of data through Earth observation (EO) has opened new pathways for efficient agricultural mapping, crop monitoring, and evaluation. Among the available EO data, vegetation indices from optical sensors like Moderate-Resolution Imaging Spectroradiometer (MODIS) are widely used as proxies to crop productivity. Complementary information conveyed by passive microwave sensors, such as Soil Moisture Ocean Salinity (SMOS) and Soil Moisture Active Passive (SMAP), can contribute to an improved continuous crop monitoring [3]. In addition, ancillary meteorological variables, such as temperature or precipitation, influence crop growth, development, and final grain yield are of paramount relevance for monitoring crops too [4], [5]. Exploiting such wealth and diversity of information in an automated manner is a challenge in itself. In recent years, machine learning (ML) methods have promised improved accuracy in yield estimation, and many methods have been actually applied for the crop yield monitoring, from random forests (RFs) to neural networks and kernel machines [3], [6], [7].

The vast majority of studies focus, however, very little on understanding and interpreting model's predictions. Only recent works such as [8] have explored this issue. In this letter, we introduce Gaussian process (GP) models to address the problem of crop yield estimation and understanding jointly. We compare GPs with other standard methods for estimation of crop yield using informative drivers from optical and passive microwave sensors as well as meteorological variables. To deal with the particularities of the time series, we introduce a composite multisource GP which can deal with nonstationary and nonlinear processes. Three crops are considered (corn, wheat, and soybean) and data are collected for years 2015–2018 over continental U.S. (CONUS). In order to deal with such heterogeneous data sources efficiently, we also introduce the use of a sparse GP model that can scale to bigger data sets while still keeping high accuracy. More importantly, we provide a ranking of covariates to assess the relevance and synergy of remote sensing and meteorological data, as well as confidence intervals for the prediction and spatially explicit relevance maps of counties and years. Finally, the GP solution helps us identify extremes and anomalies and their associated drivers.

II. GPs for Modeling

GPs are nonparametric probabilistic approaches for ML problems, mainly for regression and classification. The GP regression method [9] has proved very good performance in biophysical parameter retrieval and model emulation [10], [11].

A. Notation

Let us fix notation first. Our goal is to learn a nonparametric function f able to estimate our target variable (crop yield) at county level $y \in \mathbb{R}$ from a set of D input features 0 (e.g., satellite and meteorological), $\mathbf{x} \in \mathbb{R}^D$. We assume an additive noise model $y = f(\mathbf{x}) + \varepsilon$, where the noise is additive independent and identically Gaussian distributed with zero

Manuscript received June 15, 2020; revised July 22, 2020; accepted August 7, 2020. Date of publication August 21, 2020; date of current version November 24, 2021. This work was supported in part by the European Research Council (ERC) through the ERC-CoG-2014 Project under Grant 647423 and in part by the Spanish Ministry of Science, Innovation and Universities (MCIU/AEI/FEDER, European Union (EU)) through the Project under Grant RTI2018-096765-A-100. The work of María Piles was supported by the Ramón y Cajal contract (MINECO). *(Corresponding author: Laura Martínez-Ferrer.)*

The authors are with the Image Processing Laboratory (IPL), Universitat de València, 46010 Valencia, Spain (e-mail: laura.martinez-ferrer@uv.es; maria.piles@uv.es; gustau.camps@uv.es).

Color versions of one or more of the figures in this letter are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/LGRS.2020.3016140

mean and variance σ_n, $\varepsilon \sim \mathcal{N}(0, \sigma_n^2)$. Let us define the stacked output values $\mathbf{y} = [y_1, \ldots, y_N]^\top$ denote the test points and predictions with a subscript asterisk $\mathbf{x}_*$ and y_*, respectively. Now, the output values are distributed according to

$$\begin{pmatrix} \mathbf{y} \\ f(\mathbf{x}_*) \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K} + \sigma_n^2\mathbf{I} & \mathbf{k}_* \\ \mathbf{k}_*^\top & k_{**} \end{pmatrix}\right) \tag{1}$$

where the covariance terms of the test point $\mathbf{k}_* = [k(\mathbf{x}_*, \mathbf{x}_1), \ldots, k(\mathbf{x}_*, \mathbf{x}_N)]^\top$, $k_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$ represents the self-similarity of $\mathbf{x}_*$, and $\mathbf{K}$ is the $N \times N$ kernel matrix that contains all pairwise similarities between counties i and j with entries $[\mathbf{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.

For prediction purposes, the GP model is obtained by computing the posterior distribution over the unknown output $\mathbf{y}_*$, $p(\mathbf{y}_*|\mathbf{x}_*, \mathcal{D})$, where $\mathcal{D} \equiv \{\mathbf{x}_n, y_n | n = 1, \ldots, N\}$ is the training data set. This posterior can be shown to be a Gaussian distribution, $p(y_*|\mathbf{x}_*, \mathcal{D}) = \mathcal{N}(y_*|\mu_{\text{GP}*}, \sigma^2_{\text{GP}*})$, for which one can estimate the *predictive mean* (point-wise predictions) and the *predictive variance* (confidence intervals):

$$\mu_{\text{GP}*} = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{y} = \mathbf{k}_*^\top \boldsymbol{\alpha} \tag{2}$$

$$\sigma^2_{\text{GP}*} = k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma_n^2\mathbf{I}\right)^{-1}\mathbf{k}_* \tag{3}$$

where $\boldsymbol{\alpha}$ are model weights.

B. Interpretability With GP Models

GPs are not black boxes. They allow not only modeling but also understanding parts of the problem in different ways. First, note that, after optimization, a set of parameters $\alpha_i \in \mathbb{R}$, $i = 1, \ldots, N$ are learned for each one of the N training data examples $\mathbf{x}_i$, indicating their relevance. Second, the predictive variance $\sigma^2_{\text{GP}*}$ tell us about the confidence in the prediction, which can be useful to identify anomalous cases as well as to characterize extrapolation regimes. Third, GPs allow us to optimize several hyperparameters efficiently and hence use flexible kernel functions with *interpretable* parameters. Many kernel functions are available [12]. In this work, we propose the linear combination of kernels to account for different signal characteristics. In particular, we use

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^\top \mathbf{x}_j + \nu \exp\left(-\frac{1}{2\sigma^2}\|\mathbf{x}_i - \mathbf{x}_j\|^2\right) + \sigma_n^2\delta_{ij} \tag{4}$$

where ν is a scaling factor, σ is a dedicated parameter controlling the spread of the signal relations, σ_n is the noise standard deviation, and δ_{ij} is the Kronecker's symbol. Note that the kernel function is a combination of three kernels: a linear kernel to cope with linear features and to mimic the best linear decision, a standard squared exponential kernel to deal with locality and nonlinearities to modify the linear solution, and a noise term to regularize the solution.

Finally, the GP model can be further scrutinized following a sensitivity analysis. We suggest here the ranking features from a trained GP model by evaluating the impact of the inputs on the prediction error in the context of the other predictors. Essentially, for each feature d (or set of features associated with a particular variable), the algorithm sets to zero their values for all training samples and evaluates the prediction root-mean-square error (RMSE), which can be cast as sensitivity (relevance) of that variable. A normalized ranking of feature t is thus reported $r_d = \text{RMSE}_d / \sum_{d=1}^{D} \text{RMSE}_t$. In [12], we illustrated the usefulness of this procedure for the identification of the most relevant spectral channels for the retrieval of vegetation parameters from hyperspectral data.

C. Fast and Sparse GP Models

One of the major drawbacks of GPs is dealing with a high volume of training data. A naïve implementation requires computation which grows as $\mathcal{O}(N^3)$, where N is the number of training points. Moreover, the memory requirements to store the covariance matrix $\mathbf{K}$ grows as $\mathcal{O}(N^2)$. This makes GPs impractical when the training set is in the range of 100 000 points. Among the many available methods to improve efficiency of GPs [13], we suggest the "subset of regressors" (SoR) approach. The idea is to define the approximate solution as a function of a set of $M \ll N$ inducing variables $\mathbf{u}$. Its corresponding prior is $p(\mathbf{u}) = \mathcal{N}(0, \mathbf{K}_{M,M})$, which is the same that is set for the noise-free variables $\mathbf{f}$ of the standard GP but using the inducing inputs points. The notation $\mathbf{K}_{N,M}$ indicates a matrix $\mathbf{K}$ of N rows and M columns.

The SoR method establishes the *deterministic* relation

$$p_{\text{SoR}}(\mathbf{f}|\mathbf{u}) = \mathcal{N}\left(\mathbf{K}_{N,M}\mathbf{K}_{M,M}^{-1}\mathbf{u}, 0\right)$$

and now integrating $\mathbf{u}$, gives the prior over $\mathbf{f}$: $p_{\text{SoR}}(\mathbf{f}) = \mathcal{N}(0, \mathbf{Q}_{N,N})$, where $\mathbf{Q}_{N,N} = \mathbf{K}_{N,M}\mathbf{K}_{M,M}^{-1}\mathbf{K}_{M,N}$. The posterior $p_{\text{SoR}}(f^*|\mathbf{x}_*, \mathcal{D})$ for a new input point $\mathbf{x}_*$ is given by

$$\mu_{\text{SoR}*} = \mathbf{k}_{*,M}(\mathbf{K}_{M,N}\mathbf{K}_{N,M} + \sigma^2\mathbf{K}_{M,M})^{-1}\mathbf{K}_{M,N}\mathbf{y} \tag{5}$$

$$\sigma^2_{\text{SoR}*} = \sigma^2\mathbf{k}_{*,M}(\mathbf{K}_{M,N}\mathbf{K}_{N,M} + \sigma^2\mathbf{K}_{M,M})^{-1}\mathbf{k}_{M,*}. \tag{6}$$

This solution replaces $\mathbf{K}$ in (2) and (3) by $\mathbf{Q}$ in the standard GP posterior, which also follows from the Nyström approximation [14]. The computational complexity for training is $\mathcal{O}(NM^2)$, and $\mathcal{O}(M)$ for computing the predictive mean and $\mathcal{O}(M^2)$ for the predictive variance.

D. Inference and Bayesian Optimization

The hyperparameters of the GP model $\boldsymbol{\theta} = \{\nu, \sigma, \sigma_n\}$ are typically inferred by Type-II Maximum Likelihood, using the marginal likelihood (also called evidence) of the observations, which is also analytic (explicitly conditioning on $\boldsymbol{\theta}$)

$$\log p(\mathbf{y}|\theta, \sigma_n) = \log \mathcal{N}\left(\mathbf{y}|\mathbf{0}, \mathbf{K} + \sigma_n^2\mathbf{I}\right) \tag{7}$$

and since its derivatives are analytic, the conjugate gradient ascend is typically used for optimization. However, this procedure may end in local minima, mainly in high-noise regimes and low-to-moderate number of examples. A possibility to alleviate this would be running and sampling the GP with different initial conditions, but this can be very expensive. Instead, we searched the optimal hyperparameters using a more robust Bayesian optimization procedure [15]. We provide operational code snippets and data in https://isp.uv.es/gp4crops/ website for the sake of reproducibility.

III. Data Collection and Preprocessing

The USDS National Agricultural Statistics Services (data access: http://quickstats.nass.usda.gov/USDA-NASS) provides exhaustive surveys on crops at the county, district, and state levels across the U.S. It includes production information per crop type (e.g., area planted, area harvested, yield) as well as information of crop progress, such as the days of planting and harvest and the dates at which crops had reached specific growth stages (i.e., emergence, bloom, dropping leaves). This high level of detail makes CONUS an ideal experimental site for large-scale crop yield studies. For this work, we collected

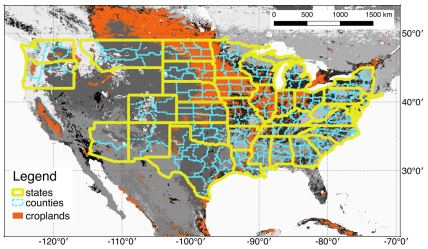

Fig. 1. Study area including the 35 states reporting soybean, corn, and wheat production during years 2015–2018 and cropland mask following the MODIS International Geosphere-Biosphere Programme (IGBP) land cover classification.

TABLE I
DATA DESCRIPTION

Product	Source	Grid	Temp. res.	Purpose
Satellite				
MODIS/EVI	MOD13C1 v6	0.05°	16 d	Phenology
VOD	SMAP [17]	EASE2 9 km	3 d	Growth
SM	SMAP [17]	EASE2 9 km	3 d	Moisture
MODIS/IGBP	MCD12C1 v6	0.05°	-	Area
Meteorologic				
T_{max}	Daymet v3	1 km x 1 km	Monthly	Temp.
Precip	Daymet v3	1 km x 1 km	Monthly	Precip.

information on soy, corn, and wheat yields (t/ha) from USDA-NASS for years 2015–2018 from 35 states, at a county level (see Fig. 1).

Time series from three satellite-based bio-geophysical variables for the study period were selected: the enhanced vegetation index (EVI) to account for vegetation chlorophyll content, the vegetation optical depth (VOD), sensitive to aboveground biomass water-uptake dynamics, and soil moisture (SM) which provides direct information of surface water conditions. We also included monthly maximum temperature ($T_{\max}$) and precipitation (*Precip*) from meteorological stations. These optical, microwave, and meteorological features constitute potential indicators of the vegetation state and drivers for crop yield estimation as shown elsewhere [3], [16]. Table I summarizes the data used in this study. Variables were first projected to a common grid (EASE2 9 km) and then a crop mask was used to identify the croplands in the study region and screen out data from mixed and nonagricultural pixels (see Fig. 1).

Satellite and meteo pixel-based information was related to the survey data at the county scale as follows. First, pixels from each county were extracted according to its geographic boundaries given by shapefile polygons and then spatially averaged to produce time series of EVI, VOD, SM, $T_{\max}$, and *Precip*. We fixed a temporal observational window from April to October of each year, which includes the growing and senescence stages of all crops in the study area. Our experimental setup takes all the available temporal information following crop progress to estimate each year's yield. All years were considered altogether to learn a single model per crop. Since we work with time series from multiple sources, different variables (covariates) have different lengths L (i.e., temporal dimension): EVI ($L = 13$), VOD, and SM ($L = 214$), $T_{\max}$ and *Precip* ($L = 7$). Stacking all the observations would make the problem intractable since the dimensionality would increase largely. Therefore, we applied

TABLE II
RESULTS FOR DIFFERENT MODELS (LR, RF, AND GP), SETS OF VARIABLES, AND MEASURES (MEAN ERROR, ME [t/ha]; ROOT-MEAN-SQUARE ERROR, RMSE [t/ha]; AND PEARSON'S CORRELATION COEFFICIENT R) OVER AN INDEPENDENT TEST SET FOR THE MARGINAL CONFIGURATION

Model	D	LR			RF			GP		
		ME	RMSE	R	ME	RMSE	R	ME	RMSE	R
Corn	N=1744									
EVI	11	0.01	1.60	0.68	0.02	1.46	0.75	0.03	1.43	**0.76**
VOD	10	0.01	1.86	0.53	0.01	1.65	0.67	-0.01	1.48	**0.74**
SM	27	0.01	1.77	0.59	-0.02	1.48	0.76	0.02	1.35	**0.79**
T_{max}	6	0.01	1.76	0.60	0.01	1.39	0.78	0.02	1.27	**0.82**
Precip	7	0.01	2.03	0.38	-0.01	1.72	0.62	0.01	1.65	**0.66**
Soy	N=2060									
EVI	11	-0.01	1.53	0.72	-0.01	1.41	0.77	-0.01	1.34	**0.79**
VOD	10	-0.01	1.81	0.57	-0.01	1.55	0.72	0.01	1.41	**0.76**
SM	27	-0.01	1.76	0.60	-0.02	1.45	0.77	-0.02	1.31	**0.80**
T_{max}	6	-0.01	1.73	0.62	-0.03	1.28	0.82	-0.01	1.20	**0.84**
Precip	7	-0.01	2.05	0.36	-0.02	1.68	0.65	0.00	1.58	**0.69**
Wheat	N=1036									
EVI	11	0.05	1.50	0.72	0.04	1.38	0.78	0.04	1.36	**0.78**
VOD	10	0.05	2.01	0.38	0.04	1.72	0.62	-0.01	1.60	**0.67**
SM	27	0.05	1.83	0.54	0.02	1.58	0.70	0.03	1.44	**0.75**
T_{max}	6	0.05	1.75	0.59	0.03	1.45	0.74	0.02	1.38	**0.77**
Precip	7	0.05	1.87	0.51	0.02	1.65	0.65	0.02	1.59	**0.68**

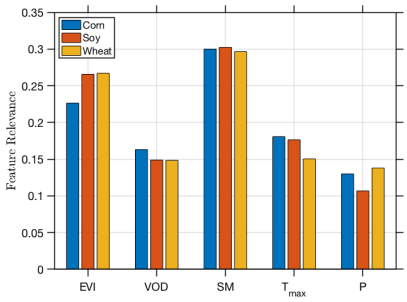

Fig. 2. Sensitivity analysis per crop.

a Principal Component Analysis (PCA) per variable to retain the same amount of variance each (95%) and stacked the projected data instead. This allowed us to project the time series in a lower-dimensional representation: EVI ($D = 11$), VOD ($D = 10$), SM ($D = 27$), $T_{\max}$ ($D = 6$), and *Precip* ($D = 7$). These components were finally stacked (leading to a final, more tractable dimensionality $D = 61$) and used for regression. A data set was obtained for each crop type, with a total of $N = 1744$ samples for corn, $N = 2060$ for soy, and $N = 1036$ for wheat.

IV. EXPERIMENTAL RESULTS

We are interested in learning accurate and efficient estimation models of crop yield, and more importantly, in inferring the most relevant drivers and climate anomalies from the proposed model.

To do this, we first compare several standard regression models (ordinary least squares linear regression (LR), RF, and the proposed GP composite model) for each variable. This helps us learning what are the most important variables and models in isolation. Second, we derive global regression models fed with all considered variables. This synergistic combination of multisensor data in the model is studied in terms of robustness, confidence intervals, sparsity and accuracy, and also by looking at the relative relevance of the drivers when considered jointly.

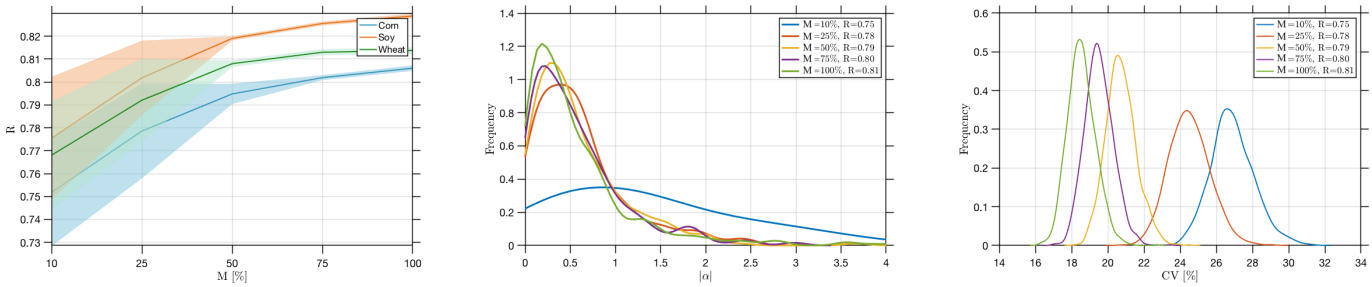

Fig. 3. (Left) Accuracy for all crop-specific GP models, (middle) density of model weights, and (right) confidence intervals for different levels for sparsity for the corn model. The number of total samples for this experiment is $N = 500$, and we vary the rate of inducing points, $M[\%]$.

TABLE III

RESULTS FOR DIFFERENT MODELS (LR, RF, AND GP), SETS OF VARIABLES AND MEASURES (MEAN ERROR, ME [t/ha]; ROOT-MEAN-SQUARE ERROR, RMSE [t/ha]; AND PEARSON'S CORRELATION COEFFICIENT R) OVER AN INDEPENDENT TEST SET FOR THE JOINT CONFIGURATION

Model	D	LR			RF			GP		
		ME	RMSE	R	ME	RMSE	R	ME	RMSE	R
EVI+VOD+SM	48									
Corn		-0.01	1.43	0.75	0.01	1.29	0.81	0.01	1.21	**0.83**
Soy		-0.02	1.35	0.78	-0.02	1.25	0.83	-0.03	1.15	**0.85**
Wheat		-0.01	1.41	0.76	-0.05	1.30	**0.80**	-0.02	1.29	**0.80**
T_{max}+*Precip*	13									
Corn		0.01	1.62	0.68	-0.01	1.37	0.79	0.00	1.30	**0.81**
Soy		-0.06	1.62	0.67	-0.04	1.27	0.82	-0.04	1.20	**0.83**
Wheat		0.01	1.61	0.67	-0.04	1.38	0.78	-0.04	1.32	**0.79**
Sat+Meteo	61									
Corn		-0.03	1.41	0.77	-0.03	1.27	0.83	-0.01	1.20	**0.84**
Soy		0.01	1.30	0.80	-0.01	1.19	0.84	0.01	1.09	**0.86**
Wheat		0.04	1.36	0.78	0.01	1.26	0.83	0.03	1.23	**0.83**

A. Marginal Approach

We analyze the information content of each individual variable for predicting crop yield separately. Results are shown in Table II. The GP model generally outperforms the rest, although results are similar to RF. In all cases, the maximum temperature gives the higher R values for any crop and regression model. In the case of corn crops, the highest R value obtained ($R = 0.82$) confirms the already demonstrated strong dependence of corn growth with temperature. Corn develops faster in warmer weather and slower in cold seasons.[1] For soy and wheat, the highest R values were obtained with $T_{\max}$ and SM, and with $T_{\max}$ and EVI, respectively. All variables can be considered informative ($R > 0.6$) and will be further considered jointly.

B. Joint Approach

Table III shows the results for all considered variable combinations: satellite only (EVI, VOD, SM), meteorological only ($T_{\max}$, *Precip*), and altogether. An important first conclusion is that, independent of the considered combination, results always improve over the marginal approach, thus suggesting that variables convey complementary information (cf. with results in Table II). Having information on the content of chlorophyll (EVI), water stress (VOD), and SM further facilitates obtaining a useful model for crop yield estimation, thus confirming the diversity in the variables nature always helps, independent of the considered crop. Also, note that, although the values obtained for precipitation have been lower than the rest of the variables (e.g., $R = 0.65$), its combination with the maximum temperature achieves very good performances, thus suggesting that meteo information could suffice. Nevertheless, we finally observe a clear improvement when meteo and remote sensing variables are combined in the models, leading to better fitted (higher R), more accurate (lower RMSE [t/ha]), and less biased (lower ME [t/ha]) models.

[1]https://www.agry.purdue.edu/ext/corn/news/timeless/heatunits.html

C. Interpretability

In all previous cases, GP models have outperformed the rest. Here we suggest several techniques to interpret what the model learned, by looking at the variables sensitivity, as well as model sparsity and the distribution of confidence intervals.

1) Sensitivity Analysis: The sensitivity analyses are shown in Fig. 2. For all three crops, the features that present a higher RMSE and hence sensitivity (relevance) are EVI and SM. It is observed that for all cases, the precipitation variable is less relevant compared to the others, as already seen in Section IV-A Table II. However, meteorological features, such as $T_{\max}$ and *Precip*, along with VOD generally reveal a considerable impact. Results confirm our observations in [3].

2) Sparsity, Accuracy, and Confidence Intervals: We run the sparse GP model for different inducing points M (10%, 25%, 50%, 75%, 100%) for all joint crop-specific models. Training was run 100 times and results are averaged in Fig. 3 (left). All models were trained using a Bayesian optimization approach with a 25% hold-out validation set. Results show that the average performance is constant for $M > 50\%$. Results degrade when using only 25% of inducing variables, but only by 7% in R. These results suggest that sparse GPs constitute an efficient way to obtain efficient GP models without sacrificing accuracy. Fig. 3 (middle) shows the density of GP model weights $|\boldsymbol{\alpha}|$ for different sparsity levels. As less inducing points M are used (higher sparsity imposed), the distribution becomes flatter, thus suggesting that all points become equally relevant, while for bigger data sets a high degree of "virtual sparsity" ($\boldsymbol{\alpha} \approx 0$) is achieved. Note, for instance, the heavy tails and deemed similar densities for $M > 25\%$. Interestingly, Fig. 3 (right) reveals a clear tradeoff between sparsity and confidence: Even if a low M can be prescribed to achieve fast and acceptable results, the associated confidence intervals are too wide in such cases, leading to coefficients of variation, $\text{CV} = 100 \times (\sigma_{\text{GP}*}/\mu_{\text{GP}*})$, larger than 25% for $M < 50\%$, which is unacceptable and does not meet the Global Climate Observing System (GCOS) and FAO recommendations of 20%. On the contrary, considering $M > 50\%$ of inducing points, accuracy becomes stable ($R \sim 0.80$) and confidence intervals are below the GCOS prescription.

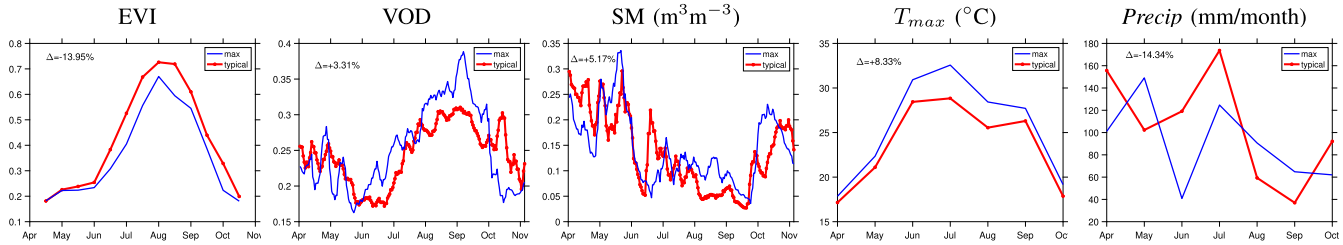

Fig. 4. Time series for the different variables of the counties with highest (most anomalous) and lowest (typical) α values.

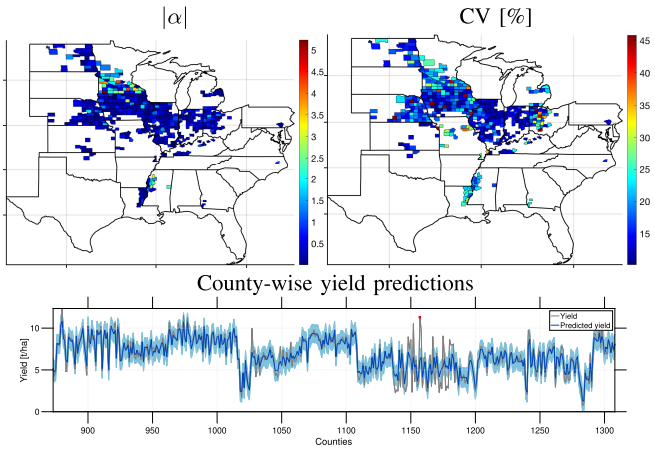

Fig. 5. Results for 2017. Top: maps of (left) α and (right) CV [%]. Bottom: county-wise actual and predicted yield with red-spotted anomaly corresponding to the Nobles county (Minnesota). Shaded region indicates the predictive variances.

D. Learning Anomalies and Hotspots

Model weights $\boldsymbol{\alpha}$ in GPs provide information about the errors of the predictions and can serve as good indicators of anomalies and hotspots. In our corn GP model, the highest $\boldsymbol{\alpha}$ values for all the years correspond to counties of the state of Minnesota, being the highest α (most anomalous) assigned to the county of Nobles for 2017. Fig. 4 shows the time series of the Nobles county and a typical county with good predictions (modal α value). That year, the temperature in Nobles was higher than in the typical county (+8.33%), associated with lower precipitation (−14.3%) and greenness (EVI was −13.95%). Actually, a considerable decrease in precipitation was observed for the summer period, where values from moderate to severe drought for the neighbor state of Iowa were recorded. The differences in VOD and SM were minimal on average ($\Delta = 5.17\%$), which probably led to poor results (high error, high α value) for this particular county. This fact, associated with high confidence (low CV[%]) makes this an anomalous county for the GP model. Maps of $\boldsymbol{\alpha}$ and CV[%] values, along the county yield and predictive variances, are shown in Fig. 5.

V. Conclusion

We introduced GPs for the estimation and understanding of crop development and yield estimation using multisensor satellite observations and meteorological data. The GP model gave noticeable gains in terms of accuracy and robustness with respect to other ML approaches for the estimation of corn, wheat, and soybean yields consistently for four years of data over continental U.S. Several strategies for understanding the main drivers of crop yield were introduced; sensitivity analysis revealed that SM and EVI were important drivers, and sparse GPs reported robust and fast results. Further work will consider assessing model's transportability with multitask GPs to Europe, where a higher crop diversity and landscape fragmentation are present.

References

[1] H. C. J. Godfray *et al.*, "Food security: The challenge of feeding 9 billion people.," *Science*, vol. 327, no. 5967, pp. 812–818, 2010.

[2] *The Future of Food and Agriculture—Trends and Challenges*, FAO, Rome, Italy, 2017.

[3] A. Mateo-Sanchis, M. Piles, J. Muñoz-Marí, J. E. Adsuara, A. Pérez-Suay, and G. Camps-Valls, "Synergistic integration of optical and microwave satellite data for crop yield estimation," *Remote Sens. Environ.*, vol. 234, Dec. 2019, Art. no. 111460.

[4] S. Siebert, H. Webber, and E. E. Rezaei, "Weather impacts on crop yields–searching for simple answers to a complex problem," *Environ. Res. Lett.*, vol. 12, no. 8, Aug. 2017, Art. no. 081001.

[5] N. Akter and M. Rafiqul Islam, "Heat stress effects and management in wheat. A review," *Agronomy Sustain. Develop.*, vol. 37, no. 5, p. 37, Aug. 2017.

[6] D. B. Lobell, W. Schlenker, and J. Costa-Roberts, "Climate trends and global crop production since 1980," *Science*, vol. 333, no. 6042, pp. 616–620, May 2011.

[7] T. Iizumi, H. Shiogama, Y. Imada, N. Hanasaki, H. Takikawa, and M. Nishimori, "Crop production losses associated with anthropogenic climate change for 1981-2010 compared with preindustrial levels," *Int. J. Climatol.*, vol. 38, no. 14, pp. 5405–5417, Aug. 2018.

[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, pp. 1–12, 2020.

[9] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*. New York, NY, USA: MIT Press, 2006.

[10] G. Camps-Valls, J. Verrelst, J. Muñoz-Marí, V. Laparra, F. Mateo-Jimenez, and J. Gomez-Dans, "A survey on Gaussian processes for Earth-observation data analysis: A comprehensive investigation," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 58–78, Jun. 2016.

[11] G. Camps-Valls, D. Sejdinovic, J. Runge, and M. Reichstein, "A perspective on Gaussian processes for Earth observation," *Nat. Sci. Rev.*, vol. 6, no. 4, pp. 616–618, Mar. 2019.

[12] J. Verrelst, L. Alonso, G. Camps-Valls, J. Delegido, and J. Moreno, "Retrieval of vegetation biophysical parameters using Gaussian process techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 5, pp. 1832–1843, May 2012.

[13] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. S. Lawrence, *Dataset Shift in Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[14] C. Williams and M. Seeger, "Using the nyström method to speed up kernel machines," in *Proc. Adv. Neural Inf. Process. Syst.*, Cambridge, MA, USA: MIT Press, 2001, pp. 682–688.

[15] D. H. Svendsen, L. Martino, and G. Camps-Valls, "Active emulation of computer codes with Gaussian processes—Application to remote sensing," *Pattern Recognit.*, vol. 100, Apr. 2020, Art. no. 107103.

[16] J. E. Adsuara, A. Pérez-Suay, J. Muñoz-Marí, 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.

[17] A. G. Konings, M. Piles, N. Das, and D. Entekhabi, "L-band vegetation optical depth and effective scattering albedo estimation from SMAP," *Remote Sens. Environ.*, vol. 198, pp. 460–470, Sep. 2017.