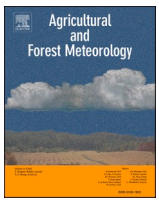
Agricultural and Forest Meteorology 351 (2024) 110022

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

journal homepage: www.elsevier.com/locate/agrformet



Beyond assimilation of leaf area index: Leveraging additional spectral information using machine learning for site-specific soybean yield prediction

Deborah V. Gaso [a,d,*], Dilli Paudel [a], Allard de Wit [b], Laila A. Puntel [c], Adugna Mullissa [a], Lammert Kooistra [a]

[a] Laboratory of Geo-Information Science and Remote Sensing, Wageningen University and Research, Wageningen 6708PB, the Netherlands
[b] Wageningen Environmental Research, Wageningen 6708PB, the Netherlands
[c] Department of Agronomy and Horticulture, University of Nebraska-Lincoln, Keim Hall, 1825N 38th Street, Lincoln, 68583-0915, NE, United States
[d] Instituto Nacional de Investigación Agropecuaria, Colonia CP 70006, Uruguay

ARTICLE INFO

Keywords:
Soybean
Remote sensing
Data assimilation
Crop modeling
Machine learning

ABSTRACT

Assimilating external observations of crop state in cropping system models is essential for making spatially explicit predictions of crop variables relevant in precision agriculture. Satellite-based leaf area index (LAI) estimates have been the most frequent variable used as a proxy of actual crop growth. However, additional information beyond LAI, like canopy N content, water content, and structure, can be retrieved from satellite observations. Including such variables by data assimilation directly is difficult because many crop models do not have corresponding state variables or the relationship between the observations and the process that regulates crop growth is unclear. Therefore, other approaches are required to include such information. In this study, we investigate the improvement in the predicted yield and feature impact on model outputs by using a hybrid approach that combines observations from Sentinel-1 and 2 time-series with the outputs from a process-based model embedded in a data assimilation framework and uses the Gradient-boosted trees regressor (GBTR) as predictive model. We used two regions with soybean fields: the US (13 K points) and Uruguay (400 K points). We found an advantage when using the GBTR as the predictive model (reduced RRMSE by ~16%) compared to data assimilation. Adding the vegetation indices had a marginal improvement (reduced RRMSE by ~1%), while the impact of adding reflectance and backscatter values was negative. The satellite-based features had a very small importance score, while features' impact on prediction was predominantly unclear, explaining the marginal predictive power added by satellite-based features. We found that features from the reproductive stages had the highest importance, while the importance of an index related to drought stress (NMDI) across the growing season provided insights for further improvement of data assimilation methods. However, more studies are required to better disentangle pathways towards further improvement in constraining crop models by ingesting satellite observations.

1. Introduction

Developing methods to estimate spatial variability in crop yield is of great relevance to address site-specific crop management and to improve resource use efficiency. Efficient site-specific crop management strategies could substantially enhance input efficiency and reduce yield gaps without intensifying the used inputs (fertilizer, pesticide, etc.), which would ensure the sustainability of the agricultural systems (Cassman and Grassini, 2020). Crop models are valuable tools because they describe the interaction between crop traits, management, growth, and environmental factors. However, applying crop growth models for site-specific crop management is usually constrained by the lack of input data available at high spatial resolution (e.g., soil parameterization) and crop management information, which introduces a high level of uncertainties in the simulated quantities (Dokoohaki et al., 2021; Folberth et al., 2016). Thus, the assimilation of remotely sensed variables into a crop model has received a lot of attention as a way to solve the lack of spatial input data required for the spatial application of crop growth

* Corresponding author.
E-mail addresses: deborah.gasomelgar@wur.nl, dgaso@inia.org.uy (D.V. Gaso).

https://doi.org/10.1016/j.agrformet.2024.110022
Received 24 October 2023; Received in revised form 6 February 2024; Accepted 18 April 2024
Available online 21 April 2024

models (Huang et al., 2019; Jin et al., 2018; Jindo et al., 2023). The key role of data assimilation consists in compensating for the missing input data and in providing information on actual crop growth status which captures the impact of abiotic stresses and yield reduction factors which are not included in the model (Huang et al., 2019).

Variables retrieved from remote sensing need to match with a simulated state (e.g. LAI, aboveground biomass) to satisfy the data assimilation requirement (Huang et al., 2019; Jindo et al., 2023). The set of variables retrieved from the satellite has been limited to leaf area index (LAI), chlorophyll concentration, leaf water content, and soil moisture in the top layer (Hajj et al., 2017; Xie et al., 2019). Out of these four, LAI has been the most common variable to be assimilated as it can directly be retrieved from optical data (Huang et al., 2019). While few studies have assimilated soil moisture (Ines et al., 2013; Kivi et al., 2022), it was found that it improved model predictions under drought limited conditions. However, remote sensing data (optical + SAR) may provide additional information on crop growth such as the impact of abiotic stresses and other yield reducing factors. Such factors (pests and diseases) are difficult to simulate by crop models because the underlying processes may not be implemented. Moreover, data assimilation may have limited value because their impact is not necessarily being captured by estimates of LAI. Thus, there is a need to identify the satellite information that allows to further constrain process-based models and find an approach to integrate them. This would have great value towards further improvements of the yield predictions from data assimilation.

For the remote sensing information that cannot be assimilated into a crop model, data-driven approaches such as machine learning (ML) models provide a way to find quantitative relationships between features (satellite-derived indicators) and target variables (yield). ML algorithms provide unique capabilities to find non-linear relationships between the target and the features that very often occur between crop traits and yield. The ML models can learn from a large amount of data and utilize diverse related features, avoiding integration complications that are usually associated with data assimilation approaches that make use of process-based models. A shortcoming of data-driven approaches is their difficulty to interpret the predictions, and thus, these ML models lack the ability to show causal relationships between input features and predicted outputs (Gevaert, 2022). To tackle this shortcoming, there has been a growing interest in hybrid approaches, which allow the synergy between data-based and knowledge-based modeling and promote applications based on causal relationships (von Rueden et al., 2020). These hybrid approaches were successfully applied for crop yield prediction by coupling outputs from a process-based model with ML models (Feng et al., 2020; Lobell et al., 2015; Paudel et al., 2021; Shahhosseini et al., 2021). Hybrid models are a promising alternative for further development of forecasting systems as their complementarity tackles the shortcomings of each of the approaches. Therefore, these hybrid models could empower exploratory studies that point at identifying complex relationships between satellite-based estimates of crop traits and crop yield.

Given the difficulties of the satellite retrievals to satisfy the requirements to be assimilated into a crop model, this study used a hybrid system that couples a biophysical model and a data assimilation framework (Gaso et al., 2021) with a ML algorithm to evaluate the added value of satellite data to improve yield predictions. Our study addressed the following three objectives: i) to assess the added value in yield prediction accuracy at pixel level when using a ML algorithm as a predictive model with features built from the data assimilation outputs (process-based model corrected by the assimilation of LAI); ii) to investigate whether the addition of features from Sentinel-1 and 2 time-series (vegetation indices vs spectral bands and backscatter signals) to the features built from the data assimilation outputs leads to improvement in the predicted yield; iii) to identify features that contribute most in explaining the yield estimates and to explain the impact of features on yield predictions.

2. Methodology

We created a hybrid framework (Fig. 1) that uses outputs from a process-based model employing data assimilation augmented with additional satellite-based features to feed the Gradient-Boosted Decision Trees (GBTR) algorithm. Our choice of the GBTR model is based on the strengths of decision trees ensemble methods in avoiding overfitting by model averaging and the fact that Gradient boosting models are generally more accurate than bagging used by random forest (Hastie et al., 2020). The GBTR showed good performance in regional crop yield forecasting (Paudel et al., 2023, 2021). We created three scenarios (Hybrid 1,2,3) for addressing our first and second objectives. In the Hybrid-1 approach several features from the data assimilation (see Table 1) plus satellite derived red edge chlorophyll index ($CI_{red\ edge}$) time-series were used as input for the GBTR algorithm for predicting yield maps. Hybrid-1 serves as a scenario which was used to evaluate the performance of GBTR compared to data assimilation when no additional satellite-derived features (beyond the $CI_{red\ edge}$) were included. In the second (Hybrid-2) and third (Hybrid 3) scenarios we augmented the input features from the Hybrid-1 with the satellite-derived indicators and used the combined input features (model with data assimilation plus satellite indicators) to predict yields with the GBTR algorithm. This approach was used to evaluate if additional satellite-derived features beyond LAI (which was already assimilated in the process model) lead to reduced error for yield prediction. The Hybrid-2 scenario used several spectral indices to augment the process-based model outputs, while the Hybrid-3 scenario used the observations from Sentinel-2 (reflectance values) and Sentinel-1 (backscatter values) directly.

2.1. Data

2.1.1. Sites

This study used 94 fields from two soybean regions: the Corn Belt in the US and the East Pampas in Uruguay. Fig. 2 showed the spatial distribution of fields in each data set. The total area from the US was 550 ha and the total area from Uruguay was 16,551 ha. The data set from the US, hereafter referred to as Soy-US, contains 12 fields (13,750 pixels), planted in 2020. The data set from Uruguay is split into two sets: 38 fields (196,977 pixels) planted in 2020 and 44 fields (216,798 pixels) planted in 2021, hereafter referred to as Soy-UY2020 and Soy-UY2021, respectively.

Soybean varieties maturity group ranged from III to IV in Soy-US and maturity group V to VI in Soy-UY2020 and Soy-UY2021. In both regions fields were managed according to optimal agronomic practices for the region and we assumed that the influence of biotic stresses (weeds, insects, and diseases) and nutrient availability was not a limiting factor.

2.1.2. Yield monitor data processing

Yield observations from the combine harvester machines were available for each field. The combine harvester machines and the measurement units differ amongst regions, thus all the monitor data were unified into consistent data formats and measurement units. Yield data was filtered by removing outliers based on frequency distribution and outliers based on the minimum and maximum agronomic yield limits (Sun et al., 2013). The yield map was generated for each soybean field by averaging the yield points within 20 × 20 m cells. Due to privacy reasons, yield data for individual fields cannot be openly shared

2.2. Data assimilation and soybean growth model run details for features generation

We generated output variables to be used as inputs to the ML models by running the soybean crop growth model and the recalibration-based methodology presented in Gaso et al. (2021). The recalibration-based method operates by optimizing a cost function (that minimizes the difference between simulated and observed LAI), which implies that

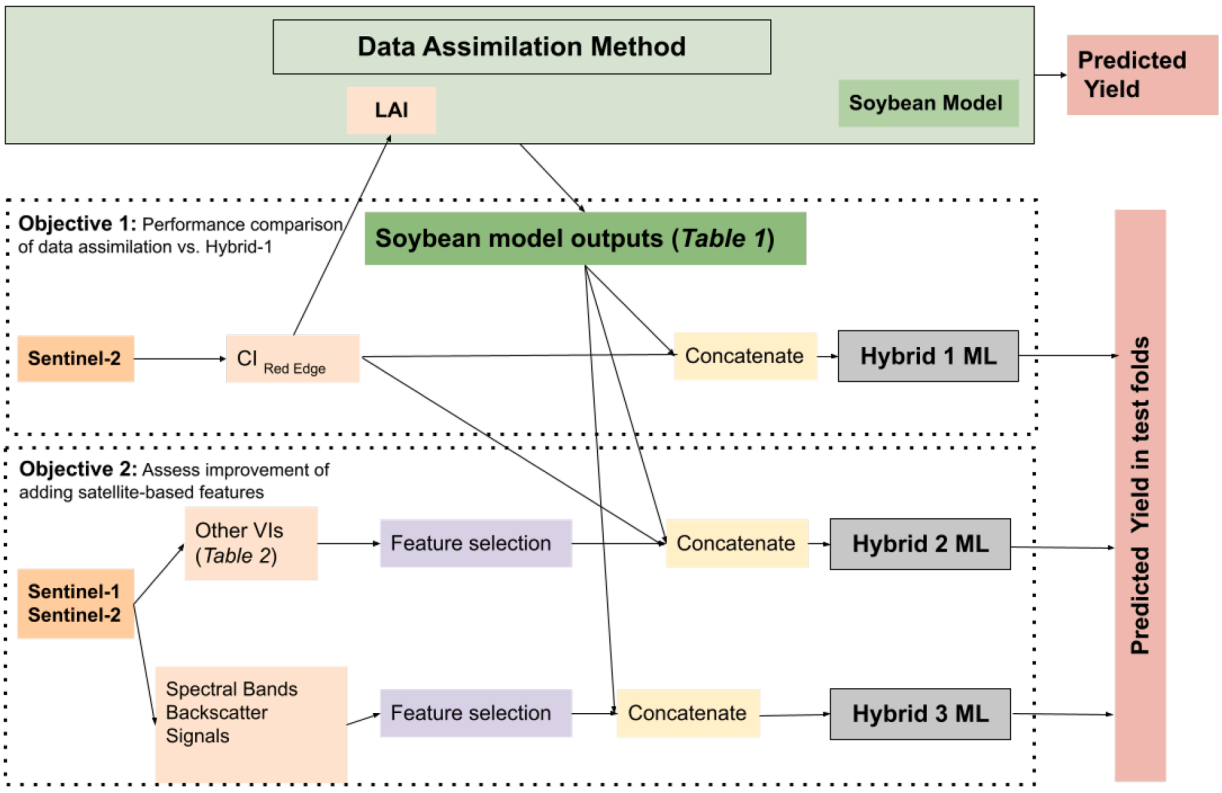


Fig. 1. Conceptual framework of the three simulation experiments (Hybrid-1, Hybrid-2 and Hybrid-3).

Table 1
Description of the crop model outputs used as features of the hybrid models.

Variable	Acronym
Cumulative water deficit in the vegetative stage	CWDv
Cumulative water deficit in the reproductive stage	CWDr
Total dry matter at flowering (R1)	TDMR1
Total dry matter at beginning of grain filling (R5)	TDMR5
Leaf area index at flowering (R1)	LAIR1
Leaf area index at beginning of grain filling (R5)	LAIR5
Cumulative crop transpiration in the vegetative stage	CTv
Cumulative crop transpiration in the reproductive stage	CTr
Total water content at flowering (R1)	TWCR1
Total water content at beginning of grain filling (R5)	TWCR5
Cumulative intercepted radiation in the vegetative stage	CRADv
Cumulative intercepted radiation in the reproductive stage	CRADr
Cumulative rainfall in the vegetative stage	CRAINv
Cumulative rainfall in the reproductive stage	CRAINr
Cumulative vapor pressure deficit in the vegetative stage	CVPDv
Cumulative vapor pressure deficit in the reproductive stage	CVPDr

uncertain model parameters must be recalibrated. Four crop model parameters were recalibrated against the LAI curves: initial LAI, soil depth, field capacity and the fraction of nitrogen translocated from leaves to seed. Weather data comes from NASA-POWER (power.larc.nasa.gov), which provides daily data (grid of 0.5° x 0.5° latitude and longitude). Planting dates and maturity group (cultivar) were not available per field, however in Soy-UY2020 and Soy-UY2021 sets, whether the field was planted as early or late (after the winter crop) was known. Thus, based on the local expertise of the region, we defined the following criteria: average planting dates and maturity group for the region, November 15th and maturity group VI, and December 15th and maturity group V for the early and late planting dates, respectively. In the Soy-US set, the planting dates were established based on the average dates for the region: May 1st and maturity group IV. By running the soybean model at the pixel level, we created a data set of features that could potentially carry the most relevant information for the yield prediction. The list of the soybean model outputs used as features of the ML models is presented in Table 1. These soybean model outputs were chosen because they synthesize the main drivers of the spatial soybean yield variability.

2.3. *Additional satellite-based features not included in the data assimilation framework*

We used known relationships between spectral indices and crop traits (Xie et al., 2019) to select a set of vegetation indices (Vis) that were used as input of the ML framework. Previous studies (Hunt et al., 2019; Perich et al., 2023; Zhang et al., 2021) have employed some of the VIs in Table 2 for crop yield prediction, but our emphasis was rather on identifying whether the addition of VIs or spectral regions improves prediction accuracy. We put particular emphasis on selecting indices that provide information from different spectral bands than those already used in the assimilation of LAI (spectral bands of Sentinel-2 in the red edge and NIR region). We used optical imagery (Level 2A) from two Sentinel-2 satellites (2A and 2B), and radar images from Sentinel-1 SAR (Synthetic Aperture Radar) dual-polarized C-band Level-1 GRD (Ground Range Detected) to build spatiotemporal satellite-based features. We used Sentinel-1 ground range detected (GRD) SAR images in dual polarization mode (VV and VH) acquired in the respective study areas from 2020 to 2021 in Google Earth Engine (GEE) platform (Gorelick et al., 2017). The Sentinel-1 SAR images were acquired in the interferometric wide swath mode (IW) with a resolution of 20 m vs 22 m in range and azimuth directions, respectively (Torres et al., 2012). The Sentinel-1 images were acquired in both ascending and descending orbits with nominal temporal resolution of 6–12 days. Prior to their ingestion into GEE, the Sentinel-1 images were processed for thermal noise removal, calibrated to sigma nought and range doppler terrain correction. We further preprocessed the images by removing remaining border noise and applying speckle filtering and radiometric terrain normalization following the methods proposed (Hoekman and Reiche,

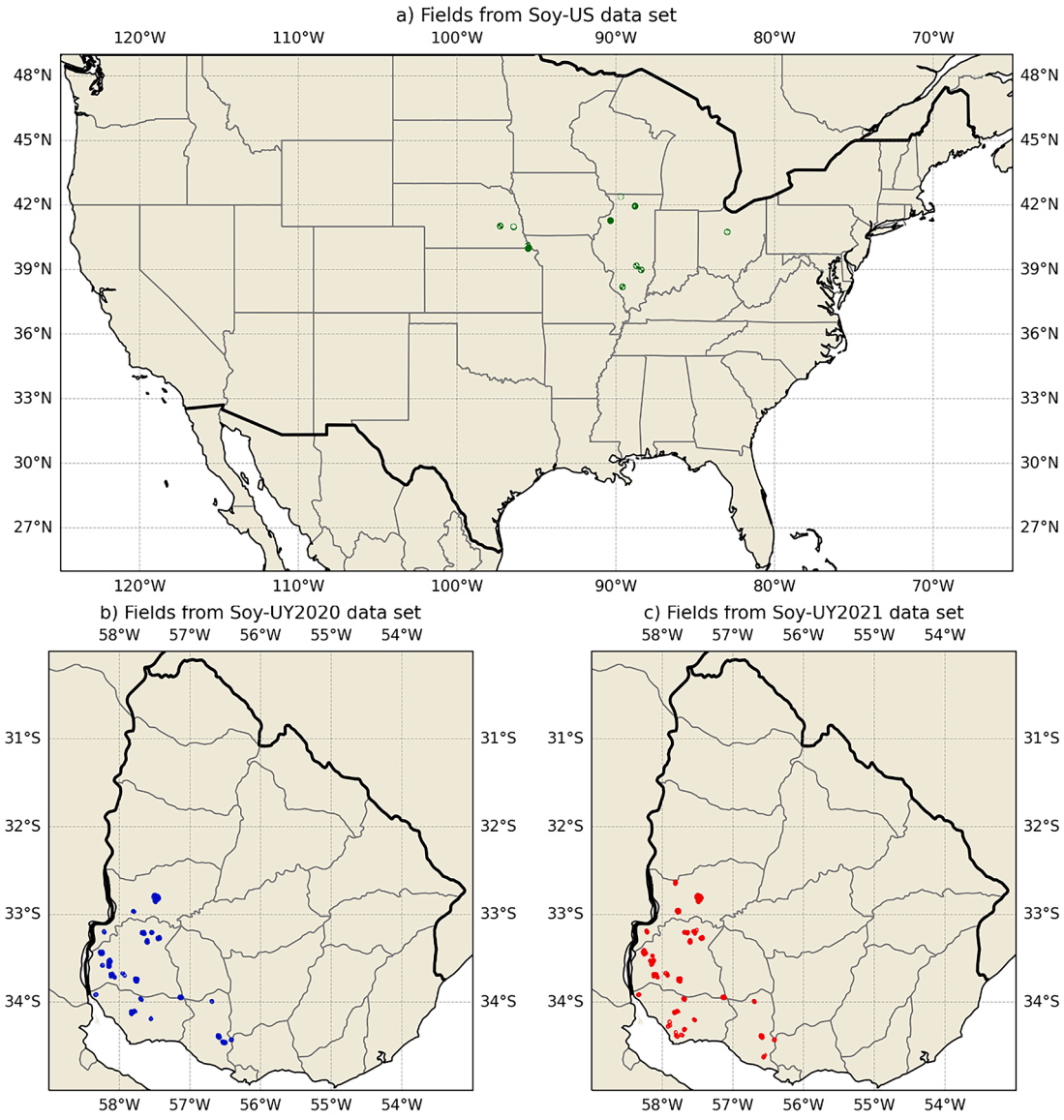


Fig. 2. Spatial distribution of the fields in each data set.

2015; Vollrath et al., 2020). The preprocessing was applied in GEE following the implementations of Mullissa et al. (2021). We made use of the public data archive available within GEE to compute and export the features variables (see: https://github.com/dgaso/Biweekly_Features_GEE)

Sentinel-2 images were ingested into GEE with atmospheric correction applied using the Sen2Corr methods (Main-Knorn et al., 2017). We further removed clouds and artifacts using the S2Cloudless approach proposed by GEE. The Sentinel-2 cloud probability image collection was employed to derive a cloud mask as a band of the Sentinel-2 surface reflectance (S2_SR) collection. We used a 10% cloud cover threshold in the cloud mask. A cloud shadow mask was computed from the dark NIR pixels (band8 in S2_SR) and it was combined with the cloud mask to produce the final cloud-shadow mask. Finally, we filtered the S2_SR by the field boundaries and the length of the growing season. The VIs from S2_SR were calculated by using the Awesome Spectral Indices package within GEE (see: https://github.com/dgaso/Biweekly_Features_GEE). Table 2 shows the VIs chosen as feature variables.

For the Sentinel-1 images, we used the Level-1 S1 GRD SAR data to compute four polarimetric descriptors (Table 2) proposed by Bhogapurapu et al. (2021). The implementation of those parameters was done within GEE. All Sentinel-1 parameters were resampled to the spatial resolution used for the Sentinel-2 features (20 m).

For each of the features, we created a temporal series of biweekly averages (eleven points) throughout the growing season. A biweekly period ensures that the crop growth curve is well-captured. We defined an equivalent length of the growing season for both study regions (the US and Uruguay). In the Soy-US region, the growing season starts on May 1st and ends on October 15. In the case of Soy-UY (Soy-UY2020 and Soy-UY2021), where the dataset was divided into early and late planting dates, we defined the growing season from November 15 to April 30 for the early planting date and from December 15 to May 30 for the late planting date. In the case of Sentinel-1, a biweekly period can have one or two images in the study area. However, in the case of Sentinel-2, the biweekly period can contain more than one image or zero. We adopted the following criteria to produce the biweekly series: i) if the biweekly has more than one image, images were averaged, ii) if the biweekly has no images, the gap was filled by linear interpolation between two consecutive (previous and subsequent) biweekly averages.

The biweekly average of the reflectance bands was computed from the following bands of the S2_SR collection: Band2 (Blue, centered at 496 nm), Band3 (Green, centered at 560 nm), Band4 (Red, centered at 665 nm), Band5 (Red Edge 1, centered at 704 nm), Band6 (Red Edge 2, centered at 740 nm), Band7 (Red Edge 3, centered at 782 nm), Band8A

Table 2
List of vegetation indexes used as features of the hybrid model.

Satellite	Description	References
Sentinel-2	Chlorophyll Index Green (CIG)	Gitelson et al. (2003)
Sentinel-2	Green Normalized Difference Vegetation Index (GNDVI)	Gitelson et al. (1996)
Sentinel-2	Normalized Multi-band Drought Index (NMDI)	Wang and Qu (2007)
Sentinel-2	Normalized Difference Vegetation Index (NDVI)	Rouse et al. (1974)
Sentinel-2	Modified Chlorophyll Absorption in Reflectance Index (MCARI)	Daughtry et al. (1999)
Sentinel-2	Transformed Chlorophyll Absorption in Reflectance Index (TCARI)	Haboudane et al. (2002)
Sentinel-2	Wide Dynamic Range Vegetation Index (WDRVI)	Gitelson (2004)
Sentinel-1	Ratio (q)	Bhogapurapu et al. (2021)
Sentinel-1	Co-pol purity (m_c)	Bhogapurapu et al. (2021)
Sentinel-1	Pseudo-scattering-type (θ_χ)	Bhogapurapu et al. (2021)
Sentinel-1	Pseudo scattering entropy (H_c)	Bhogapurapu et al. (2021)

(Red Edge 4, centered at 864 nm), Band11 (SWIR 1, centered at 1614 nm) and Band12 (SWIR 1, centered at 2202 nm). While the biweekly average of backscatter signals was computed from the Sentinel-1 SAR analysis ready (Mullissa et al., 2021) and we used the following signal: VV (Single co-polarization, vertical transmit/vertical receive), VH (Dual-band cross-polarization, vertical transmit/horizontal receive) and VVVH (ratio between VV and VH). All backscatter signals of Sentinel-1 were resampled to the spatial resolution used for the Sentinel-2 bands (20 m).

2.4. Hybrid models

2.4.1. Simulation experiments set up

Our first objective was to evaluate the advantage of using the GBTR to predict yield at the pixel level compared to the data assimilation method. To address this objective, we developed the first hybrid model (Hybrid-1), which used outputs of the soybean model after data assimilation (Table 1) and the time series of the vegetation index ($CI_{red\ edge}$) employed to estimate LAI in data assimilation (Fig. 1). Our second objective was to assess whether additional remote sensing derived features improve the accuracy of yield estimations compared to those that were obtained with assimilating LAI in a process-based model. Therefore, we developed two additional hybrid models to predict yields based on different sets of input features (Fig. 1). The second hybrid model (Hybrid-2) used the input features from Hybrid-1 plus additional VIs derived from Sentinel-1 and Sentinel-2 (Table 2). The third hybrid model (Hybrid-3) used the soybean model outputs from Hybrid-1 only plus the reflectance from all spectral bands from Sentinel-2 and all microwave signals from Sentinel-1. The satellite information from Hybrid-1 ($CI_{red\ edge}$) was considered not necessary in Hybrid-3 as this information is contained in the spectral bands. While the Hybrid-2 set up adds VIs related to crop traits different from LAI, Hybrid-3 set up directly adds reflectance information and backscattering eliminating the need for computing spectral indices.

2.4.2. Machine learning model

We chose the GBTR algorithm as the predictive model of the hybrid models. The GBTR is an ensemble method which uses gradient boosting to grow the trees (Friedman, 2001). The gradient boosting creates multiple weak models (trees) where their combination is powerful enough to find nonlinear relationships between the target and features. We used GBTR from Spark library (Spark MLlib, https://spark.apache.org/mllib/). For training and evaluation, the complete dataset was split into 5 training and test splits using GroupKFold ($k = 5$, groups = field identifiers) of the scikit-learn library (Pedregosa et al., 2011). When using a regular 5-fold cross-validation, training and test points are allocated randomly; points from the same field, with strong spatial correlation, can end up in the training and test sets. GroupKFold avoids this issue, thus helping to evaluate generalization across fields. To optimize the hyperparameters of GBTR, the training sets were further split into training and validation sets again using GroupKFold ($k = 5$, groups = field identifiers). The final predictions were obtained by refitting the GBTR model with the optimized hyperparameters on the original training set (before the second GroupKFold split).

2.4.3. Feature selection

Feature selection was performed to reduce the dimensionality of features in the case of Hybrid-2 and Hybrid-3 using the training sets. We started with the time series of VIs (121 features, 11 VIs x 11 time steps) for Hybrid-2 and spectral bands and backscattering (132 features) for Hybrid-3. The top fifteen features were selected from a GBTR model trained on each of the five training sets based on the *featureImportances* attribute. This method computes the average importance of each feature across all trees in the ensemble. The criteria for selecting the features in the internal trees is based on an impurity-based method (mean decrease in variance across all trees in the ensemble); thus, the relative importance indicates whether the feature has a relatively high depth in the decision node of the trees (models) of the GBTR. The importance vector that contains the importance value is normalized to sum to one. The final set of features included those ranked in the top fifteen by three out of five models. This process of feature selection was conducted independently per data set.

2.5. Evaluation

2.5.1. Performance comparison

The mean absolute error (MAE), root mean squared error (RMSE) and Relative RMSE (RRMSE) were computed on each testing fold, per simulation experiment (three data sets x three hybrid set up). The RRMSE was calculated using the average observed yield value in each test fold. We reported the mean and standard deviation of the accuracy metrics from five testing folds in each simulation experiment.

To address our first objective, we compared accuracy metrics from the data assimilation to the ones obtained from Hybrid-1. We then used simulations from Hybrid-1 as a baseline experiment to address our second objective (Fig. 1). We evaluated the accuracy improvement by computing the differences in RRMSE between Hybrid-1 and Hybrid-2 or Hybrid-3. The improvement from Hybrid-1 set up was expressed as the percentage of increase or decrease in RRMSE.

We computed Pearson correlation between residuals of Hybrid-1 and the other two settings (Hybrid-2 and 3). A high correlation between residuals indicates similarities amongst simulations experiments, and thus, a lack of improvement from Hybrid-1 to Hybrid-2 or Hybrid-3. The Mann-Whitney U statistical test was used to assess whether the difference between the pairs (residuals from Hybrid-1 vs Hybrid-2 and residuals from Hybrid-1 vs Hybrid-3) follows a symmetric distribution around zero. We used the Mann-Whitney U test to determine if the pairs of residuals are significantly different from each other, meaning that the increase or decrease in accuracy metrics was significant.

2.5.2. Feature importance determination with SHAP

We conducted the SHAP (Shapley Additive exPlanations) analysis to determine feature importance and to get insights into the relationship between the value of a feature and the impact on the prediction. We extracted feature importance from the SHAP method (Lundberg and Lee, 2017). Feature importance in SHAP is based on the magnitude of feature attributions and is computed as the mean absolute Shapley values. It differs from permutation feature importance which is based on the

decrease in model performance. We run the SHAP method on each fold of the test set. We computed the SHAP method 10 times to account for randomness. The final SHAP values were obtained by averaging the 10 runs. We then concatenated the SHAP values from each fold of the test set to obtain the SHAP value per feature for all the predictions (13 K points in Soy-US, 196 K points in Soy-UY2020 and 216 K points in Soy-UY2021). Further assessment of feature effects was performed to explore the impact of each feature on model output. We pointed to investigate feature directionality, meaning the relationship between the value of a feature and the impact on the prediction.

3. Results

3.1. Performance comparison between data assimilation and Hybrid-1

The Hybrid-1 improved yield predictions at pixel level compared to the data assimilation (reduced RRMSE by 16%, weighted average by dataset size). There was divergent behavior when using GBTR as predictive model. Hybrid-1 gave no improvement over data assimilation in Soy-US, while the benefit of Hybrid-1 was consistent in Soy-UY2020 (reduced RRMSE by ~21%) and Soy-UY2021 (reduced RRMSE by ~12%). There was also a shrinking of the variability around line 1:1 when comparing data assimilation with Hybrid-1, although a significant portion of the yield variability remains unexplained by both methods as indicated by numerous data points that deviate significantly from the 1:1 line (Fig. 3). Grids with the highest density of points aligned closely with the 1:1 line, indicating that the Hybrid-1 method had an added value compared to the data assimilation method (Fig. 3). This was evident in the largest data sets (Soy-UY2020 and Soy-UY2021), where the environmental variability is reduced due to the close proximity of fields (Fig. 2).

3.2. Selected satellite-based features

Derived features from Sentinel-2 (Table 2) were present within the topmost important variables for predicting soybean yield (Fig. 4a). Within those Sentinel-2 based features the most important ones were derived from the reproductive stage (biweekly six to eleven, Fig. 4c), where the critical period for soybean yield formation is located. Specifically, during biweekly period eight, which aligns with the onset of grain filling, we observed a pronounced high frequency across all features. The NMDI, an index related to plant and soil water content, was selected as important across the whole growing season. We found that the NMDI for the biweekly one and ten had high selection frequency, with values close to 1. It indicates that NMDI played an important role in capturing factors influencing crop yield early and late in the growing season. Furthermore, NMDI was the unique VIs with high presence during the vegetative stage. The MCARI, an index related to pigment content, was consistently selected with high frequency in the mid of the growing season (Fig. 4a). The VIs derived from Sentinel-1 were not within the topmost important features (Fig. 4a), except for Hc in the biweekly one.

Results from feature importance analysis run on Hybrid-3 set up (Fig. 4b) confirmed that the most relevant information in the temporal series comes from the reflectance bands of Sentinel-2 during the reproductive stage. The high frequency of the shortwave infrared bands (B11 and B12 in Fig. 4b) in biweekly one was accordingly to the frequency of the selected VIs (NMDI). The backscattering of Sentinel-1 (VH, VV and VVVH in Fig. 4b) was presented within the top fifteen in a few cases (VV and VH in Fig. 4b), which aligns with the absence of VIs formulated from Sentinel-1 in Fig. 4a.

Overall, the relative importance of the top fifteen features was generally very low (lower than 0.15, Figure S2) across the two hybrid models and three data sets. The low relative importance means that none of the features had a relatively high depth in the decision trees of the GBTR and explained a low proportion of the total variance in the target variable (yield observation).

3.3. Performance comparison between Hybrid models

There was no consistent yield prediction accuracy improvement by adding satellite-based features in Hybrid-2 and 3 across regions and datasets (Fig. 5). Overall, Hybrid-2 model reduced the RRMSE by 1% (weighted average by dataset size) compared to Hybrid-1, and Hybrid-3 model increased the RRMSE compared to Hybrid-1.

The uncertainty level was high amongst the hybrid models and data

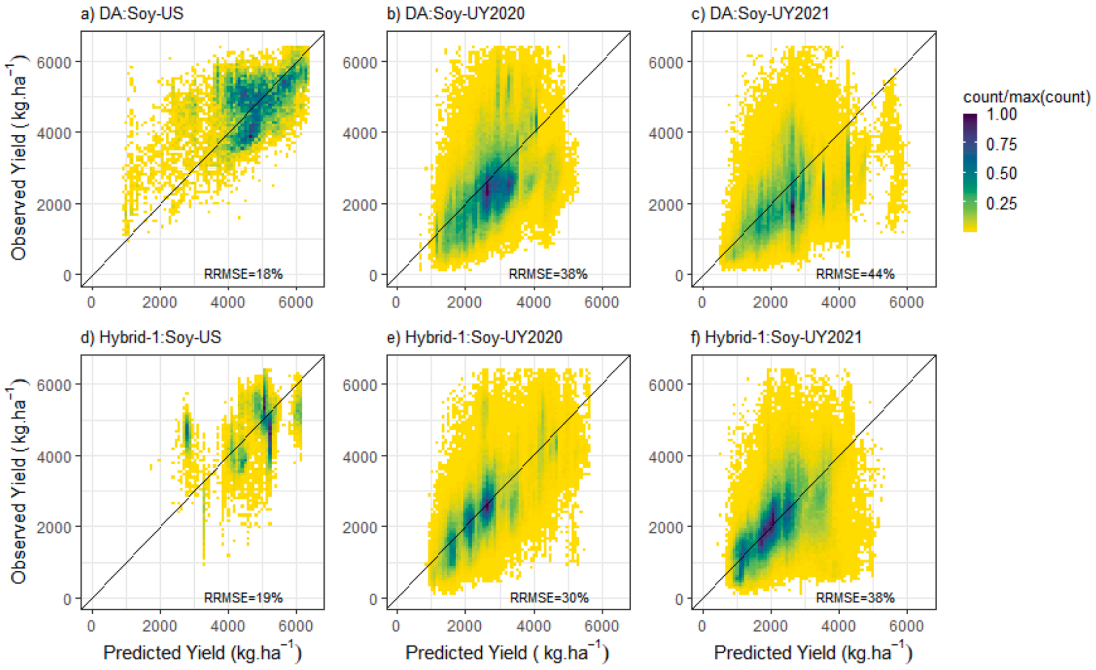


Fig. 3. Relationship between observed and predicted yield from data assimilation method (DA) and Hybrid-1, for the five test folds in each region (13 K pixels in Soy-US, 196 K pixels in Soy-UY2020 and 216 K pixels in Soy-UY2021). Color bars represent the relative number of points within the cell.

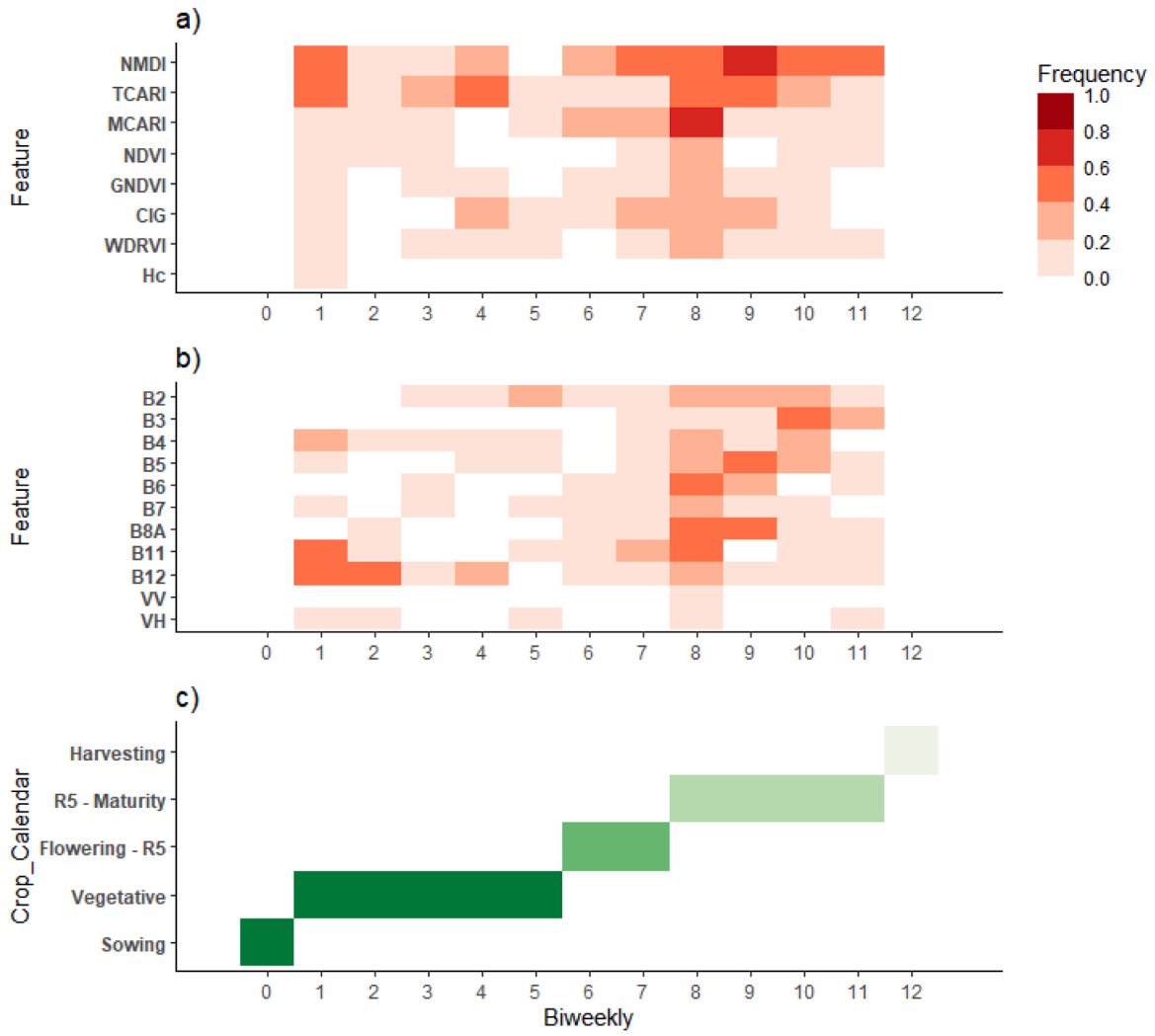


Fig. 4. Frequency of an input feature was selected during the feature importance analysis run on each fold of the training of Hybrid-2 (a) and Hybrid-3 (b). The X-axis indicates the number of the biweekly average in the growing season (temporal dimensionality). The Y-axis indicates the feature (a and b) and soybean phenology (c). R5 represents the beginning of the grain filling.

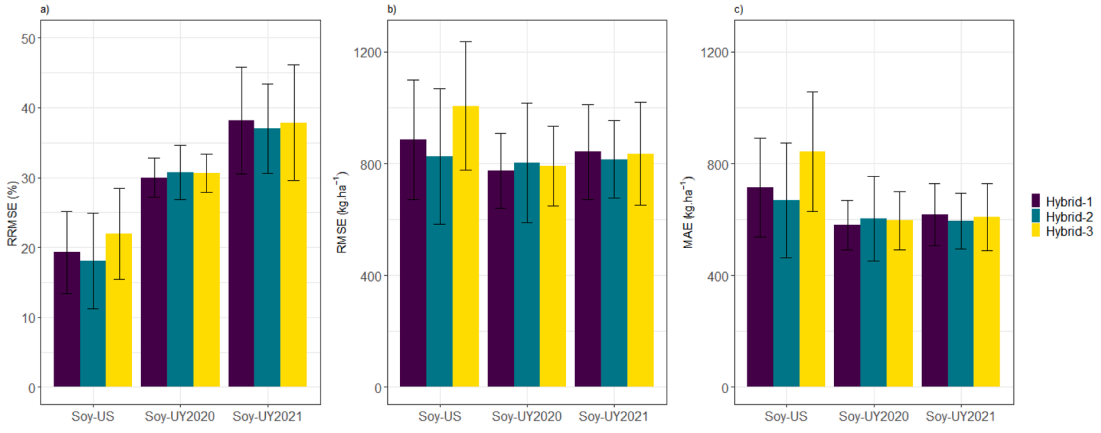


Fig. 5. Relative root mean square error (RRMSE), RMSE and mean absolute error (MAE) from Hybrid-1, Hybrid-2 and Hybrid-3, for test set in each region. Error bars represent the standard deviation across the five folds.

sets (error bars in Fig. 5). The impact of adding features in Hybrid-2 and 3 on the uncertainty level of the GBTR tended to be negative as the standard deviation increased in some cases. It means that those satellite-based features were not informative as they increased uncertainty.

Pearson correlation between residuals distribution of Hybrid-1 and the other two settings (Hybrid-2 and 3) was high in all data sets (~0.9, Figure S3). However, the p-value from the Mann-Whitney U test was less than 0.01 in all data sets and comparison between the pairs of residuals.

It indicated that the difference (improvement or reduction) in the accuracy metrics between Hybrid-1 and the experiments with added satellite information (Hybrid-2 or Hybrid-3) was significant.

3.4. Feature importance with SHAP

From the SHAP analysis we determined feature importance and feature directionality. The latter can be visualized from Fig. 6 by comparing feature values with SHAP values: whether a positive or negative effect on prediction is related to feature values. Overall, feature importance score showed that the added satellite-based features had small score value (lower than 300 kg. ha^{-1}, Figure S4). Features added in Hybrid-2 and 3 were generally important late in the season (grain filling in Fig. 6). It was also apparent that many features had a mixed impact on the model prediction (unclear features' directionality). This mixed impact of feature on predictions suggests that the relationship between the feature and the yield is non-linear. In Fig. 6, we showed the SHAP results from one of the data sets (Soy-UY2021), the other two sets can be seen in Figure S5 and S6.

Features added to Hybrid-2: TCARI and NMDI, VIs related to chlorophyl content and drought stress, were selected across the season but their importance varied during the growing season. Those indexes had less effect during the vegetative period as several points are placed quite close to 0 in Fig. 6, while their influence was higher around flowering and grain filling periods (e.g. NMDI in biweekly 7 and 8). Amongst features added in Hybrid-2, it was apparent that most of the VIs had a mixed impact on the model as NMDI in biweekly seven where high value were associated with negative and positive impact on predictions. The predominance of a mixed impact of the selected VIs on predictions was also observed in the other two data set (Figure S5 and S6).

Features added to Hybrid-3: Similar to the features added in Hybrid-2, the importance of the reflectance bands varied during the growing season. The spectral bands during grain filling period showed the largest importance, while bands in the vegetative had a small importance (points are placed close to 0). The NIR region during grain filling (B8A in biweekly 8) displayed the largest importance among the selected bands. Similar to the impact of the VIs, no clear directionality arose from Fig. 6 in the added spectral bands, where lower values in the NIR region (blue points in Fig. 6) negatively affected predictions (negative SHAP) in some cases, while it had a positive effect in other cases (positive SHAP). There was certain correspondence between the chosen spectral bands and the spectral region employed in constructing the selected VIs. For instance, spectral bands B11, B12, and B8A, present during grain filling, are in the formulation of the NMDI.

4. Discussion

To our knowledge this is the first study exploring whether the inclusion of additional satellite signals to the reflectance bands used to estimate LAI could lead to enhanced yield predictions accuracy. We built a framework that goes beyond traditional data assimilation methods. Our framework used a hybrid model that combined the data assimilation system presented by Gaso et al. (2021), satellite-based features and a ML algorithm (Fig. 1). We first evaluated the advantage of using a ML algorithm as a predictive model compared to the data assimilation only. We then compared three experiments set up (Hybrid-1, 2 and 3) to assess the benefits of adding satellite-based features. Finally, we analyzed the importance and impact of the satellite-based features across the growing season.

4.1. Performance comparison between data assimilation and Hybrid-1

We found an advantage of using the GBTR as the predictive model in Hybrid-1 as it reduced RRMSE by 16% (weighted average by dataset size) compared to data assimilation method (Fig. 4). The advantage of the GBTR lies in efficiently learning information from those features (data assimilation outputs plus the vegetation index used to estimate LAI), leading to enhanced accuracy metrics. This observation suggests the crop model corrected by data assimilation might not fully capture key underlying mechanisms, such as the effects of pests, diseases, or abiotic stresses, that significantly impact crop yield. Nevertheless, in the Soy-US set, a smaller dataset covering a vast geographic area, no enhancement was detected as Hybrid-1 deteriorated performance. This observation underscored the limitations of the data-driven approaches to be generalized under diverse environments and small dataset as the case of Soy-US set.

In this study, we employed a rigorous model evaluation that tests GBTR models using unseen randomly sampled fields as it would reflect real-life applications. By testing the GBTR under this more rigorous criteria, we ensure a robust test of the capability of this ML model to extrapolate complex patterns in space. However, splitting by space has mostly been used when using ML techniques to predict yield, such as the study by Perich et al. (2023) in which yield pixels from each individual field were used for training and testing. When this splitting approach is applied, the spatial autocorrelation increases as geographical close points are in train and test sets, resulting in overoptimistic model performance (Ploton et al., 2020). We proved the relevance of the splitting criteria by testing the performance of the GBTR when splitting by space for Soy-US and Soy-UY2020 sets (Figure S7). We found that model accuracy metrics on the test set were too optimistic when the split procedure was performed by space (reduced RRMSE by ~50%), meaning that points within the train and test set come from the same field.

It is also important to note that we observed diminished performance of the data assimilation method in comparison to our previous investigations within the same regions. For instance, the RRMSE in the Soy-US dataset was approximately twice the value reported by Gaso et al. (2023). We attributed this performance deterioration to the absence of crop management information, which lead us to standardize

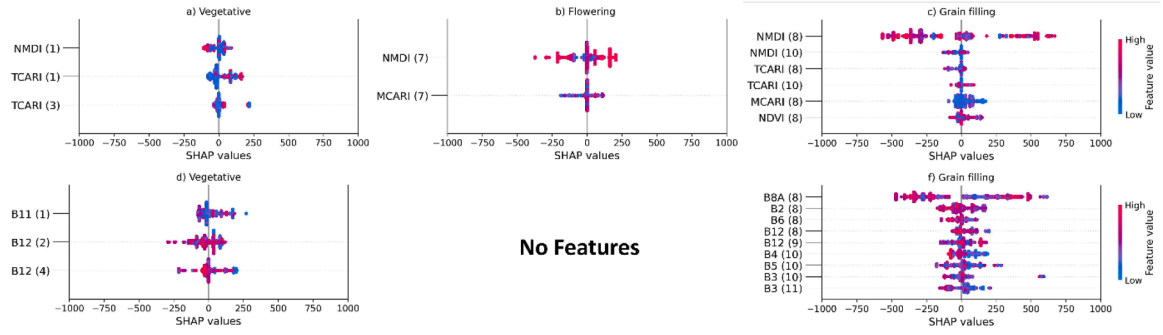


Fig. 6. Beeswarm plot of SHAP (SHapley Additive exPlanations) values for the added features in Hybrid-2 (a, b and c) and Hyrbid-3 (d, e and f), in Soy-UY2021 set, run on test sets. Features were plotted based on crop calendar: first column is vegetative, second is around flowering and third is grain filling. The number between brackets indicates the biweekly number. Each dot corresponds to a model prediction (216 K points).

the input data for the crop growth model. We defined a fixed length and standard beginning of the growing season, along with a fixed cultivar, to run the data assimilation method. This way of standardizing data assimilation inputs by making coarse assumptions for model settings could likely differ from reality, and consequently, it introduces uncertainty. When using crop models for spatially explicit soybean yield predictions, the largest source of uncertainty is explained by cultivar parameters followed by management and climate drivers (Dokoohaki et al., 2021). Thus, the lack of observed crop management is a crucial source of uncertainty in the data assimilation method. In this context, using a data-driven algorithm as a predictive model (Hybrid-1) proved advantageous for the most extensive datasets (Soy-UY2020 and 2021), characterized by lower environmental variability (Figure S1).

Previous studies (Feng et al., 2020; Shahhosseini et al., 2021) have employed hybrid approaches by integrating features from process-based model (APSIM) into ML models to predict maize and wheat yield. These studies obtained higher performance compared to our study when predicting maize yield at the aggregation of the county-level in the US (RRMSE lower than 10%, Shahhosseini et al. (2021)) and wheat yield from experiment conducted at plot-scale in Australia (RRMSE ~20% one month before harvesting, Feng et al. (2020)). They highlighted the relevance of integrating process-based features in ML models for more reliable crop yield predictions. For instance, the study from Shahhosseini et al. (2021) showed that adding water stress features to ML models significantly enhanced maize yield predictions and pointed out that ML models need more features related to the underlying mechanisms that determines crop yield. The inclusion of process-based features at the pixel-level in ML models is infrequent, as the acquisition of extensive and high-resolution yield data sets is often constrained by privacy concerns. Deines et al. (2021) demonstrated a pixel-level application, revealing scale-dependent accuracy in their ML model: it accounted for 40% of maize yield variation at the pixel level (30 m resolution in Landsat) compared to 69% at the county level. Their study stressed the relevance of quantifying accuracy at fine resolution to support crop management actions. Nevertheless, it is also crucial to highlight that generating process-based features for within-field yield predictions may be unfeasible in operational frameworks due to high computational demands. Therefore, more efficient alternatives, such as metamodels from crop growth models (Zhao et al., 2022), will be required.

4.2. Performance assessment of the satellite-based features

Including the selected satellite-based features in the hybrid models did not lead to consistent yield prediction accuracy improvement (Fig. 5). The features added to Hybrid-2 slightly reduced RRMSE (~1%), while features added to Hybrid-3 tended to be negative. It means that the GBTR struggled to extract meaningful patterns from the added features. The marginal incremental predictive power of the added features can be explained by the low score of these features added to Hybrid-2 and 3 (Figure S2 and S4). We associated the marginal predictivity added by the selected features with the capability of the $CI_{red\ edge}$ (index used to estimate LAI in the data assimilation) to be an accurate estimator of chlorophyll content (Clevers and Gitelson, 2012) and LAI (Nguy-Robertson et al., 2012). Thus, a substantial portion of the spatial variability was effectively captured by the features that were assimilated into the process-based model, and the selected indexes based on importance measures, such as TCARI and MCARI, exhibit a certain degree of redundancy in their information content.

The predictive capacity of these hybrid models is expected to improve when features are built by accounting for the observed crop calendar. The uncertainties in crop phenology due to the lack of information on soybean management lowers the predictive power of the time series of satellite-based features. Thus, this uncertainty could result in noisy predictors or irrelevant information. Estimating the starting of the growing season from satellite information might be a way to reduce this uncertainty. It is also essential to note the crop-specific (soybean) challenge, as previously reported by Dado et al. (2020), who observed markedly poorer performance in soybean compared to wheat. This crop-specific characteristic of soybean crop is probably related to the indeterminate growth habit (vegetative growth continues after flowering) compared to cereal crops. This indeterminate habit likely results in temporal features with weak signals.

4.3. Features importance and interpretability analysis

We found that features from Sentinel-1 did not contain predictive information, and thus, they were rarely included in the feature selection (Fig. 4). We explained the lack of Sentinel-1 features by the fact that the signals from the spectral bands of Sentinel-2 are stronger and more effective than the backscattering signal. The importance of the selected VIs and spectral bands was higher late in the season (Fig. 4). The VIs related to LAI, chlorophyll content and its interaction, such as TCARI and MCARI, showed higher importance from flowering time onwards (Fig. 4) but their impact on prediction was mostly unclear (Fig. 6). Similarly, the reflectance in the NIR and red edge regions (Fig. 4) were important during grain filling. The latter is probably related to variability in senescence rates across the field which leads to yield variability due to the tight link between the length and rate of the senescence (drop in chlorophyll concentration) and soybean yield formation. It reaffirmes the predictive power of features closely related to the soybean critical period (pod formation to end of grain filling, Monzon et al. (2021)). Nonetheless, it also highlights the challenge of early forecasting as the most predictive features come from the reproductive stage. Our findings agree with other studies forecasting mid-season soybean yield (Khaki et al., 2021; Schwalbert et al., 2020), where model degradation was found when removing the satellite predictors from flowering onwards, confirming the importance of these predictors.

The VIs associated with crop traits different from LAI, chlorophyll content and its interaction, could give us insight into the characteristics not accounted for by index used to estimate LAI in the data assimilation ($CI_{red\ edge}$). For instance, the NMDI was selected frequently in the beginning, middle and end of the season. This index is sensitive to drought severity, being well suited to estimate both soil and vegetation moisture (Wang and Qu, 2007). Thus, the importance of NMDI early and late in the season with predominance of bare soil could be connected with soil moisture content. In the middle of the season, with high LAI values, the importance of NMDI is linked to vegetation water content rather than soil moisture. However, it is also worth to emphasize that the contribution of the VIs to the predictive performance of the model was minimal. Therefore, additional studies are required to untangle pathways towards further improvement in within-field soybean yield predictions.

4.4. Future application of Hybrid models for yield predictions

Mixing approaches that combine data-driven and process-based techniques offer a promising alternative for future development of methods aiming at forecasting crop yield (Maestrini et al., 2022). While data-driven approaches usually fail in simulating unseen conditions, biophysical approaches, where the “transfer learning” relies on biophysical principles, are valuable tools that provide complementarity. Nevertheless, biophysical models have clear limitations and have difficulties when dealing with yield reducing factors that are difficult to model such as the impact of pest and disease. On one hand, biophysical process-based model can ensure transferability of the model while on the other hand, the application of data-driven models can be used to learn local features or yield reducing impacts that are not part of the biophysical model.

Deep learning techniques have proven to be a powerful method that allows us to extract useful information from the raw data and reach high predictive accuracy (Saleem et al., 2021). Complex models (like CNNs or

LSTM networks) have great potential for improving the prediction capabilities, although these models have limited application for understanding complex interactions amongst variables underpinning crop yield. Due to these limitations of the data-driven approaches, concerns about the need for explainable models have arisen in recent studies (Gevaert, 2022). Incorporating domain knowledge into data-driven model is one way to move towards explainable data-driven models. Constraining ML algorithms by the inclusion of domain knowledge will lead to more reliable data-driven models. Thus, these mixing approaches would contribute to achieve more reliable and understandable ML algorithms. It is also worth mentioning that merging approaches usually implies generating features from a process-based model, which comes at the cost of high computing resources. Therefore, future mixing approaches applications should focus on including domain knowledge throughout metamodels from crop growth models, which are computationally more efficient, and thus, an effective way to integrate them.

5. Conclusions

Our study built a hybrid system that goes beyond data assimilation methods to assess the added value of using a ML algorithm and the improvement when adding satellite-based features, in a large dataset from two regions (~400 K points in Uruguay and 13 K points in the US). We found an advantage when using the GBTR as the predictive model in Hybrid-1 (reduced RRMSE by ~16%, weighted average by dataset size). Thus, the ML algorithm was able to extract meaningful information and explain part of the remainder variability. However, there was no improvement by using the GBTR in the dataset that contains the largest intrinsic variability (Soy-US), which pointed to the limitation of the data-driven model to be generalized under diverse environments and a small dataset.

Adding the VIs to the features built from the data assimilation had a marginal improvement (reduced RRMSE by ~1%), while the impact of adding reflectance and backscatter values was negative. Satellite features derived from the reproductive stages proved to be the most important ones, which confirmed the relevance of capturing this crop growth period for accurate yield estimation and pointed at the challenges of early in-season forecasting. However, it is also worth to highlight that the impact of satellite-based features on model prediction was mostly unclear (unclear directionality), which might have resulted in noisy predictors. The predictive power of the satellite-based features could be enhanced by estimating field-specific emergence dates which would reduce this uncertainty. In other agricultural systems, there may be other limiting factors, particularly those not so strongly reflected in LAI, such as certain pest and disease impacts. In those systems, additional satellite features beyond LAI could contribute more significantly to unraveling yield variability, and may therefore be more successful in enhancing predictive accuracy. Thus, our findings highlighted that more studies are required to better disentangle pathways towards further improvement in constraining crop models by ingesting satellite observations. We believe that merging prediction tools by using mixed approaches is a promising tool for future development of yield prediction frameworks as the data-driven model benefits from the process-based method that rely on biophysical principles.

CRediT authorship contribution statement

Deborah V. Gaso: Conceptualization, Formal analysis, Writing – original draft. **Dilli Paudel:** Conceptualization, Formal analysis, Methodology, Writing – review & editing. **Allard de Wit:** Conceptualization, Investigation, Methodology, Supervision, Writing – review & editing. **Laila A. Puntel:** Conceptualization, Writing – review & editing. **Adugna Mullissa:** Data curation, Resources, Software. **Lammert Kooistra:** Conceptualization, Supervision, Writing – review & editing.

Declaration of competing interest

The authors have declared that there is no conflict of interests.

Data availability

The authors do not have permission to share data.

Acknowledgments

This research was funded by the Instituto Nacional de Investigación Agropecuaria de Uruguay and a Ph.D. fellowship provided by Agencia Nacional de Investigación e Innovación (ANII, scholarship code: POS_EXT_2017_1_147121). We would like to thank ProNutrition Agrotecnologías, USDA-NRCS Conservation Innovation Grant (Award Number NR213A7500013G021) and USDA NIFA-AFRI Food Security Program Coordinated Agricultural Project for sharing the field data.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.agrformet.2024.110022.

References

Bhogapurapu, N., Dey, S., Bhattacharya, A., Mandal, D., Lopez-Sanchez, J.M., McNairn, H., López-Martínez, C., Rao, Y.S., 2021. Dual-polarimetric descriptors from Sentinel-1 GRD SAR data for crop growth assessment. ISPRS J. Photogramm. Remote Sensing 178, 20–35. https://doi.org/10.1016/j.isprsjprs.2021.05.013.

Cassman, K.G., Grassini, P., 2020. A global perspective on sustainable intensification research. Nat. Sustain. 3, 262–268. https://doi.org/10.1038/s41893-020-0507-8.

Clevers, J.G.P.W., Gitelson, A.A., 2012. Using the Red-Edge Bands on Sentinel-2 For Retrieving Canopy Chlorophyll and Nitrogen Content. European Space Agency, (Special Publication) ESA SP.

Dado, W.T., Deines, J.M., Patel, R., Liang, S.Z., Lobell, D.B., 2020. High-resolution soybean yield mapping across the us midwest using subfield harvester data. Remote Sens. 12, 1–22. https://doi.org/10.3390/rs12213471.

Daughtry, C.S.T., Walthall, C.L., Kim, M.S., Brown de Colstoun, E., McMurtrey III, J.E., 1999. Estimating corn leaf chlorophyll concentration from leaf and canopy reflectance. Remote Sens. Environ.

Deines, J.M., Patel, R., Liang, S.Z., Dado, W., Lobell, D.B., 2021. A million kernels of truth: insights into scalable satellite maize yield mapping and yield gap analysis from an extensive ground dataset in the US Corn Belt. Remote Sens. Environ. 253 https://doi.org/10.1016/j.rse.2020.112174.

Dokoohaki, H., Kivi, M.S., Martinez-Feria, R., Miguez, F.E., Hoogenboom, G., 2021. A comprehensive uncertainty quantification of large-scale process-based crop modeling frameworks. Environ. Res. Lett. 16 https://doi.org/10.1088/1748-9326/ac0f26.

Feng, P., Wang, B., Liu, D.L., Waters, C., Xiao, D., Shi, L., Yu, Q., 2020. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. Agric. For. Meteorol. 107922, 285–286. https://doi.org/10.1016/j.agrformet.2020.107922.

Folberth, C., Elliott, J., Müller, C., Balkovic, J., Chryssanthacopoulos, J., Izaurralde, R.C., Jones, C.D., Khabarov, N., Liu, W., Reddy, A., Schmid, E., Skalský, R., Yang, H., Arneth, A., Ciais, P., Deryng, D., Lawrence, P.J., Olin, S., Pugh, T.A.M., Ruane, A.C., Wang, X., 2016. Uncertainties in global crop model frameworks: effects of cultivar distribution, crop management and soil handling on crop yield estimates. Biogeosci. Discuss. 1–30. https://doi.org/10.5194/bg-2016-527.

Friedman, J., 2001. Greedy function approximation : a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Gaso, D.V., de Wit, A., Berger, A.G., Kooistra, L., 2021. Predicting within-field soybean yield variability by coupling Sentinel-2 leaf area index with a crop growth model. Agric. For. Meteorol. 308–309, 108553 https://doi.org/10.1016/j.agrformet.2021.108553.

Gaso, D.V., De Wit, A., De Bruin, S., Puntel, L.A., Berger, A.G., Kooistra, L., 2023. Efficiency of assimilating leaf area index into a soybean model to assess within-field yield variability. Eur. J. Agron. 143, 126718 https://doi.org/10.1016/j.eja.2022.126718.

Gevaert, C.M., 2022. Explainable AI for earth observation: a review including societal and regulatory perspectives. Int. J. Appl. Earth Observ. Geoinform. 112, 102869 https://doi.org/10.1016/j.jag.2022.102869.

Gitelson, A.A., 2004. Wide dynamic range vegetation index for remote quantification of biophysical characteristics of vegetation. J. Plant Physiol.

Gitelson, A.A., Gritz, Y., Merzlyak, M.N., 2003. Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves. J. Plant Physiol.

Gitelson, A.A., Kaufman, Y.J., Merzlyak, M.N., Blaustein, J., 1996. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. Remote Sens. Environ.
Gorelick, N., Hancher, M., Dixon, M., Ilyushchenko, S., Thau, D., Moore, R., 2017. Google Earth Engine: planetary-scale geospatial analysis for everyone. Remote Sens. Environ. 202, 18–27. https://doi.org/10.1016/j.rse.2017.06.031.
Haboudane, D., Miller, J.R., Tremblay, N., Zarco-Tejada, P.J., Dextraze, L., 2002 Integrated narrow-band vegetation indices for prediction of crop chlorophyll content for application to precision agriculture.
Hajj, M.El, Baghdadi, N., Zribi, M., Bazzi, H., 2017. Synergic use of Sentinel-1 and Sentinel-2 images for operational soil moisture mapping at high spatial resolution over agricultural areas. Remote Sens. 9, 1–28. https://doi.org/10.3390/rs9121292.
Hastie, T., Tibshirani, R., Friedman, J., 2020. The Elements of Statistical Learning. Data Mining, Inference, and Prediction, Models for Ecological Data. Springer US. https://doi.org/10.2307/j.ctv15r5dgv.9.
Hoekman, D.H., Reiche, J., 2015. Multi-model radiometric slope correction of SAR images of complex terrain using a two-stage semi-empirical approach. Remote Sens. Environ. 156, 1–10. https://doi.org/10.1016/j.rse.2014.08.037.
Huang, J., Gómez-Dans, J.L., Huang, H., Ma, H., Wu, Q., Lewis, P.E., Liang, S., Chen, Z., Xue, J.H., Wu, Y., Zhao, F., Wang, J., Xie, X., 2019. Assimilation of remote sensing into crop growth models: current status and perspectives. Agric. For. Meteorol. 107609, 276–277. https://doi.org/10.1016/j.agrformet.2019.06.008.
Hunt, M.L., Blackburn, G.A., Carrasco, L., Redhead, J.W., Rowland, C.S., 2019. High resolution wheat yield mapping using Sentinel-2. Remote Sens. Environ. 233 https://doi.org/10.1016/j.rse.2019.111410.
Ines, A.V.M., Das, N.N., Hansen, J.W., Njoku, E.G., 2013. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. Remote Sens. Environ. 138, 149–164. https://doi.org/10.1016/j.rse.2013.07.018.
Jin, X., Kumar, L., Li, Z., Feng, H., Xu, X., Yang, G., Wang, J., 2018. A review of data assimilation of remote sensing and crop models. Eur. J. Agron. 92, 141–152. https://doi.org/10.1016/j.eja.2017.11.002.
Jindo, K., Kozan, O., de Wit, A., 2023. Data Assimilation of Remote Sensing Data into a Crop Growth Model, pp. 185–197. 10.1007/978-3-031-15258-0_8.
Khaki, S., Pham, H., Wang, L., 2021. Simultaneous corn and soybean yield prediction from remote sensing data using deep transfer learning. Sci. Rep. 11, 1–14. https://doi.org/10.1038/s41598-021-89779-z.
Kivi, M., Vergopolan, N., Dokoohaki, H., 2022. A comprehensive assessment of in situ and remote sensing soil moisture data assimilation in the APSIM model for improving. Hydrology and Earth System Sciences Discussions, pp. 1–33. https://doi.org/10.5194/hess-27-1173-2023.
Lobell, D.B., Thau, D., Seifert, C., Engle, E., Little, B., 2015. A scalable satellite-based crop yield mapper. Remote Sens. Environ. 164, 324–333. https://doi.org/10.1016/j.rse.2015.04.021.
Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. Adv. Neural Inf. Process. Syst. 2017, 4766–4775.
Maestrini, B., Mimić, G., van Oort, P.A.J., Jindo, K., Brdar, S., van Evert, F.K., Athanasiados, I., 2022. Mixing process-based and data-driven approaches in yield prediction. Eur. J. Agron. 139 https://doi.org/10.1016/j.eja.2022.126569.
Main-Knorn, M., Pflug, B., Louis, J., Debaecker, V., Müller-Wilm, U., Gascon, F., 2017. Sen2Cor for Sentinel-2. In: Conference Paper 3. https://doi.org/10.1117/12.2278218.
Monzon, J.P., Cafaro La Menza, N., Cerrudo, A., Canepa, M., Rattalino Edreira, J.I., Specht, J., Andrade, F.H., Grassini, P., 2021. Critical period for seed number determination in soybean as determined by crop growth rate, duration, and dry matter accumulation. Field Crops. Res. 261 https://doi.org/10.1016/j.fcr.2020.108016.
Mullissa, A., Vollrath, A., Odongo-Braun, C., Slagter, B., Balling, J., Gou, Y., Gorelick, N., Reiche, J., 2021. Sentinel-1 sar backscatter analysis ready data preparation in google earth engine. Remote Sens. 13, 5–11. https://doi.org/10.3390/rs13101954.
Nguy-Robertson, A., Gitelson, A., Peng, Y., Viña, A., Arkebauer, T., Rundquist, D., 2012. Green leaf area index estimation in maize and soybean: combining vegetation indices to achieve maximal sensitivity. Agron. J. 104, 1336–1347. https://doi.org/10.2134/agronj2012.0065.
Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., Athanasiadis, I.N., 2021. Machine learning for large-scale crop yield forecasting. Agric. Syst. 187, 103016 https://doi.org/10.1016/j.agsy.2020.103016.
Paudel, D., de Wit, A., Boogaard, H., Marcos, D., Osinga, S., Athanasiadis, I.N., 2023. Interpretability of deep learning models for crop yield forecasting. Comput. Electron. Agric. 206 https://doi.org/10.1016/j.compag.2023.107663.
Pedregosa, F., Michel, V., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Vanderplas, J., Cournapeau, D., Pedregosa, F., Varoquaux, G., Gramfort, A., Thirion, B., Grisel, O., Dubourg, V., Passos, A., Brucher, M., Perrot, M., Duchesnay, É., 2011. Scikit-learn: machine learning in python. J. Mach. Learn. Res.
Perich, G., Turkoglu, M.O., Graf, L.V., Wegner, J.D., Aasen, H., Walter, A., Liebisch, F., 2023. Pixel-based yield mapping and prediction from Sentinel-2 using spectral indices and neural networks. Field. Crops. Res. 292, 108824 https://doi.org/10.1016/j.fcr.2023.108824.
Ploton, P., Mortier, F., Réjou-Méchain, M., Barbier, N., Picard, N., Rossi, V., Dormann, C., Cornu, G., Viennois, G., Bayol, N., Lyapustin, A., Gourlet-Fleury, S., Pélissier, R., 2020. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. Nat. Commun. 11, 1–11. https://doi.org/10.1038/s41467-020-18321-y.
Rouse, R.W.H., Haas, J.A.W., Deering, D.W., 1974. Monitoring Vegetation Systems in the Great Plains with Erts.
Saleem, M.H., Potgieter, J., Arif, K.M., 2021. Automation in Agriculture by Machine and Deep Learning Techniques: A Review of Recent Developments, Precision Agriculture. Springer US. https://doi.org/10.1007/s11119-021-09806-x.
Schwalbert, R., Amado, T., Nieto, L., Corassa, G., Rice, C., Peralta, N., Schauberger, B., Gornott, C., Ciampitti, I., 2020. Mid-season county-level corn yield forecast for US Corn Belt integrating satellite imagery and weather variables. Crop. Sci. 60, 739–750. https://doi.org/10.1002/csc2.20053.
Shahhosseini, M., Hu, G., Huber, I., Archontoulis, S.V., 2021. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Sci. Rep. 11, 1–15. https://doi.org/10.1038/s41598-020-80820-1.
Sun, W., Whelan, B., McBratney, A.B., Minasny, B., 2013. An integrated framework for software to provide yield data cleaning and estimation of an opportunity index for site-specific crop management. Precis. Agric. 14, 376–391. https://doi.org/10.1007/s11119-012-9300-7.
Torres, R., Snoeij, P., Geudtner, D., Bibby, D., Davidson, M., Attema, E., Potin, P., Rommen, B.Ö., Floury, N., Brown, M., Traver, I.N., Deghaye, P., Duesmann, B., Rosich, B., Miranda, N., Bruno, C., L'Abbate, M., Croci, R., Pietropaolo, A., Huchler, M., Rostan, F., 2012. GMES Sentinel-1 mission. Remote Sens. Environ. 120, 9–24. https://doi.org/10.1016/j.rse.2011.05.028.
Vollrath, A., Mullissa, A., Reiche, J., 2020. Angular-based radiometric slope correction for Sentinel-1 on Google Earth Engine. Remote Sens. https://doi.org/10.3390/rs12111867.
von Rueden, L., Mayer, S., Sifa, R., Bauckhage, C., Garcke, J., 2020. Combining Machine Learning and Simulation to a Hybrid Modelling Approach: Current and Future Directions, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer International Publishing. https://doi.org/10.1007/978-3-030-44584-3_43.
Wang, L., Qu, J.J., 2007. NMDI: a normalized multi-band drought index for monitoring soil and vegetation moisture with satellite remote sensing. Geophys. Res. Lett. 34 https://doi.org/10.1029/2007GL031021.
Xie, Q., Dash, J., Huete, A., Jiang, A., Yin, G., Ding, Y., Peng, D., Hall, C.C., Brown, L., Shi, Y., Ye, H., Dong, Y., Huang, W., 2019. Retrieval of crop biophysical parameters from Sentinel-2 remote sensing imagery. Int. J. Appl. Earth Observ. Geoinformation. 80, 187–195. https://doi.org/10.1016/j.jag.2019.04.019.
Zhang, L., Zhang, Z., Luo, Y., Cao, J., Xie, R., Li, S., 2021. Integrating satellite-derived climatic and vegetation indices to predict smallholder maize yield using deep learning. Agric. For. Meteorol. 311 https://doi.org/10.1016/j.agrformet.2021.108666.
Zhao, Y., Han, S., Meng, Y., Feng, H., Li, Z., Chen, J., Song, X., Zhu, Y., Yang, G., 2022. Transfer-learning-based approach for yield prediction of winter wheat from planet data and SAFY model. Remote Sens. 14 https://doi.org/10.3390/rs14215474.