



Potential Use of Data-Driven Models to Estimate and Predict Soybean Yields at National Scale in Brazil

Leonardo A. Monteiro^{1,2,3} · Rafael M. Ramos⁴ · Rafael Battisti⁵ · Johnny R. Soares¹ · Julianne C. Oliveira⁶ · Gleyce K. D. A. Figueiredo¹ · Rubens A. C. Lamparelli⁷ · Claas Nendel^{8,9,10} · Marcos Alberto Lana¹¹

Received: 27 April 2022 / Accepted: 19 July 2022 / Published online: 7 September 2022
© Springer Nature Switzerland AG 2022

Abstract

Large-scale assessment of crop yields plays a fundamental role for agricultural planning and to achieve food security goals. In this study, we evaluated the robustness of data-driven models for estimating soybean yields at 120 days after sow (DAS) in the main producing regions in Brazil; and evaluated the reliability of the “best” data-driven model as a tool for early prediction of soybean yields for an independent year. Our methodology explicitly describes a general approach for wrapping up publicly available databases and build data-driven models (multiple linear regression—MLR; random forests—RF; and support vector machines—SVM) to predict yields at large scales using gridded data of weather and soil information. We filtered out counties with missing or suspicious yield records, resulting on a crop yield database containing 3450 records (23 years × 150 “high-quality” counties). RF and SVM had similar results for calibration and validation steps, whereas MLR showed the poorest performance. Our analysis revealed a potential use of data-driven models for predict soybean yields at large scales in Brazil with around one month before harvest (i.e. 90 DAS). Using a well-trained RF model for predicting crop yield during a specific year at 90 DAS, the RMSE ranged from 303.9 to 1055.7 kg ha⁻¹ representing a relative error (rRMSE) between 9.2 and 41.5%. Although we showed up robust data-driven models for yield prediction at large scales in Brazil, there are still a room for improving its accuracy. The inclusion of explanatory variables related to crop (e.g. growing degree-days, flowering dates), environment (e.g. remotely-sensed vegetation indices, number of dry and heat days during the cycle) and outputs from process-based crop simulation models (e.g. biomass, leaf area index and plant phenology), are potential strategies to improve model accuracy.

Keywords Large-scale analysis · Machine learning approaches · Public databases · Geospatial and temporal variability · Climatic and soil variables

✉ Leonardo A. Monteiro
monteiroleonardo6@gmail.com

¹ School of Agricultural Engineering (FEAGRI), Campinas, São Paulo, Brazil

² Department of Plant and Soil Sciences, University of Kentucky, Lexington, USA

³ Food and Agriculture Organization (FAO), Rome, Italy

⁴ University UNIEURO, Brasilia, Brazil

⁵ College of Agriculture, Federal University of Goiás, Goiânia, Brazil

⁶ Division of Environmental Systems Analysis, Department of Technology Management and Economics, Chalmers University of Technology, 412 96 Göteborg, Sweden

⁷ Center of Energy Planning (NIPE), Campinas, São Paulo, Brazil

⁸ Leibniz Centre for Agricultural Landscape Research (ZALF), Eberswalder-Str. 84, 15374 Müncheberg, Germany

⁹ Institute of Biochemistry and Biology, University of Potsdam, Am Mühlenberg 3, 14476 Potsdam, Germany

¹⁰ Global Change Research Institute, The Czech Academy of Science, Bělidla 986/4a, 603 00 Brno, Czech Republic

¹¹ Department of Crop Production Ecology, Swedish University of Agricultural Sciences, Uppsala, Sweden

Introduction

The soybean crop [*Glycine max* (L.) Merr.] plays a strategic role in the food and energy security issues, being one of the most important legume species cultivated worldwide (~ 120.5 million ha) (FAOSTAT, 2021). Brazil is the largest producer of this particular crop, with approximately 137.2 million tons of grains on 38.9 million hectares harvested during the last growing season (i.e. 2020/21) (Conab, 2021). The average soybean yield in Brazil is around 3.000 kg ha⁻¹, but due to the high technology associated to optimum management practices used at some farms, farmers can reach yields greater than 10.000 kg ha⁻¹ under commercial conditions (Battisti et al., 2018).

During the last decades, many efforts have been made for better understanding the geospatial and temporal variability of crop yields at large-scales (regional or national). A snapshot of the past and actual agronomic and climatic scenarios is essential regarding resources allocation, efficient market strategies, and socioeconomic policies towards closing gaps in agricultural production systems. Crop yield—the production (e.g. soybean grains, sugarcane stem and grassland biomass) per unit of land area (e.g. hectare)—is one of the mostly used metrics to indicate the level of agricultural development from a particular region (Lobell et al., 2009). However, its estimation at large scales is one of the major challenges that policymakers and governmental agencies have faced for draw efficient agriculture strategies (van Klompenburg et al., 2020). Uncertainties associated with uneven distribution of yield data collected from farmer's surveys, the spatial variability of soil, relief and weather even at small scales, the heterogeneity of inputs and genotypes being used to achieve those yields, and the need to account for gradual changes of the latter over time (plant breeding, technology adoption, policy changes) still pose considerable challenges (Hampf et al., 2020; van Bussel et al., 2015).

Agricultural models are powerful tools for assessing the effect of different environmental and management conditions on crop yields. The use of them has become popular in the last decades, following the pronounced advances in technology, since the access to data resources and computational processing also have substantially increased along last decades (Jones et al., 2017). Moreover, those tools are fundamental to identify opportunities for enhancing global food production, mitigate the GHG emissions, and shrink food insecurity in a sustainable way (Cassman & Grassini, 2020; Ewert et al., 2015).

Process-based crop simulation models have been developed and tested for better understanding of the relationships involved on crop growth, environmental conditions and management practices (Jones et al., 2017; Nendel

et al., 2014). They are particularly interesting in evaluating the impact of environmental conditions or management strategies on multiple target variables, and their trade-offs, simultaneously. Process-based models are often developed under experimental field conditions and require detailed information for running the simulations (e.g. weather, soil and management), which in many agricultural regions are still rarely found (Ramirez-Villegas & Challinor, 2012). Furthermore, appropriate calibration of these models still pose a major challenge (Wallach et al., 2021).

On the other hand, data-driven models (machine learning algorithms or statistical models) have been also massively used during the last years due their flexibility concerning inputs required. This group of models is often used for investigate the relationship between a target variable (e.g. crop yield) and a set of explanatory variables (e.g. crop, weather, soil, management and vegetation indices) (Kang et al., 2020; Webber et al., 2020). Although there are some limitations of data-driven models due its intrinsic characteristics (e.g. they do not allow to understand a particular crop growth process), they present some advantages compared to process-based models. For example, data-driven models are flexible regarding its inputs, i.e. they do not require a previously established set of inputs (daily weather records, detailed soil and management information) as needed by modelling platforms like DSSAT (Jones et al., 2003), APSIM (Holzworth et al., 2018) or MONICA (Nendel et al., 2011). Another advantage lies in the possibility of estimating yields with daily, monthly or even yearly weather records. Additionally, there is the possibility of including categorical variables like soil type and level of management (low, mid and high) in the set of explanatory variables. Thus, due to the lack of detailed inputs for running process-based crop models (e.g. cultivar choice, sowing dates, planting density, fertilization rates, etc.) at large geospatial scales, data-driven models have appeared and tested as a valuable alternative for yield estimation at regional and national scales (Jiang et al., 2020; Lobell & Burke, 2010; Schwalbert et al., 2020).

In Brazil, several studies have investigated aspects associated to the sustainability (Sentelhas et al., 2015), impact of climate change (da Silva et al., 2021) and impact of management practices (Nóia Júnior and Sentelhas, 2019) of soybean crop often through process-based modelling approaches at point-basis and field experiments. On the other hand, fewer studies have assessed the impact of climate change and advances in agricultural technology in soybean cropping systems (Hampf et al., 2020), as well as the effect of economic and operational costs at soybean yields (Vera-Diaz et al., 2008) using process- and regression-based models, respectively. In addition, hybrid methods have used remote sensing products

merged with agrometeorological models to estimate soybean yields at the regional-scale (De Melo et al., 2008; Silva Fuzzo et al., 2020), becoming a potential alternative for mapping yields at a fairly low cost.

The use of publicly available national databases towards large-scale assessments is not common. A potential source of long-term databases with information regarding agricultural information (including crop yield) is available at the Brazilian Institute of Geography and Statistics through the Survey of Agricultural Production website (IBGE/SIDRA, <https://sidra.ibge.gov.br/>). There, a range of crop production information (e.g. crop yield, harvested area, total production) is spatially aggregated from municipality to national-scale, and can be accessed for large periods.

In this study, we hypothesize that data-driven models are suitable tools for both early prediction and end-of-season soybean yields estimations at large scales based on publicly available agro-climatic information. In order to test our hypothesis, the performance of data-driven models feed with publicly available databases of soybean yields and agro-climatic data at county scale during 23 years (1996–2018) was investigated. We calibrated and validated data-driven models to estimate and further make predictions of soybean yields using an independent dataset. Thus, the objectives of this study were: (i) to evaluate the robustness of data-driven models for early prediction of soybean yields at 30, 60 and 90 days after sow (DAS), and further compare it with end-of-season (120 DAS) yield estimation in the main producing regions in Brazil; (ii) to investigate the suitability of the “best” data-driven model as a tool for predict the soybean yield for an independent year, based on publicly available databases of crop yield, weather and soil.

Material and Methods

Soybean Yield Database

A publicly available database containing county-scale soybean yield records (in kg ha⁻¹) was downloaded from the Brazilian Institute of Geography and Statistics (IBGE) during 23 growing seasons (1996–2018). The raw dataset had initially records from 558 counties, and it was submitted through a quality control to identify and further remove suspicious and unrealistic yield records. The following steps were applied to the raw crop yield dataset: (i) counties with at least one missing year were removed; and (ii) counties with identical yield records in consecutive years were either removed, since it is unlikely that it happens, due to year-to-year variability of meteorological conditions throughout crop cycle.

Weather Database

Monthly weather data containing records of maximum and minimum air temperature (T_{max} and T_{min}, respectively, °C), and precipitation (Prec, mm) were downloaded from the gridded weather database Worldclim (<https://www.worldclim.org>) for the period of 1995–2018. Worldclim is a monthly time-step product, downscaled from CRU-TS-4.03. The WorldClim data records are stored as GeoTiff files for the years 1960–2018, covering the whole globe at ~5-km spatial resolution. We downloaded the weather variables T_{max}, T_{min} and Prec, and cropped them spatially (counties selected) and temporally (November to February, during 23 years of analysis) using Quantum-GIS software.

Soil Database

Soil information were taken from the SoilGrids database (<https://www.isric.org/>)—a widely used soil information database for agro-ecological modelling studies. Soil characteristics available in SoilGrids were generated based on machine learning approaches developed through circa 150,000 soil profiles around the world, which around 5,000 are located in Brazil (Cooper et al., 2005). SoilGrids raster files cover the whole worlds on a 250-m spatial resolution at 6 standard depths (0–5, 5–15, 15–30, 30–60, 60–100 and 100–200 cm). However, we downloaded the top soil (i.e. 0–30 cm) products related to soil texture (i.e. sand and clay content), in order to add soil characteristics to the inputs of the multivariate models. The soil information was accounted through the first 30 cm of depth (averaged weight) and further geospatially aggregated for each of the county unit.

Crop Cycle and Preprocessing of Explanatory Data

For simplicity of our analysis, a typical soybean cycle of 120 days (sow to harvest) was considered. Previous assessments have highlighted that soybean sown between October and November is likely to reduce yield losses due water deficit in Brazil (Battisti & Sentelhas, 2015; Nóia Júnior and Sentelhas, 2019). Therefore, we synthetically simulated the soybean growing period starting on 1st of November (305 DOY) for all 23 growing seasons considered (1996–2018).

Crop, weather and soil information were used as input data for building the models. The time series of crop yield were de-trended in order to minimize the effects of different agronomic characteristics (herein considered as technological level) that are not available at county-scale for whole Brazil, such sow dates, maturity groups, water management (rainfed and irrigated cropping systems) and other factors that might be tricky to easily find at large scales. Although there are several approaches in the literature to de-trend data series, we used the method described by (Heinemann

& Sentelhas, 2011), where the yields are de-trended in relation to the last year in the time series. The last year of the time series theoretically represent the year where farmers use the most advanced technology in their fields, whereas the other years are de-trended based on that year, following the Eqs. 1, 2 and 3.

$$Y_{\text{regression}} = a + bx \quad (1)$$

$$C_{\text{residual}} = \frac{Y_{\text{obs}} - Y_{\text{predicted}}}{Y_{\text{predicted}}} \quad (2)$$

$$Y_{\text{detrended}} = (1 + C_{\text{residual}}) \times Y_{\text{obs}}^n \quad (3)$$

where “a” is the linear coefficient; b is the slope of the regression; x is the id number representing each of the years (1, 2, 3, ...); $Y_{\text{regression}}$ is the yield calculated through the linear equation (kg ha^{-1}) (Eq. 1); C_{residual} is the relative deviation between the linear (Eq. 1) and observed (Y_{obs}) yields; $Y_{\text{detrended}}$ is the yield (kg ha^{-1}) theoretically without effect of agronomic technology (i.e. driven only by environmental factors); and Y_{obs}^n is the yield with theoretically highest technology from the observed time series.

The explanatory variables presented different levels of magnitude, and therefore we standardized them through max–min procedure (ranging between 0 and 1) for further feeding the data-driven models (Shahhosseini et al., 2021). The variables were chosen according to their widely known importance for driving crop yield and photosynthesis rates (e.g. air temperature) and soil–water availability (e.g. precipitation and soil texture) (Hatfield et al., 2001; Lobell et al., 2009; Monteith, 1977).

We choose to use easily available explanatory variables for feeding the models (e.g. air temperature, precipitation and soil texture) to make the methods useful for decision-makers, farmers and other stockholders whom maybe do not have access or familiarity in manipulating large-scale databases, being well aware that richer, public available information at global geospatial scale of weather, soil and crop-related data exists.

Data-Driven Models

The data-driven models tested herein follow different approaches, but they were generically adjusted to the natural logarithm of the observed yields as a function of the explanatory variables (Lobell & Burke, 2010), as presented in the Eq. 4:

$$\log(Y_{\text{obs}}) = f(\text{weather and soil information}) \quad (4)$$

where Y_{obs} is the observed yield (kg ha^{-1}); weather and soil information are the inputs of the data-driven models

spatially aggregated at county-scale. The estimated yields through the data-driven models were back-transformed through exponential function for future investigation of the model performance through the statistical metrics.

We investigated the performance of two widely used data-driven models: random forests (RF) and support vector machines (SVM). In addition, the performance of multiple linear regression (MLR) was also investigated and assumed as our baseline.

MLR is a widely used statistical technique, typically accounting to the linear combination effects of the input variables to explain the variations at the response variable. Due to its simplicity and handling use, it has been massively use for agronomic applications since last decades (Olson & Olson, 1986).

RF is a broadly used machine learning method based on the ensemble of multiple trees for resolve classification and regression problems (Breiman, 2001). This method is based on producing multiple random trees, which theoretically will “vote” in the most popular class within a given set of characteristics. For regression purposes, in particular, the output generated through the RF algorithm is the average output from all trees built. RF was implemented in R software (R Core Team, 2020), using the package “randomForest” (Liaw & Wiener, 2002). The RF models were trained using 100 trees ($n_{\text{tree}} = 100$), since the error is almost constant beyond this number of trees (Figure S1), and the number of variables randomly sampled at each split equal to 3 ($m_{\text{try}} = 3$). The parameters “n_{tree}” and “m_{try}” are included in the *randomForest* function used.

SVM are robust and largely used machine learning algorithms for both classification and regression problems. SVM is currently applied in order to find an optimal hyperplane that maximize the distance between samples (or classes), separating groups with similar characteristics (support vectors) (Cortes & Vapnik, 1995). SVM approach was also implemented in R software, through “e1071” package (Meyer et al., 2021), where a radial basis kernel was adopted. SVM models were built considering the default parameters, since the tuning functions did not show considerable improvement for the results from previous studies (Lischeid et al., 2022).

Modelling Strategies

Two steps were used to verify the potential suitability of data-driven models to predict soybean yields in Brazil. First, we built the models with explanatory variables being temporally aggregated at 30, 60, 90 and 120 days after sow (DAS), to verify how earlier the soybean yields could be predicted, according to the statistical performance of the models. Since the “best” model was determined, we performed a “leave-one-year-out” cross-validation (LOYOCV)

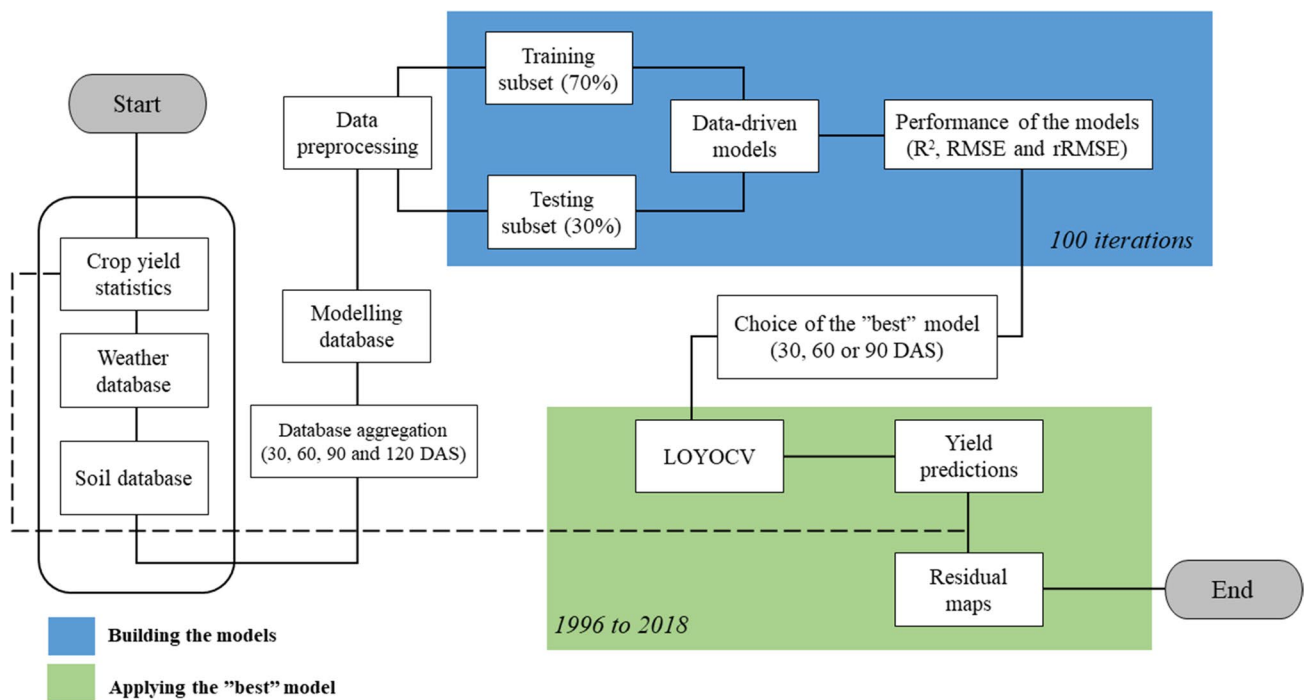


Fig. 1 Flowchart showing the main steps used to build and evaluate the performance of data-driven models for estimate soybean yields at large-scale in Brazil

strategy to predict soybean yield for each of the selected counties. Figure 1 shows the steps considered during our modelling process.

Early Prediction and End-of-Season Estimation of Soybean Yields

The robustness of the models was tested for early prediction of the soybean yields at 30, 60 and 90 days after sow (DAS). Furthermore, the end-of-cycle soybean yield (i.e. 120 DAS) was estimated and considered our “baseline” for checking how earlier the models would be accurately suitable for estimate soybean yields. The performance of the models was measured through the coefficient of determination (R^2), to account how much of the variance of captured by the model fitted; the root mean squared error (RMSE, in kg ha^{-1}), to determine the absolute error of the model, and by the mean-weighted RMSE (rRMSE, in %), in order to represent the relative error.

The data-driven models were build using a standard strategy for split the whole dataset (3450 samples) in training (2415 samples—70%) and testing (1035 samples—30%) subsets. The aforementioned selection was randomly performed 100 times, aiming to minimize potential effects of sampling selection. Since each iteration was completed, the model performance was determined for each of the

models investigated (MLR, RF and SVM) through statistical metrics.

Leave-One-Year-Out Cross-Validation Approach (LOYOCV)

Once the “best” model (i.e. the model that showed the best statistical coefficients and the number of days after sow) was chosen, the LOYOCV approach was performed. Thus, we investigated whether a given model would be suitable for estimate soybean yields according to the environmental characteristics for a specific and independent year. The residues (i.e. difference between the predicted and observed yields) will be geospatially presented at county-scale for each of the years evaluated in this study, as well as the relationship between predicted and observed yields.

Statistical Metrics for Model Evaluation

The performance of the data-driven models was evaluated through standard statistical coefficients broadly used in agro-ecological modelling studies. In our study, further than the coefficient of determination (R^2), the root mean square error (RMSE, kg ha^{-1}) and the mean-weighted root mean squared error (rRMSE, %) were calculated to determine the robustness of the models regardless the choice of the samples, through the Eqs. 5 and 6.

$$RMSE (kg \text{ ha}^{-1}) = \sqrt{\frac{\sum_{i=1}^n (Y_{est} - Y_{obsi})^2}{N}} \quad (5)$$

$$rRMSE (\%) = 100 \times \frac{RMSE}{\overline{Y_{obs}}} \quad (6)$$

where $\overline{Y_{obs}}$ is the average of observed yields.

Results

Selection of High-Quality Soybean Yield Datasets

Following the criteria described at the “Soybean Yield Database” section, a total of 150 counties remained (~27%) and composed our so-called “high-quality” soybean yield database. Thus, the data-driven models were fed with a total of 3450 records (150 counties \times 23 years), where the soybean yield represented the response variable from our models. The average of soybean yields within the

last five years ranged from less than 2000 to more than 3000 kg ha⁻¹, averaging 3063.1 kg ha⁻¹ (Fig. 1c). The geographic distribution of the soybean yields during the last 5-years of the time series evaluated in our study, as well as the yearly variability of soybean yield, and its frequency distribution are shown in Fig. 2.

These results are likely to provide insights about how diverse and challenging can be the large-scale modeling of soybean yields, given different cropping systems, genotypes (maturity groups, harvest timing, diseases and drought resistances), environmental conditions (air temperature and precipitation patterns) and agronomic practices (sow and harvest dates, plant density, row spacing, fertilization types and rates) across the country. The average technological progress of soybean is 45.7 kg ha⁻¹ yr⁻¹ (Fig. 1b), but a large diversity in levels of technology can be seen in Brazil, ranging from 10 to 105 kg ha⁻¹ year⁻¹. This highlights the different cropping systems that soybean is carry out along the last decades along the country (Figure S2).

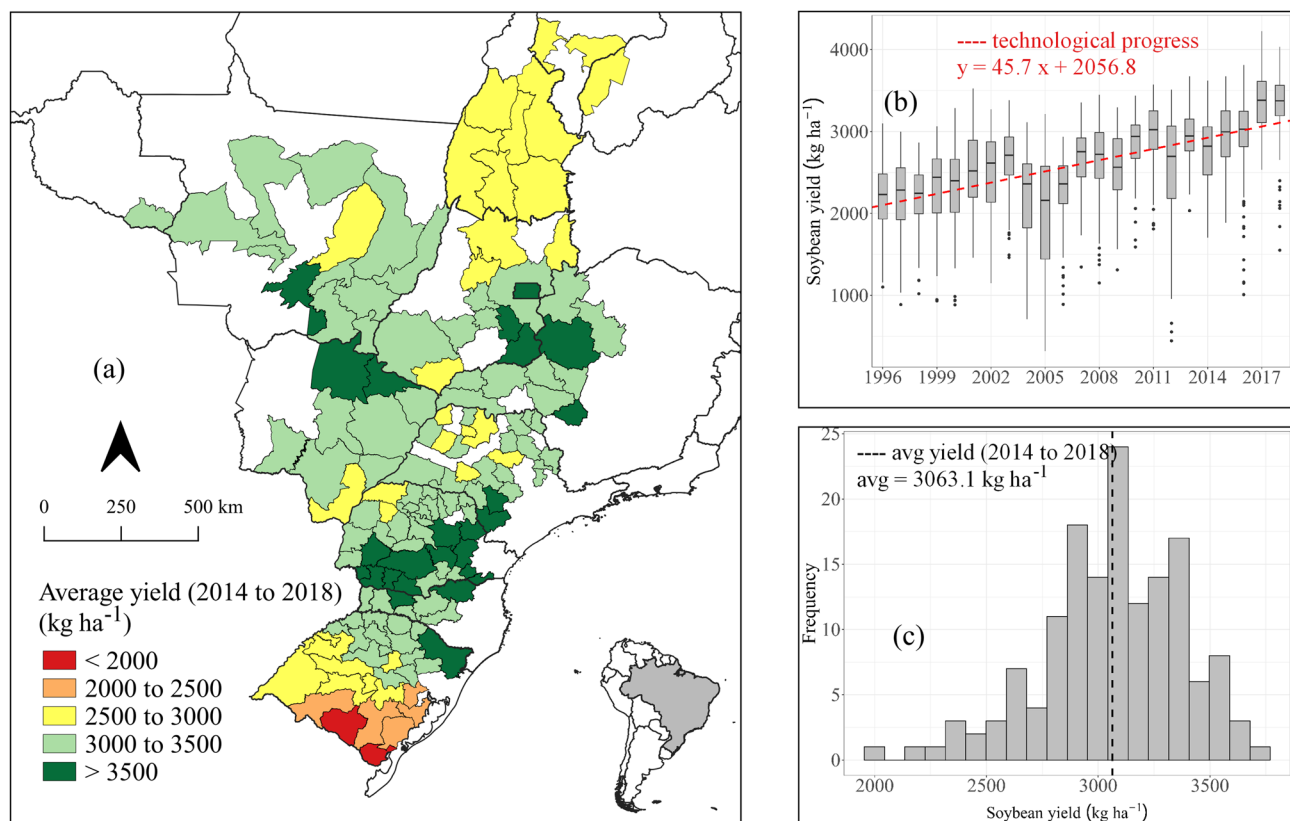


Fig. 2 Spatial variability of the average soybean yield (2014–2018) at the 150 “high-quality” counties (a); year-to-year variability of soybean yields and its technological progress (dashed line) throughout the period analysed in Brazil (1996–2018). The linear equation

address the relationship between of crop yields and years, while the slope of the trend line (x) represents the general technological progress of soybean (45.7 kg ha⁻¹ year⁻¹) (b); and the frequency analysis of the soybean average yields (2014–2018) (c)

Performance of the Data-Driven Models: Calibration and Validation Steps

The narrow distribution of the statistical metrics during the calibration step strongly suggest that our models have high robustness, regardless the choice of the samples (performed 100-folds), despite early (30, 60 or 90 DAS) and end-of-cycle (120 DAS) scenarios. Nevertheless, during the validation step the distribution curves are more scattered. In Fig. 3, the coloured histograms shows the distribution of the RMSE metric (kg ha^{-1}), given the choice of the samples for building the prediction and estimation models. Similar curves are shown in Figures S3 and S4 representing, respectively, the variability of R^2 and rRMSE.

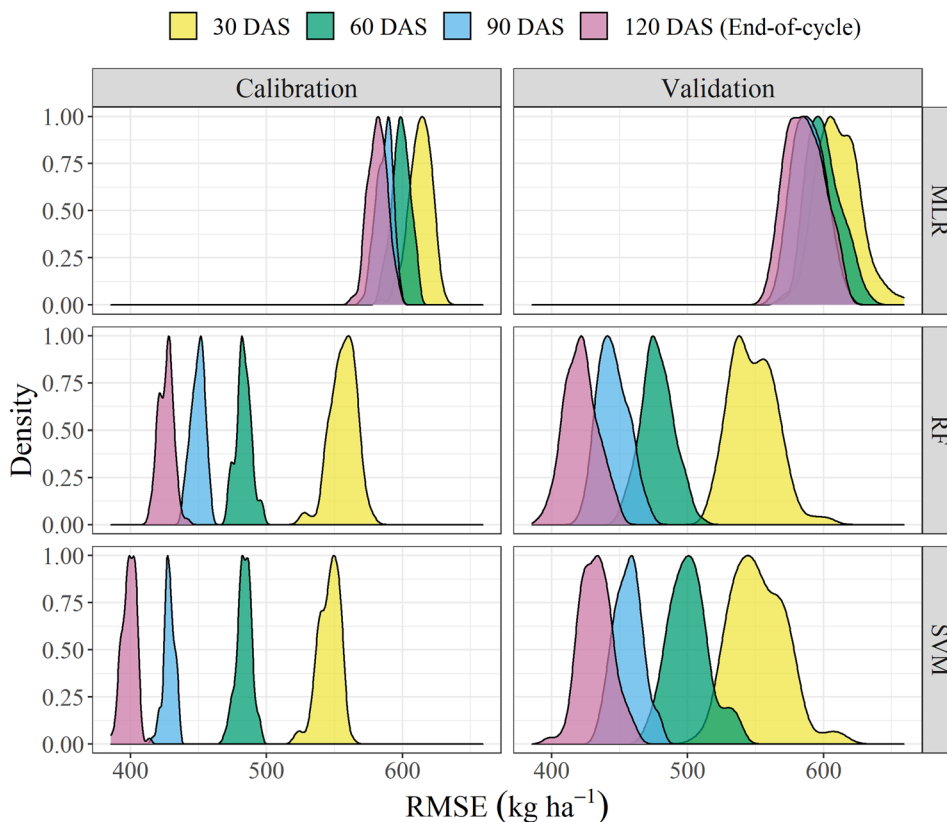
The models showed a progressive increment on their performances, since the number of days systematically increased until the whole crop cycle (120 DAS). The MLR models presented the poorest performance probably due its linear approach. MLR yielded always RMSE greater than 500 kg ha^{-1} for calibration and validation steps, representing relative deviation slightly below 18% (Figure S3). On the other hand, the MLR visually presented the largest share of the RMSE curves overlapping each other, highlighting therefore the robustness of our models, regardless the choice of the samples for building them.

In contrast, RF and SVM machine learning models presented better results than MLR, possibly due their non-linear approaches and higher capacity to better detect patterns and relationships between explanatory and response variables. Only the earliest prediction scenario (30 DAS) generated RMSE greater than 500 kg ha^{-1} during the calibration step for both RF and SVM models. The other scenarios (60, 90 and 120 DAS), however, came with RMSEs usually ranging from 400 to 500 kg ha^{-1} (12–15%, Figure S3) and only few combinations yielded RMSE smaller than 400 kg ha^{-1} (< 12%, Figure S3) (SVM model). In the validation step, similarly to the MLR models, the distribution curves of RF and SVM had slightly larger variability (more scattered) for all the statistical metrics evaluated (Fig. 3, S3 and S4). Furthermore, few differences can be identified at the distribution of RF and SVM considering the validation RMSE curves, suggesting similar performances of these methods.

Performance of the Data-Driven Models: Choosing the “Best” Model

The curves representing in our scenarios (30, 60, 90 and 120 DAS) are very similar in their shapes and position relatively to the x-axis at the validation step. Nevertheless, the 60, 90 and 120 DAS curves representing RF models overlap apparently more than those from SVM models. Therefore,

Fig. 3 Variability of the RMSE (kg ha^{-1}) for 100-fold choice of the calibration and validation subsets for building the data-driven models for predict and estimate soybean yields



we selected RF for making yield predictions using the LOY-OCV approach. Regardless the potential use of the models built with 60 DAS for predict soybean yields, there is only a tiny portion of those curves overlapping each other, suggesting a higher risk of highly skewed predicting soybean yields using models built up that early, matching with the most critical crop phases (flowering and grain filling) (Steduto et al., 2012).

Performance of the Data-Driven Models: LOYOCV Approach

At country-scale, the RF model was partially able to capture the effects of agro-climatic conditions on soybean yields, since the averages of predicted (at 90 DAS) and observed yields were nearly similar (Table 1).

The RF model tended to underestimate soybean yields, where 15 years had negative residues (Table 1). The highest

deviation is observed in 2005, where the residues achieved 650.6 kg ha^{-1} (25.6%). On the other hand, negative residues lower than -400 kg ha^{-1} ($\sim 11\%$) were not observed, indicating a potential use of data-driven models for crop yield analysis at large scales using few input data for feed the models.

Additionally, the performance of RF model to estimate soybean yields for a particular and independent year is presented at county-scale for the 23 years evaluated in our study, where is shown the geospatial and temporal distribution of the residues (Fig. 4). In general, there is a large share of white (i.e. residues between $\pm 250 \text{ kg ha}^{-1}$) or light-coloured ($\pm 500 \text{ kg ha}^{-1}$) areas throughout the years. In contrast, particular years such 2005 and 2006 come with predominantly darker-coloured (either green or brown) regions, indicating a poor performance of the model for those particular years (residues higher than 1500 kg ha^{-1}). This underperformance of the model in those years can be associated with factors that were not considered as explanatory variables, such the occurrence of extreme climate conditions during the crop cycle, resulting in poor performance of the model to capture yield variation at those particular years (Figure S5). Also, the relatively short time series used for train the model and further make predictions might have only few years with those particular conditions. For example, the accumulated precipitation in southern Brazil during 2005, 2006 and 2012 was much lower (less than 30%) than the average precipitation during the simulated soybean cycle (1996–2018, Figure S6). Additionally, the maximum air temperature presented positive deviation in large part of southern Brazil, particularly in 2005, 2006 and 2014 (Figure S7), likely affecting yields (Hatfield & Prueger, 2015). Although soybean is unlikely to be affected due low temperature in Brazil, we also investigated its geospatial and temporal variability during the period assessed in this study (Figure S8).

Discussion

Technological Progress of Soybean Yields

In this study, we aimed to investigate the performance of data-driven models for early prediction and end-of-season soybean yield estimation at large scales in Brazil. Given the continental extent of the country, naturally there are several soybean production systems, in which farmers adopt different technologies at their fields and regions, yielding different technological advances across regions. Technological progress is typically associated to the gradual change in technology and management practices adopted by farmers in a given region over time (Figueiredo, 2016).

The yield dataset was de-trended to minimize the effects of technological progress along different regions, assuming a linear gain (in $\text{kg ha}^{-1} \text{ year}^{-1}$, Figure S2) for all set

Table 1 Overview of the yearly variability of soybean yields in Brazil

Years	Soybean yield (kg ha^{-1})		Residues	
	Predicted	Observed	(kg ha^{-1})	(%)
1996	3327.8	3476.9	- 149.1	- 4.3
1997	3343.1	3420.7	- 77.5	- 2.3
1998	3371.8	3379.8	- 7.9	- 0.2
1999	3299.4	3394.8	- 95.4	- 2.8
2000	3225.2	3341.5	- 116.2	- 3.5
2001	3277.5	3624.1	- 346.6	- 9.6
2002	3087.1	3455.4	- 368.3	- 10.7
2003	3297.1	3693.9	- 396.9	- 10.7
2004	3339.8	2946.4	393.4	13.4
2005	3192.3	2541.7	650.6	25.6
2006	3048.2	2977.9	70.3	2.4
2007	3414.5	3423.6	- 9.1	- 0.3
2008	3329.4	3320.5	8.8	0.3
2009	3008.5	3114.6	- 106.0	- 3.4
2010	3367.7	3443.1	- 75.4	-2.2
2011	3304.0	3537.8	- 233.9	- 6.6
2012	3094.0	2939.9	154.1	5.2
2013	3243.4	3380.3	- 136.9	- 4.0
2014	3115.4	3080.2	35.3	1.1
2015	3360.3	3309.3	50.9	1.5
2016	3332.1	3248.3	83.8	2.6
2017	3295.9	3634.9	- 339.0	- 9.3
2018	3340.4	3548.0	- 207.5	- 5.8
Average ¹	3261.5	3314.5	- 53.0	- 1.0

Predicted yields were calculated through RF model for 90 DAS, using the LOYOCV approach, while the residues were calculated through the difference between the predicted and observed yields

¹Predicted and observed yields were submitted to the t-test, do not presenting statistic difference between them (p value = 0.3973)

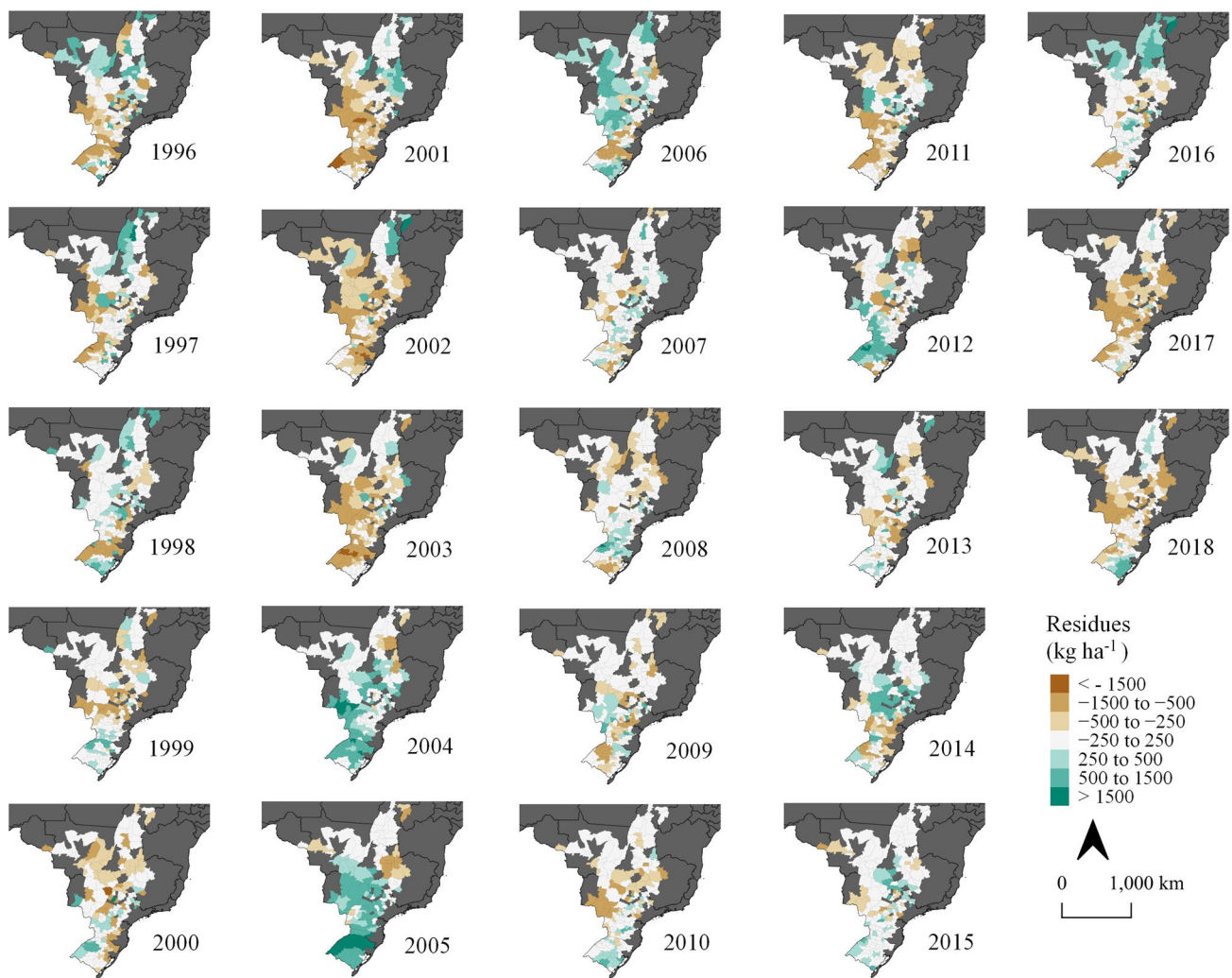


Fig. 4 Geospatial and temporal distribution of the soybean yield residues at the “high-quality” counties. Residues were calculated through the difference between estimated and observed soybean yields

of high-quality counties evaluated. Nevertheless, regions with a high level of technology probably present non-linear genetic gains along the years, while other regions where soybean is expanding, farmers are forced to adopt more suitable practices (e.g. sowing date), use new cultivars or even replace old cultivars for others more adapted to the environmental conditions (Umburanas et al., 2022). Thus, we identified large variability of technological packages in Brazil, and therefore the technological progress averaged $45.7 \text{ kg ha}^{-1} \text{ year}^{-1}$ (Fig. 1b). However, since Brazil is a country with continental dimension, there is a broader range of technological progress in the soybean producing areas, varying from 10 to $105 \text{ kg ha}^{-1} \text{ year}^{-1}$ (Figure S2). That variability is likely to be related to the advances in plant breeding and introduction of modern genotypes at the commercial fields (Rogers et al., 2015; Umburanas

et al., 2022). Also, management practices such optimized water use in soybean fields (da Silva et al., 2019), adjustment of sow dates to reduce the risk of crop failure due to water deficit on flowering and grain filling periods (Nóia Júnior & Sentelhas, 2019), and adoption of new cultivars adapted to the new agricultural frontiers such Amazon forest, might increase crop resilience under climate change scenarios (Hampf et al., 2020). Therefore, due to the broad range of factors that might affect yield gains through level of technology adopted by farmers, the yield dataset was de-trended (Figure S2). We used this, because our main goal in this study was to use the data-driven models for make short-term yield predictions. In this case, the impact of technology is unlikely to significantly affect yields as showed under long-term yield predictions, as demonstrated by Hampf et al. (2020) in Brazil.

Large-Scale Yield Simulation: The Data-Driven Models

Although several studies have applied data-driven models for scaling up crop yields at large areas (e.g. country) there is still several issues associated with the methodology used, especially regarding the models' structure and parameterisation impacting the outputs and the optimal strategy to split the dataset for training and validation. In this study, we used maybe one of most common approaches regarding split the datasets in calibration and validation subsets: 70 and 30%'s. Paudel et al., (2022) also used the 70–30% subset ratios for split the dataset and further build machine learning models for investigate yield patterns and trends at six crops in nine countries in Europe successfully. Although these authors included explanatory variables related to crop phenology (i.e. vegetative and reproductive phases), the ranges in rRMSE (10–30%) were similar to the ones we found (9.2–41.5%). Using up to 28 explanatory variables for predict corn yields at 10 states in the USA Corn-Belt region, Jiang et al. (2020) tested several data-driven approaches, where their relative errors when RF model was evaluated ranged from 7–33%. Other approaches for data splitting were investigated in Germany, where the data-driven models (RF and SVM) were feed with weather data and process-based model outputs. Considering 90% of the dataset for calibrate the model and 10% for test, they were able for capture up to 70% of the crop yield variability at national-scale (Lischheid et al., 2022). This highlights further room for including explanatory variables such remote sensing products for instance from Landsat or Sentinel constellations, and potentially improve model accuracy. Another factor that supports our results in terms of model robustness is the sampling choice. The 100-fold sampling process that we selected was likely to reduce the skewness probability and increasing the random effects of sample choice. However, this is not so clear in most of the papers using machine learning for crop yield assessments.

Large-Scale Yield Simulation: Model Performance

Recently, regional analyses have been made using machine learning methods for crop yield prediction in Brazil (dos Santos et al., 2021; Fernandes et al., 2017; Schwalbert et al., 2020) The models that we tested, regardless their simplicity, performed similarly well as compared to previous studies using well-calibrated process-based models under experimental field conditions obtained. For example, using process-based simulation models, Battisti et al. (2017) tested the performance of different models and obtained RMSE ranging from 262 to 2010 kg ha⁻¹, whereas we had the same metric ranging between 400 and 500 kg ha⁻¹, when RF and SVM were used (Fig. 3). In central Brazil, (Carauta et al., 2017)

assessed the performance of MONICA model under different field conditions, and coupled with a micro-agent simulation model (MPMAS), finding a RMSE around 480 kg ha⁻¹ for soybean yield. In contrast, when using data-driven models, Schwalbert et al. (2020) coupled remote sensing indices (i.e. NDVI and EVI) with weather data for feeding machine learning models and predict soybean yield at typical soybean region in south Brazil, found RMSE figures varying around 390 to 570 kg ha⁻¹ using RF models. These authors also identified large yield deviation in some years (e.g. 2005), highlighting the need of longer data series (i.e. where a large number of “atypical” samples are potentially found), and then the data-driven models can easily learn from this atypical condition. In “Cerrado” region, dos Santos et al. (2021) investigated the suitability of several data-driven models to estimate soybean yields in that region, finding out that RF showed the best performance. In that study, further than weather variables, they included crop phenology and outputs from soil–water balance, yielding RMSE often lower than 200 kg ha⁻¹.

Uncertainties and Potential Improvements

Our results benchmark that data-driven models are powerful tools to predict and monitor crop yields and environmental impact assessment at large-scales with public available information. However, several aspects are likely to produce different perspectives in terms of model output uncertainties, and herein we addressed some of them. For example, lack of information regarding how does the crop yield dataset was collected and harmonized by IBGE system and detailed geospatial datasets regarding agricultural management practices (fertilizer rates, sow and harvest timing, impact of insects and diseases) that play a fundamental factor for determining crop production. On the other hand, datasets have been made available to characterize water resources and irrigation practices at global scale, although uncertainties associated to input data, changes in geographic distribution and lack of temporal and spatial pattern in some regions (for example, developing areas) should be considered (Siebert et al., 2015).

Regarding the choice of the weather data, many studies have shown that the source of meteorological data, as well as how it is aggregated has a significant impact on the modelled outputs (e.g. yield) (Hoffmann et al., 2016; Van Wart et al., 2013; Zhao et al., 2015). Here, we used the weather datasets available by WorldClim, which is a product from Climate Research Unit (CRU) at monthly time-step, and our models resulted in satisfactory results, given the coarseness that the analysis were carried out. However, further analysis are likely to be performed considering daily time-step weather products available at AgERA5 database (<https://cds.climate.copernicus.eu/cdsapp#!/dataset/sis-agrometeorological-indicators?tab=overview>), where “netcdf” files are available from 1979 to near-present

covering the whole globe at 0.1° lat-long regular grid. Thus, inclusion of explanatory variables considering number of dry days, number of heat days, for example, are likely to be included in our analysis aiming to improve model accuracy.

Finally, it is very attractive the idea of coupling of remote sensing products and process-based crop simulation models—so-called hybrid models—for large-scale yield monitoring. Nowadays, cloud platforms such Google Earth Engine are fundamental for accurate large-scale assessments that rely on land monitoring, allowing rapid access and processing of remotely-sensed satellite-derived products. Recent approaches have successfully used hybrid approaches for investigate the main factors that drive the yield variability of corn in the USA merging vegetation indices and outputs from validated crop simulation models (Deines et al., 2021; Kang et al., 2020; Lobell et al., 2015). As previously mentioned, studies approaching the use of hybrid models for yield prediction in Brazil are rare. Part of it is due to the lack of observed input datasets for calibrate and validate process-based simulation models beyond the experimental fields located at the research or universities centres. Those tools have proved to be suitable for generate “pseudo” yield observations at fine scale, further than other potential explanatory variables for build data-driven models. Hence, that kind of hybrid approaches are likely to be considered in future analysis of short-term crop yield monitoring at large-scales, since factors that control plant growth, development, water and health status can also be monitored through those products at fine spatial and temporal resolutions. Furthermore, the hybrid approaches seem highly interesting, since it might provide benefits from the capacity of process-based models to systematically generate crop yields for long-term future scenarios, what we cannot have only with data-driven models.

Conclusions

The main findings of this study highlighted the potential use of data-driven models for crop yield prediction at large scales given the publicly available databases in Brazil, which few studies have had explored those datasets for similar purpose. Although RF and SVM models showed a certain robustness for predicting soybean yield (R^2 from 0.17 to 0.68, nRMSE ranging from 9.2 to 41.5%) already at an early stage (90 DAS), it was a general analysis, where publicly available datasets were considered to explain the spatial and temporal variation of soybean yield in Brazil. Therefore, there is still room for enhancing the accuracy of these models through the integration of more complex sources of data (i.e. remote sensing products). In addition, hybrid approaches—for instance, combining outputs from process-based crop models (e.g. growing degree-days, flowering dates) and environment characteristics associated to

extreme climate events, number of dry and heat days during the cycle, might be a valuable option for increase the accuracy and usefulness of data-driven models. Additionally, the inclusion of remote sensing products like vegetation indices (e.g. NDVI, EVI), which are available at cloudy platforms might be an alternative for increase the explanation power of such models we used in this study. Hence, combinations of data-driven and process-based models with real-time sensor data may become an interesting approach for enhanced crop yield monitoring and improved development of decision-making strategies at large scales in a near future.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s42106-022-00209-0>.

Acknowledgements The authors would like to thank the funding support by the Sao Paulo Research Foundation (FAPESP), through the Grant numbers: [2014/26767-9 and 2017/08970-0].

Declarations

Conflict of interest We declare that there are no conflicts of interest.

References

- Battisti, R., & Sentelhas, P. C. (2015). Drought tolerance of brazilian soybean cultivars simulated by a simple agrometeorological yield model. *Experimental Agriculture*, 51, 285–298. <https://doi.org/10.1017/S0014479714000283>
- Battisti, R., Sentelhas, P. C., & Boote, K. J. (2017). Inter-comparison of performance of soybean crop simulation models and their ensemble in southern Brazil. *Field Crops Research*, 200, 28–37. <https://doi.org/10.1016/j.fcr.2016.10.004>
- Battisti, R., Sentelhas, P. C., Pascoalino, J. A. L., Sako, H., de Sá Dantas, J. P., & Moraes, M. F. (2018). Soybean yield gap in the areas of yield contest in Brazil. *International Journal of Plant Production*, 12, 159–168. <https://doi.org/10.1007/s42106-018-0016-0>
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Carauta, M., Libera, A.A.D., Hampf, A., Chen, R.F.F., Silveira, J.M.F.J., Berger, T. (2017). On-farm trade-offs for optimal agricultural practices in Mato Grosso, Brazil. *Revista de Economia e Agronegócio*. <https://doi.org/10.25070/rea.v15i3.505>
- Cassman, K. G., & Grassini, P. (2020). A global perspective on sustainable intensification research. *Nature Sustainability*, 3, 262–268. <https://doi.org/10.1038/s41893-020-0507-8>
- Conab. (2021). Brazilian Food Supply Company.
- Cooper, M., Mendes, L. M. S., Silva, W. L. C., & Sparovek, G. (2005). A national soil profile database for brazil available to international scientists. *Soil Science Society of America Journal*, 69, 649–652. <https://doi.org/10.2136/sssaj2004.0140>
- Cortes, C., & Vapnik, V. (1995). Support-Vector Networks. *Machine Learning*, 20, 273–297. <https://doi.org/10.1111/j.1747-0285.2009.00840.x>
- da Silva, E. H. F. M., Gonçalves, A. O., Pereira, R. A., Fattori Júnior, I. M., Sobenko, L. R., & Marin, F. R. (2019). Soybean irrigation requirements and canopy-atmosphere coupling in Southern Brazil. *Agricultural Water Management*, 218, 1–7. <https://doi.org/10.1016/j.agwat.2019.03.003>

- da Silva, E. H. F. M., Silva Antolin, L. A., Zanon, A. J., Soares Andrade, A., Antunes de Souza, H., dos Santos Carvalho, K., Aparecido Vieira, N., & Marin, F. R. (2021). Impact assessment of soybean yield and water productivity in Brazil due to climate change. *European Journal of Agronomy*, *129*, 126329. <https://doi.org/10.1016/j.eja.2021.126329>
- De Melo, R. W., Fontana, D. C., Berlato, M. A., & Ducati, J. R. (2008). An agrometeorological-spectral model to estimate soybean yield, applied to southern Brazil. *International Journal of Remote Sensing*, *29*, 4013–4028. <https://doi.org/10.1080/01431160701881905>
- de N6ia J6nior, R. S., & Sentelhas, P. C. (2019). Soybean-maize succession in Brazil: Impacts of sowing dates on climate variability, yields and economic profitability. *The European Journal of Agronomy*, *103*, 140–151. <https://doi.org/10.1016/j.eja.2018.12.008>
- 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 Sensing of Environment*, *253*, 112174. <https://doi.org/10.1016/j.rse.2020.112174>
- del Vera-Diaz, M. C., Kaufmann, R. K., Nepstad, D. C., & Schlesinger, P. (2008). An interdisciplinary model of soybean yield in the Amazon Basin: The climatic, edaphic, and economic determinants. *Ecological Economics*, *65*, 420–431. <https://doi.org/10.1016/j.ecolecon.2007.07.01510.1016/j.ecolecon.2007.07.015>
- dos Santos VB, dos Santos AMF, da Silva Cabral Dextx Moraes, JR, de Oliveira Vieira IC, de Souza Rolim G (2021). Machine learning algorithms for soybean yield forecasting in the Brazilian Cerrado. *Journal of the Science of Food and Agriculture*<https://doi.org/10.1002/jsfa.11713>
- Ewert, F., R6tter, R. P., Bindi, M., Webber, H., Trnka, M., Kersebaum, K. C., Olesen, J. E., van Ittersum, M. K., Janssen, S., Rivington, M., Semenov, M. A., Wallach, D., Porter, J. R., Stewart, D., Verhagen, J., Gaiser, T., Palosuo, T., Tao, F., Nendel, C., ... Asseng, S. (2015). Crop modelling for integrated assessment of risk to food production from climate change. *Environmental Modelling and Software*, *72*, 287–303. <https://doi.org/10.1016/j.envsoft.2014.12.003>
- FAOSTAT. (2021). Food and Agriculture Organization - FAOSTAT.
- Fernandes, J. L., Ebecken, N. F. F., & Esquerdo, J. C. D. M. (2017). Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. *International Journal of Remote Sensing*, *38*, 4631–4644. <https://doi.org/10.1080/01431161.2017.1325531>
- Figueiredo, P. N. (2016). New challenges for public research organisations in agricultural innovation in developing economies: Evidence from Embrapa in Brazil's soybean industry. *The Quarterly Review of Economics and Finance*, *62*, 21–32. <https://doi.org/10.1016/j.qref.2016.07.011>
- Hampf, A. C., Stella, T., Berg-Mohnicke, M., Kawohl, T., Kilian, M., & Nendel, C. (2020). Future yields of double-cropping systems in the Southern Amazon, Brazil, under climate change and technological development. *Agricultural Systems*. <https://doi.org/10.1016/j.agsy.2019.102707>
- Hatfield, J. L., & Prueger, J. H. (2015). Temperature extremes: Effect on plant growth and development. *Weather Clim. Extrem.*, *10*, 4–10. <https://doi.org/10.1016/j.wace.2015.08.001>
- Hatfield, J. L., Sauer, T. J., & Prueger, J. H. (2001). Managing soils to achieve greater water use efficiency: A review. *Agronomy Journal*, *93*, 271–280. <https://doi.org/10.2134/agronj2001.932271x>
- Heinemann, A. B., & Sentelhas, P. C. (2011). Environmental group identification for upland rice production in central Brazil. *Science in Agriculture*, *68*, 540–547. <https://doi.org/10.1590/s0103-90162011000500005>
- Hoffmann, H., Zhao, G., Asseng, S., Bindi, M., Biernath, C., Constantin, J., Coucheney, E., Dechow, R., Doro, L., Eckersten, H., Gaiser, T., Grosz, B., Heinlein, F., Kassie, B. T., Kersebaum, K. C., Klein, C., Kuhnert, M., Lewan, E., Moriondo, M., ... Ewert, F. (2016). Impact of spatial soil and climate input data aggregation on regional Yield Simulations. *PLoS One*, *11*, 1–23. <https://doi.org/10.1371/journal.pone.0151782>
- Holzworth, D., Huth, N. I., Fainges, J., Brown, H., Zurcher, E., Cichota, R., Verrall, S., Herrmann, N. I., Zheng, B., & Snow, V. (2018). APSIM Next Generation: Overcoming challenges in modernising a farming systems model. *Environmental Modelling and Software*, *103*, 43–51. <https://doi.org/10.1016/j.envsoft.2018.02.002>
- Jiang, Z., Liu, C., Ganapathysubramanian, B., Hayes, D. J., & Sarkar, S. (2020). Predicting county-scale maize yields with publicly available data. *Science and Reports*, *10*, 1–12. <https://doi.org/10.1038/s41598-020-71898-8>
- Jones, J. W., Antle, J. M., Basso, B., Boote, K. J., Conant, R. T., Foster, I., Godfray, H. C. J., Herrero, M., Howitt, R. E., Janssen, S., Keating, B. A., Munoz-Carpena, R., Porter, C. H., Rosenzweig, C., & Wheeler, T. R. (2017). Brief history of agricultural systems modeling. *Agricultural Systems*, *155*, 240–254. <https://doi.org/10.1016/j.agsy.2016.05.014>
- Jones, J. W., Hoogenboom, G., Porter, C. H., Boote, K. J., Batchelor, W. D., Hunt, L. A., Wilkens, P. W., Singh, U., Gijsman, A. J., & Ritchie, J. T. (2003). The DSSAT cropping system model. *European Journal of Agronomy*. [https://doi.org/10.1016/S1161-0301\(02\)00107-7](https://doi.org/10.1016/S1161-0301(02)00107-7)
- Kang, Y., Ozdogan, M., Zhu, X., Ye, Z., Hain, C., & Anderson, M. (2020). Comparative assessment of environmental variables and machine learning algorithms for maize yield prediction in the US Midwest. *Environmental Research Letters*. <https://doi.org/10.1088/1748-9326/ab7df9>
- Liaw, A., & Wiener, M. (2002). Classification and Regression by random. *Forest R News*, *2*, 18–22.
- Lischeid, G., Webber, H., Sommer, M., Nendel, C., & Ewert, F. (2022). Machine learning in crop yield modelling: A powerful tool, but no surrogate for science. *Agricultural and Forest Meteorology*, *312*, 108698. <https://doi.org/10.1016/j.agrformet.2021.108698>
- Lobell, D. B., & Burke, M. B. (2010). On the use of statistical models to predict crop yield responses to climate change. *Agricultural and Forest Meteorology*, *150*, 1443–1452. <https://doi.org/10.1016/j.agrformet.2010.07.008>
- Lobell, D. B., Cassman, K. G., & Field, C. B. (2009). Crop yield gaps: Their importance, magnitudes, and causes. *Annual Review of Environment and Resources*, *34*, 179–204. <https://doi.org/10.1146/annurev.enviro.041008.093740>
- Lobell, D. B., Thau, D., Seifert, C., Engle, E., & Little, B. (2015). A scalable satellite-based crop yield mapper. *Remote Sensing of Environment*, *164*, 324–333. <https://doi.org/10.1016/j.rse.2015.04.021>
- Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.-C., & Lin, C.-C. (2021). Package ‘e1071.’
- Monteith, J. L. (1977). Climate and the efficiency of crop production in Britain. *Philosophical Transactions of the Royal Society B*, *281*, 277–294. <https://doi.org/10.1098/rstb.1977.0140>
- Nendel, C., Berg, M., Kersebaum, K. C., Mirschel, W., Specka, X., Wegehenkel, M., Wenkel, K. O., & Wieland, R. (2011). The MONICA model: Testing predictability for crop growth, soil moisture and nitrogen dynamics. *Ecological Modelling*, *222*, 1614–1625. <https://doi.org/10.1016/j.ecolmodel.2011.02.018>
- Nendel, C., Kersebaum, K. C., Mirschel, W., & Wenkel, K. O. (2014). Testing farm management options as climate change adaptation strategies using the MONICA model. *European Journal of Agronomy*, *52*, 47–56. <https://doi.org/10.1016/j.eja.2012.09.005>
- Olson, K. R., & Olson, G. W. (1986). Use of multiple regression analysis to estimate average corn yields using selected soils and climatic data. *Agricultural Systems*, *20*, 105–120. [https://doi.org/10.1016/0308-521X\(86\)90062-4](https://doi.org/10.1016/0308-521X(86)90062-4)

- Paudel, D., Boogaard, H., de Wit, A., van der Velde, M., Claverie, M., Nisini, L., Janssen, S., Osinga, S., & Athanasiadis, I. N. (2022). Machine learning for regional crop yield forecasting in Europe. *Field Crops Research*, 276, 108377. <https://doi.org/10.1016/j.fcr.2021.108377>
- R Core Team. (2020). A language and environment for statistical computing.
- Ramirez-Villegas, J., & Challinor, A. (2012). Assessing relevant climate data for agricultural applications. *Agricultural and Forest Meteorology*, 161, 26–45. <https://doi.org/10.1016/j.agrformet.2012.03.015>
- Rogers, J., Chen, P., Shi, A., Zhang, B., Scaboo, A., Smith, S. F., & Zeng, A. (2015). Agronomic performance and genetic progress of selected historical soybean varieties in the southern USA. *Plant Breeding*, 134, 85–93. <https://doi.org/10.1111/pbr.12222>
- Schwalbert, R. A., Amado, T., Corassa, G., Pott, L. P., Prasad, P. V. V., & Ciampitti, I. A. (2020). Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. *Agricultural and Forest Meteorology*, 284, 107886. <https://doi.org/10.1016/j.agrformet.2019.107886>
- Sentelhas, P. C., Battisti, R., Câmara, G. M. S., Farias, J. R. B., Hampf, A. C., & Nendel, C. (2015). The soybean yield gap in Brazil—Magnitude, causes and possible solutions for sustainable production. *Journal of Agricultural Science*, 153, 1394–1411. <https://doi.org/10.1017/S0021859615000313>
- 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. *Science and Reports*, 11, 1–15. <https://doi.org/10.1038/s41598-020-80820-1>
- Siebert, S., Kumm, M., Porkka, M., Döll, P., Ramankutty, N., & Scanlon, B. R. (2015). A global data set of the extent of irrigated land from 1900 to 2005. *Hydrology and Earth System Sciences*, 19, 1521–1545. <https://doi.org/10.5194/hess-19-1521-2015>
- Silva Fuzzo, D. F., Carlson, T. N., Kourgiyalas, N. N., & Petropoulos, G. P. (2020). Coupling remote sensing with a water balance model for soybean yield predictions over large areas. *The Earth Science Informatics*, 13, 345–359. <https://doi.org/10.1007/s12145-019-00424-w>
- Steduto, P., Hsiao, T.C., Fereres, E., & Raes, D. (2012). Crop yield response to water.
- Umburanas, R. C., Kawakami, J., Ainsworth, E. A., Favarin, J. L., Anderle, L. Z., Dourado-Neto, D., & Reichardt, K. (2022). Changes in soybean cultivars released over the past 50 years in southern Brazil. *Science and Reports*, 12, 1–14. <https://doi.org/10.1038/s41598-021-04043-8>
- van Bussel, L. G. J., Grassini, P., Van Wart, J., Wolf, J., Claessens, L., Yang, H., Boogaard, H., de Groot, H., Saito, K., Cassman, K. G., & van Ittersum, M. K. (2015). From field to atlas: Upscaling of location-specific yield gap estimates. *Field Crops Research*, 177, 98–108. <https://doi.org/10.1016/j.fcr.2015.03.005>
- van Klompenburg, T., Kassahun, A., & Catal, C. (2020). Crop yield prediction using machine learning: A systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709. <https://doi.org/10.1016/j.compag.2020.105709>
- Van Wart, J., Grassini, P., & Cassman, K. G. (2013). Impact of derived global weather data on simulated crop yields. *Global Change Biology*, 19, 3822–3834. <https://doi.org/10.1111/gcb.12302>
- Wallach, D., Palosuo, T., Thorburn, P., Hochman, Z., Gourdain, E., Andrianasolo, F., Asseng, S., Basso, B., Buis, S., Crout, N., Dibari, C., Dumont, B., Ferrise, R., Gaiser, T., Garcia, C., Gayler, S., Ghahramani, A., Hiremath, S., Hoek, S., ... Seidel, S. J. (2021). The chaos in calibrating crop models: Lessons learned from a multi-model calibration exercise. *Environmental Modelling and Software*. <https://doi.org/10.1016/j.envsoft.2021.105206>
- Webber, H., Lischeid, G., Sommer, M., Finger, R., Nendel, C., Gaiser, T., & Ewert, F. (2020). No perfect storm for crop yield failure in Germany. *Environmental Research Letters*. <https://doi.org/10.1088/1748-9326/aba2a4>
- Zhao, G., Hoffmann, H., Van Bussel, L. G. J., Enders, A., Specka, X., Sosa, C., Yeluripati, J., Tao, F., Constantin, J., Raynal, H., Teixeira, E., Grosz, B., Doro, L., Zhao, Z., Nendel, C., Kiese, R., Eckersten, H., Haas, E., Vanuytrecht, E., Ewert, F. (2015). Effect of weather data aggregation on regional crop simulation for different crops, production conditions, and response variables. *Climate Research*, 65, 141–157. <https://doi.org/10.3354/cr01301>

Springer Nature or its licensor holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.