University of São Paulo
"Luiz de Queiroz" College of Agriculture

A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling

Daniel Alves da Veiga Grubert

Thesis presented to obtain the degree of Doctor in Science. Area: Agricultural Systems Engineering

Piracicaba
2023

Daniel Alves da Veiga Grubert
Bachelor in Agronomy

A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling

versão revisada de acordo com a Resolução CoPGr 6018 de 2011

Advisor:
Prof. Dr. **FELIPE GUSTAVO PILAU**

Thesis presented to obtain the degree of Doctor in Science. Area: Agricultural Systems Engineering

Piracicaba
2023

Dados Internacionais de Catalogação na Publicação
DIVISÃO DE BIBLIOTECA – DIBD/ESALQ/USP

Grubert, Daniel Alves da Veiga

A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling / Daniel Alves da Veiga Grubert. - - versão revisada de acordo com a Resolução CoPGr 6018 de 2011. - - Piracicaba, 2023.

116 p.

Tese (Doutorado) - - USP / Escola Superior de Agricultura “Luiz de Queiroz”.

1. Modelagem híbrida 2. Predição de produtividade agrícola 3. Modelagem preditiva. 4. APSIM-Sugar 5. Simulação de cana-de-açúcar 6. Algoritmos de aprendizagem de máquina I. Título

ACKNOWLEDGMENTS

To God, for everything in my life.

To my beloved Bruna Orsi, for being by my side for so long. Thank you is not enough for all the support you have given me. We have gone through many difficulties, joys, and achievements, from graduation to the dream of becoming doctors at ESALQ/USP. If I had to start again, I wouldn't change anything as long as it was with you.

To my mother Virta Maria and my brother Willian Grubert, for the example you have set for me and for supporting my academic and professional development.

To my advisor Prof. Dr. Felipe Gustavo Pilau, for the trust placed in me, the opportunities provided, the teachings and guidance during my master's and doctoral studies, and above all, for the great friendship.

To CAPES for granting me a scholarship.

To the professors of the PPG Agricultural Systems Engineering for the teachings.

SUMMARY

RESUMO

Uma abordagem sinérgica para a previsão de produtividade da cana-de-açúcar usando aprendizado de máquina, sensoriamento remoto e modelagem baseada em processos

Previsões precisas e acuradas da produtividade de culturas agrícolas são fundamentais para agricultores e tomadores de decisão. Este estudo tem como objetivo avaliar uma abordagem híbrida que envolve dados de sensoriamento remoto, modelagem de culturas com modelos baseados em processos e algoritmos de aprendizado de máquina para melhorar as previsões de produtividade da cana-de-açúcar. Para isso, foi desenvolvido uma abordagem híbrida de previsão de produtividade que combina várias fontes de dados, incluindo variáveis simuladas de solo e da planta do modelo APSIM (um modelo de cultura baseado em processos), dados meteorológicos e índices de vegetação. Esses dados foram utilizados como entrada em modelos de aprendizado de máquina para prever a produtividade final da cana-de-açúcar. Neste estudo, foram avaliados 16 modelos de regressão para prever a produtividade da cana-de-açúcar no final da safra ao nível municipal, no estado de São Paulo, Brasil, durante o período 2010-2020. Os resultados indicaram que a abordagem híbrida desenvolvida utilizando o algoritmo K-Neighbors Regressor apresentou a melhor performance estatística, resultando no menor erro absoluto médio (MAE) de 3.26 t ha^{-1}, com um erro percentual absoluto médio (MAPE) de 4.54%. As previsões de produtividade da cana-de-açúcar proporcionaram maior grau de precisão entre 1-2 meses antes da colheita. Além disso, determinou-se quais variáveis exerceram maior influência para a previsão da produtividade da cana-de-açúcar, excluindo parcialmente algumas variáveis do modelo de previsão. Os resultados mostraram que a adição de variáveis simuladas pelo modelo baseado em processos (APSIM) como variáveis de entrada para modelos de aprendizado de máquina, pode reduzir o erro quadrático médio (RMSE) da previsão de produtividade, variando entre 7,7 e 26,9%, enquanto que, os índices de vegetação tiveram o menor impacto nas previsões. A análise mostrou que os dados meteorológicos têm um impacto maior na previsão da produtividade quando fornecidos ao modelo baseado em processos do que quando usados diretamente em algoritmos de aprendizado de máquina. Esse resultado indica que as variáveis simuladas fornecidas pelo APSIM oferecem uma descrição biofísica mais completa da interação entre solo, planta e atmosfera.

Palavras-chave: Modelagem híbrida, Predição de produtividade agrícola, Modelagem preditiva, APSIM-Sugar, Simulação de cana-de-açúcar, Algoritmos de aprendizado de máquina

ABSTRACT

A synergistic approach to sugarcane yield forecasting using machine learning, remote sensing, and process-based modeling

Accurate and precise crop yield forecasts are essential for farmers and decision-makers. This study aims to assess a hybrid approach involving remote sensing data, crop modeling with process-based models, and machine learning algorithms to improve sugarcane yield predictions. To achieve this, a hybrid yield forecasting approach was developed, combining various data sources, including simulated soil and plant variables from the APSIM model (a process-based crop model), meteorological data, and vegetation indices. These data were used as inputs in machine learning models to forecast end-season sugarcane yield. In this study, 16 regression models were evaluated to forecast sugarcane yield at the municipal level in the state of São Paulo, Brazil, during the period 2010-2020. The results indicated that the hybrid approach developed using the K-Neighbors Regressor algorithm showed the best statistical performance, resulting in the lowest Mean Absolute Error (MAE) of 3.26 t ha^{-1}, with a Mean Absolute Percentage Error (MAPE) of 4.54%. Sugarcane yield predictions were most accurate 1-2 months before harvesting. Furthermore, the study determined which variables had the greatest influence on sugarcane productivity prediction by partially excluding some variables from the prediction model. The results showed that adding variables simulated by the process-based model (APSIM) as input variables for machine learning models could reduce the Root Mean Square Error (RMSE) of yield prediction, ranging from 7.7% to 26.9%, while vegetation indices had the least impact on predictions. The analysis revealed that meteorological data had a greater impact on yield prediction when provided to the process-based model than when directly used in machine learning algorithms. This result suggests that the simulated variables provided by APSIM offer a more comprehensive biophysical description of the interaction between soil, plant, and atmosphere.

Keywords: Hybrid modeling, Agricultural yield prediction, Predictive modeling, APSIM-Sugar, Sugarcane simulation, Machine learning algorithms

1. INTRODUCTION

The prediction of crop yield through data-driven or process-based approaches has certainly evolved in recent decades, resulting in a more comprehensive understanding of crop growth and development regarding environmental variability. The advances in the fields of machine learning and process-based crop modeling, coupled with the great evolution in data generation by digital agriculture, have led to the possibility of integrating crop models and data-driven approaches, offering new opportunities to enhance yield prediction accuracy of agricultural crops (KEATING; THORBURN, 2018; FENG et al., 2020). While these technologies have already proven their unique capabilities and have advanced prediction performance, there is a growing interest in integrating them to further improve prediction accuracy (EVERINGHAM et al., 2016; CASADEBAIG; DEBAEKE; WALLACH, 2020; CORRALES et al., 2022; KHAN et al., 2022). However, in order to successfully combine these models, it is important to understand the real-world context in which they are applied. This understanding can be gained through observation and the development of models that accurately represent the essential features of the system being studied. In this sense, hybrid modeling leverages the strengths of both data-driven and process-based methods to gain a deeper understanding of the factors influencing crop yields, and can serve as a basis for creating new crop models and guide the integration of both approaches for more accurate predictions in agriculture (BATOOL et al., 2022; MAESTRINI et al., 2022).

Crop yields are influenced by various factors, such as soil quality, weather conditions, plant genotype, crop management, biotic effects, and the interactions between them. These factors can cause significant variations in crop yields, both at large regional scales and even within small-scale in-field conditions (OLIVER; ROBERTSON; WONG, 2010; WACHOWIAK et al., 2017). Adding to this, early forecasts, while desirable for planning, management, and decision-making, can add even more complexity and uncertainty in yield forecast due to unknown weather variability, pests, weeds and diseases outbreaks, as well as climate changes (HATFIELD et al., 2008; BROWN et al., 2018; CHALONER; GURR; BEBBER, 2021). This is particularly significant in semi-perennial crops like sugarcane, as they remain in the field for over a year and are exposed to various weather conditions, as well as the impact of past inadequate management and biotic and abiotic stresses, which can affect subsequent yields (VITTI et al., 2007; RAMBURAN et al., 2013; DLAMINI; ZHOU, 2022).

Brazil is the world's largest producer of sugarcane, in 2021 it produced 715 millions of tons that accounts for 40% of the global production, mainly in the state of São Paulo (IBGE, 2021; VANDENBERGHE et al., 2022). Sugarcane is a major source of sugar and biofuel, with Brazil

being the largest producer of sugarcane-based ethanol, corresponding to 30% of the world's ethanol production (KARP et al., 2021). Given the importance of sugarcane, early and accurate crop yield forecasting is important for the sugarcane production chain, supporting food security policies, market stabilization, and planning in various scales. Producers can optimize agronomic management with timely forecasts, while the agri-food industry can optimize processing, storage, transport, and marketing (JOHNSON, 2014; CROCI et al., 2023; PRIYA et al., 2023). Thus, quantifying productive variability and developing sugarcane yield estimates contributes for sustainable agricultural development. In this sense, numerous studies have focused on develop yield forecasts based on different techniques such as data-driven, process-based and remote sensing techniques, by relating yield to the environment and crop management (SCHEPEN; EVERINGHAM; WANG, 2020; AKBARIAN et al., 2022; HAN; BISHOP; FILIPPI, 2022; VERMA et al., 2023).

The process-based crop models simulates the process of plant growth and development and are used to predict yield, phenological stages, and water stress using management, crop cultivar traits, soil parameters and weather inputs (HE et al., 2017). In this way, the temporal variability of the input variables is incorporated by the models, which is one of its advantages over other productivity estimation approaches. These models are pre-trained using experimental data from various environments and are calibrated for accurate predictions. However, the variability over large regions, and also in-field, of input variables and other effects not currently simulated such as pests, weeds and diseases effects are the main constraints in process-based modeling yield prediction (BASSO et al., 2001; THORP et al., 2008; O'LEARY et al., 2016; DONATELLI et al., 2017). In contrast, remote sensing-based approaches offer an advantage in capturing biomass variability by measuring spectral differences in crop canopies, which can indirectly relate to stress factors and overcome the limitations of process-based models (FRANKE; MENZ, 2007; JIANG et al., 2014; OLIVEIRA et al., 2018; BHATTARAI; SCHMID; MCCORNACK, 2019). In this sense, vegetation indices are well-known tools to monitor crop growth and development in high spatial resolution, and therefore can be combined with process-based models to improve crop yield estimates (HUANG et al., 2015; AZZARI; JAIN; LOBELL, 2017; FENG et al., 2020). More recently, data-driven approaches based on machine learning have emerged in agriculture. These methods use algorithms trained on data to predict outputs based on inputs, making them more easily applicable than process-based models. They do not require expert knowledge or user calibration skills, and they have lower runtimes and data storage requirements (SHAHHOSSEINI et al., 2021). Several studies have shown that machine learning algorithms can accurately predict various agricultural domains, including disease and pest forecasts, as well as predictions of soil

parameters, water content, and crop yield (CHLINGARYAN; SUKKARIEH; WHELAN, 2018; LIAKOS et al., 2018; YU et al., 2019; BERTALAN et al., 2022). Thus, the main objective of this study is to develop a hybrid approach that combine relevant features that impact crop yield, such as remote sensing and process-based modeling outputs, with machine learning, to improve the accuracy of sugarcane yield forecasts.

2. OBJECTIVES

The objective of this study is to develop a hybrid approach that combines process-based modeling, remote sensing vegetation indices, and machine learning algorithms to improve crop yield prediction for the growing conditions of Southeastern Brazil. Thus, it can be subdivided into specific objectives, as follows:

i) Determine if the hybrid approach, including process-based crop modeling + remote sensing data and machine learning algorithms, can result in better sugarcane yield predictions in the top cities producing sugarcane in São Paulo state, Brazil;

ii) Identify which machine learning algorithms provide the most accurate predictions in the hybrid approach;

iii) Establish which features from the crop modeling, weather data, or vegetation indices are most relevant for use by machine learning in predicting sugarcane yield;

iv) Assess the extent to which sugarcane yield forecasts can be made earlier while maintaining a reasonable accuracy.

3. LITERATURE REVIEW

3.1. Crop yield forecasting overview

Crop yield forecasting plays a crucial role in ensuring global food security. With the world's population projected to reach more than 9 billion by 2050 (FAO, 2019), it is essential to have an accurate understanding of crop yields in order to plan for and prevent potential food shortages (HASEGAWA et al., 2021). Forecasting yield of agricultural crops accurately before harvesting is one of the main challenges in agriculture, as it allows anticipate and optimize planning, crop management and decision making in a national, regional and local levels. At a local level, it allows farmers to make informed decisions about planting, harvesting, and pricing their crops, while at a regional and national levels, an accurately yield forecasting, can help policymakers to make informed decisions about trade policies and food aid distribution (HATFIELD et al., 2008; BROWN et al., 2018). Additionally, accurate crop yield forecasting can help mitigate the impact of extreme weather events, pests, and diseases on food production, which can ultimately benefit the global food supply chain.

Seasonal yield forecast varies across temporal and spatial scales. Primarily, crop yields are affected by numerous complex factors, that includes landscapes, soil quality, genotype, pest infestations, water availability, weather conditions, management and harvest planning, and other interrelated factors (SINGELS, 2013). Additionally, the processes that determines the crop yield are time-specific and in essence they are nonlinear, adding to the complexity of the forecast process (LI et al., 2019). Various techniques exist for yield forecasting, including traditional methods where crop status is evaluated by experts, empirical and statistical models, time-series analysis, remote sensing, process-based crop models and machine learning algorithms. Traditional methods rely on expert's knowledge and are conducted throughout the crop growing season, where observations and measurements are taken, such as tiller number, biomass production, number of plants per meter square, percentage of damage from pests and fungi, percentage of weeds infestation, among others. After that, regression methods or local expertise knowledge can be used to forecast the yield from the data collected (BASSO; LIU, 2019). Thompson (1969) conducted one of the earliest crop yield forecasts by using regression analysis to identify the correlation between average regional yields and weather patterns, revealing a general trend in crop yields in the United States corn belt region. The traditional method is on the basis of yield forecast of Brazilian National Supply Company (Conab) and the Institute of Geography and Statistics (IBGE). These two institutions are responsible for supplying information on the crop status during the season. Conab and IBGE conduct field surveys and provide annual yield forecasts on a state-level, as well as estimations on

a municipality-level (which is exclusively released by IBGE) after harvest (MONTEIRO et al., 2013).

Yield forecasting using traditional experts' knowledge and *in situ* surveys that evaluate crop status have some benefits. They are easily accessible and cost-effective, they provide real-time and accurate site-specific information on crop growth, development, and yield. Also, they provide valuable insights into the potential yield and enable farmers to make informed decisions about crop management practices. Traditional methods often incorporate local knowledge and expertise that cannot be obtained through remote sensing or other modern technologies (BASSO; LIU, 2019). Despite those benefits, they present some limitations, those traditional methods are time-consuming and require manual labor, the accuracy of yield forecasting depends on the experience and expertise of the individual experts conducting the surveys. Besides that, yield forecasts may be limited to small areas, making it difficult to extrapolate the data to larger areas or regions. In situ surveys can be affected by weather conditions and other external factors that may influence the growth and development of crops, and ultimately, these methods may not be suitable for predicting long-term yield trends or changes in crop production systems (SCHAUBERGER; JÄGERMEYR; GORNOTT, 2020). However, even survey measurements based on ground measurements of the crops may present large errors (KOSMOWSKI et al., 2021).

In the past, farmers relied on their experiences and historical information for crop yield prediction and made significant cultivation decisions based on these predictions. However, given the complexity of the forecast process and the multitude of factors involved, more sophisticated methods are required to accurately forecast crop yield and complement or substitute those traditional methods. Statistical methods have been widely used for decades to forecast crop yield. One common approach involves using agrometeorological data as inputs to a statistical regression model to generate a seasonal yield forecast (SCHWALBERT et al., 2020). These models are relatively simple to construct by using historical yield data and various agrometeorological parameters such as precipitation and temperature (ROBERTS et al., 2017). However, their simplicity is also their limitation, as they may not be able to accurately forecast yields beyond the boundaries of the observed data (BECKER-RESHEF et al., 2010). Moreover, these models are poor predictors when spatial variability in soils, stresses or management practices are present (LABUS et al., 2002). Also, the increase in climate variability and extreme events has rendered these models less effective in forecasting future yields (SCHMITT et al., 2022).

Recent advances in sensor technologies and data acquisition have led to the development of more sophisticated statistical models that can extract insights from historical influences on past yields. However, these models still struggle to capture the dynamic interactions between climate,

soil, plants, and management practices (BASSO; CAMMARANO; CARFAGNA, 2013). Factors such as soil type and spatial and temporal variability are critical determinants of crop yield and need to be taken into account. While it may be challenging to include all influencing factors, integrating more comprehensive parameters in the models can improve their predictive power (SCHAUBERGER; JÄGERMEYR; GORNOTT, 2020). To further improve yield forecasting accuracy, it may be necessary to combine statistical methods with other forecasting techniques. In this scenario, the remote sensing technique played an important role to improve the yield forecast of agricultural crops, turning the process more rapid, scalable and accurate (GEIPEL et al., 2014).

Remote sensing and vegetation indices have proven to be valuable tools in predicting agricultural crop yield. Indirectly, vegetation indices, obtained from satellite and aerial imagery, can give information about crop growth and health, soil moisture, and weather patterns, and thus, can be utilized in crop yield status monitoring and yield forecast (REMBOLD et al., 2013). Vegetation indices, such as the normalized difference vegetation index (NDVI) and Enhanced vegetation index (EVI), provide a measure of plant health and can be used to identify areas of stress or poor growth (LABUS et al., 2002). This information can then be used to make predictions about crop yield potential, allowing farmers to adjust their management practices accordingly. More recently, with the ease of access to satellite and unmanned aerial vehicles, the use of remote sensing data has been widely disseminated and it became a routine practice for assessing plant status variables and monitoring agricultural areas, with studies demonstrating a good correlation between estimated and observed yield (BECKER-RESHEF et al., 2010; ZHOU et al., 2017). In that way, it is undeniable how much remote sensing has contributed to understanding the productive variability of agricultural areas, and how these developments have provided a better understanding of the interactions between plants, soil, environment, and management practices over the years (ATZBERGER, 2013; LU et al., 2020).

Remote sensing and vegetation indices offer several benefits, including the ability to cover large areas quickly and non-destructively, as well as the potential for real-time monitoring (SEGARRA et al., 2020). These tools also provide a more objective measure of crop health compared to traditional methods, which rely on visual inspections or expert opinions. However, there are some limitations to these methods, such as the need for local or over time calibration and validation, also this approach tends to be specific to each crop and region and is susceptible for errors due to cloud cover or sensor malfunctions (AZZARI; JAIN; LOBELL, 2017). Overall, remote sensing and vegetation indices are powerful tools for predicting crop yield and improving agricultural management practices.

All the previous mentioned yield forecasting methods does not model the dynamic processes and relations of plant soil and atmosphere variables that will impact crop yield at harvesting. This explains why these methods have limiting applications to other environments, leading to uncertainties in crop yield forecasting (BASSO et al., 2001). In this sense, one approach that can overcome these limitations for yield forecasting, and is widely accepted and used, are the Process-based crop models (PBM). PBM's are computational algorithms that represents the process of plant growth and development, varying according to soil and climate conditions and management practices (HE et al., 2017). Crop models such as APSIM (KEATING et al., 2003) and DSSAT (JONES et al., 2003) require meteorological data on a daily scale, and thus, the temporal variability of the input variables is incorporated by the models, one of which is its advantages over other yield forecasting approaches. PBM can be used to simulate crop yield variability in space and time, however, the requirement to perform simulations with finer spatial resolution is that the model input variables needs to be measured at a higher resolution (BASSO et al., 2001; THORP et al., 2008).

Yield forecasting methods have their benefits and limitations. To address the deficiencies of each approach, some techniques can be combined to provide a more accurate and comprehensive solution (SCHAUBERGER; JÄGERMEYR; GORNOTT, 2020). However, accurate yield predictions require input variables, and combining methods can lead to a substantial increase in data collection, depending on the desired temporal and spatial resolution. With the technological advancements in agriculture, opportunities for monitoring operations and yield forecasting have increased, resulting in a large volume of data generated from various sensors. Despite this, the amount of data can be overwhelming and beyond the capacity of human analysis, which has led to the adoption of emerging technologies such as machine learning (ML). ML algorithms can identify patterns in large datasets, and their application in agriculture has resulted in the monitoring of crop quality, the detection of diseases, pests, and weeds, and the prediction of crop yields (BEHMANN et al., 2015; ELAVARASAN et al., 2018; BEHERA; RATH; SETHY, 2021; WALDAMICHAEL et al., 2022).

To accurately predict the outcome of the target variable in different scenarios, machine learning requires sufficient training data to capture essential processes. The integration of data generated by remote sensing, variables simulated in PBM's, data collected from machines, crop management and surveys of environmental variables *in loco* is particularly the case in which the data is not homogeneous, because they are generated by sensors with different spatial, temporal and spectral resolutions. Common statistical methods cannot be applied in such cases because they are based on statistical assumptions and data distributions (CHLINGARYAN; SUKKARIEH;

WHELAN, 2018). In addition, agricultural systems' non-linearity and complexity make simple and linear correlation analyses ineffective. Machine learning algorithms can capture non-linear dependencies between response and predictor variables (SAHOO et al., 2017) and are, therefore, ideal for estimating and predicting crop yield (GONZALEZ-SANCHEZ; FRAUSTO-SOLIS; OJEDA-BUSTAMANTE, 2014; PANTAZI et al., 2016; RAMOS et al., 2017). Studies have shown that the combination of remote sensing monitoring methods, simulation of plant variables, and machine learning algorithms can improve crop yield forecasting (LOBELL et al., 2015; SHAHHOSSEINI et al., 2021; CORRALES et al., 2022; REN et al., 2023).

3.2. Advances of remote sensing in agriculture and crop yield forecasting

Remote sensing is a technique used to obtain information about objects without direct contact with them by utilizing electromagnetic radiation waves of different wavelengths, ranging from visible light to microwave bands (ALI et al., 2022). The electromagnetic radiation is used to capture data about the target object or area of interest and this radiation can come from many sources, including the sun, ground-based sensors, or active sensors that emit their own radiation (ERDLE; MISTELE; SCHMIDHALTER, 2011; SHANMUGAPRIYA et al., 2019). This technique is employed in a variety of devices for agricultural applications, including field sensors, drones, aircraft, unmanned aerial vehicles (UAVs), LIDAR, and RADAR sensors, cameras, and orbiting satellites (WÓJTOWICZ; WÓJTOWICZ; PIEKARCZYK, 2016).

The sensors used to collect remote sensing data are typically classified into passive and active ones. Passive sensors include multi and hyperspectral technologies that collect optical data using natural illumination. On the other hand, active sensors like Radio Detection and Ranging (RADAR) and Light Detection and Ranging (LiDAR) use their own illumination sources, enabling them to operate at night and in shaded areas (KAHRAMAN; BACHER, 2021). RADAR technology uses radio waves to detect and measure the distance to objects on the ground, and it can penetrate clouds and vegetation, making it useful for crop monitoring in areas with frequent cloud cover or dense vegetation (SETIYONO et al., 2018). The main applications of RADAR data include estimating crop height, canopy cover, aboveground biomass, and detecting changes in soil moisture levels that can impact crop growth and yield prediction. Synthetic Aperture Radar (SAR) data, which is a type of radar data, provides high-resolution images with detailed information about the structure and composition of crops. Polarimetric Synthetic Aperture Radar (PolSAR) data, which measures the polarization of radio waves, can provide additional information about the physical properties of crops, such as their height and moisture content (SIVASANKAR et al.,

2018). Hosseini et al. (2022) evaluated the use of satellite Sentinel-1 dual-polarimetric data for soybean yield forecasting one month before the harvest at field scales in central Argentina. The results showed that the SAR data from Sentinel-1 had a high potential for soybean yield forecast, with a coefficient of determination (R^2) of 0.81, and mean absolute error (MAE) of 581.65 kg ha^{-1}. Although RADAR systems have the ability to accurately predict yields and detect objects that are farther away, due to its wavelength, their lower resolution can be a limiting factor.

In this sense, LiDAR is a remote sensing technology that uses infrared laser pulses to create highly accurate 3D information about the Earth's surface. This technology has a shorter range, compared to RADAR, but it offers higher precision measurements at close range due to its shorter wavelength. For example, it can distinguish elevated features such as buildings and plants, and provide high-accuracy structural data for land and forestry applications (KAHRAMAN; BACHER, 2021). In agriculture, it can describe the height and structure of vegetation, and thus, can be used to estimate plant height, canopy cover, and biomass, which can be used to predict yield. Eyre et al. (2021) evaluated within-field yield prediction in cereal crops using LiDAR-derived topographic attributes in Canada. Their findings demonstrated that these attributes were effective in explaining yield variation of several crops, with general coefficient of determination (R^2) values of 0.80, 0.73 and 0.71 for corn, wheat and soybeans, respectively. This demonstrates the importance of LiDAR in precision agriculture for accurate yield estimation and optimized agricultural management. Other applications of LiDAR-derived features include Digital Surface Model, Digital Terrain Model, Canopy Height Models, and Ground Height Difference (KAHRAMAN; BACHER, 2021).

Passive sensors such as optical data can be further divided into visible and non-visible wavelengths, which correspond to different parts of the electromagnetic spectrum. The most useful wavelengths in remote sensing include visible light (VIS), near infrared (NIR), shortwave infrared (SWIR), thermal infrared (TIR), and microwave bands. NIR wavelengths are particularly useful in agriculture because they are highly reflective of healthy vegetation, making them ideal for vegetation index calculations (WÓJTOWICZ; WÓJTOWICZ; PIEKARCZYK, 2016). A widely used technique in remote sensing is thermal imaging, which involves using thermal infrared (TIR) radiation to capture images of vegetation canopies. This technique measures differences in total radiant energy and allows for temperature calculations, therefore images can then be used to assess crop health, predict yield, and identify variations in crop temperature caused by differences in water availability, nutrient levels, or other environmental factors that can affect crop growth (ZHOU et al., 2021).

In addition to the type of data, the sensor used to capture the data is also an important consideration. Different sensors have different spatial and spectral resolutions, which affect the level of detail and accuracy of the data. For example, high-resolution sensors can capture detailed information about individual plants, while low-resolution sensors can provide broader coverage of large areas (LOBELL, 2013). Overall, understanding the different types of remote sensing data and sensors is crucial to understand their applicability in geology, forestry, environmental monitoring, agriculture and ultimately in crop yield forecasting (NAVALGUND; JAYARAMAN; ROY, 2007). By selecting the appropriate data and sensor, researchers and farmers can improve their ability to monitor and predict crop performance, leading to more effective management strategies and higher yields.

The data acquired through remote sensing is crucial for the agriculture, with a wide range of applications. Some of the significant uses are the identification and monitoring of agricultural management practices (EDALAT; NADERI; EGAN, 2019), soil characteristics (SUN et al., 2012; KUNKEL; WELLS; HANCOCK, 2022), estimation of atmospheric variables such as solar radiation, temperature and rainfall (DINKU et al., 2018; HOOKER; DUVEILLER; CESCATTI, 2018; PELOSI; CHIRICO, 2021), crop growth and canopy health status (WANG et al., 2016; HASSAN et al., 2019), and to predict crop yields as well (MA et al., 2001; MARESMA et al., 2020; SKAKUN et al., 2021).

Remote sensing may have been limited in the past due to the high cost and limited availability of high-resolution data, however, this obstacle is rapidly diminishing. Additionally, remote sensing and satellite imagery data has become a feasible option due to the emergence of cloud platforms such as Google Earth Engine (GEE) (GORELICK et al., 2017) that offer easier access to vast amounts of satellite and weather data, and significantly boost processing capabilities with parallel computing resources (SCHWALBERT et al., 2020). Of all the remote sensing techniques, satellite imagery has become one of the most common, and widely used, data source for obtaining within-season information to generate crop yield forecasting (LOBELL, 2013). The most frequently employed public satellite imagers for this purpose include SPOT-Vegetation, Advanced Very High-Resolution Radiometer (AVHRR), Landsat, MODIS and Sentinel. Satellites provide bands with different wavelengths of the electromagnetic spectrum reflected by a surface, and in the case of a vegetated surface, this can be used as a valuable data for assessing crop health in agriculture. Spectral data from different bands, such as red and NIR, can be combined to calculate vegetation indices, which are more informative for crop biomass estimation than reflectance values from individual bands (BANNARI et al., 1995; BASSO; LIU, 2019).

The Normalized Difference Vegetation Index (NDVI) is the most widely used vegetation index (VI) in research and was first proposed by Rouse (1973), it is calculated as the ratio of the difference and sum of the reflectance values in the near-infrared (NIR) and red wavelengths. The sensors capture the differences reflected by the vegetation in the NIR region, because the green parts of plants reflect strongly due to scattering in the leaf mesophyll and absorb red and blue light via chlorophyll (PETTORELLI et al., 2005). Although the NDVI is the most commonly used vegetation index, there are other indices or satellite products that use different bands and wavelengths, such as the Soil Adjusted Vegetation Index (SAVI) (VENANCIO et al., 2019; NAGY et al., 2021), Enhanced Vegetation Index (EVI) (GUSSO et al., 2013; KOUADIO et al., 2014), normalized difference water index (NDWI), Green Normalized Difference Vegetation Index (GNDVI), the Burned Area Index (BAI) and the biophysical parameters of LAI and fAPAR (LÓPEZ-LOZANO et al., 2015). SAVI is a VI developed to limit the influence of soil on remotely sensed vegetation data by adding a soil adjustment factor (L), and it was first proposed by Huete (1988). The EVI is another VI extensively used in agriculture, which was designed to incorporate reflectance in the blue portion of the spectrum, in addition to the red and NIR of the NDVI calculation (LIU; HUETE, 1995). As result, he EVI exhibits better sensitivity to high biomass areas, and has improved capability in monitoring vegetation. This is achieved by decoupling the canopy background signal and reducing the impact of atmospheric conditions (HUETE et al., 1999; MATSUSHITA et al., 2007).

The NDWI is an index which can be of great interest as a support to decision making. The NDWI was designed to estimate soil moisture and canopy water content by measuring the interaction of liquid water molecules in vegetation canopies with incoming solar radiation, it incorporates a short-wave infrared (SWIR) band in its calculation, which enhances its ability to detect liquid water (GAO, 1996). This index is particularly sensitive to changes in liquid water and is influenced by local climate and soil properties that regulate water availability (JACKSON et al., 2004). Gitelson et al. (1996) proposed the GNDVI, which is another relevant VI in agriculture. Their study showed that using the green band is more effective in identifying nutritional status variations of plant canopies. They found that the GNDVI index is five times more sensitive to chlorophyll concentration than the NDVI. The index also estimates chlorophyll and leaf N content, indicating that the green band is a better option for this purpose. Another VI of interest is BAI, it is a spectral index specifically designed for discriminating burned land in the red-near-infrared spectral domain. It was tested on Landsat Thematic Mapper (TM) and NOAA Advanced Very High Resolution Radiometer (AVHRR) images, and it showed a higher discrimination ability among the indices tested, such as NDVI, SAVI, and GEMI (CHUVIECO; MARTÍN;

PALACIOS, 2002). BAI is important in agriculture as it can help identify burned areas and monitor post-fire vegetation recovery, which is crucial for land management and preventing future fires. The utility of vegetation indices varies depending on the constituent bands, wavelengths, and parameters utilized in their calculation. In summary, they can be effectively employed in diverse applications, including, but not limited to, mitigating soil background and atmosphere reflectance on spectral measurements, assessing plant water stress, identify burned area, mitigating vegetation saturation arising from high biomass among other uses.

Many studies have investigated the use of various vegetation indices in relation to crop yield forecasts. The most basic method for estimating crop yields involves establishing a correlation between ground-based yield measures and vegetation indices (VI's) measured on a single date or integrated over the growing season. Previous studies on wheat and maize showed that VI variations can account for more than 65% of the observed variations in crop yields within individual fields (SHANAHAN et al., 2001; VANNOPPEN; GOBIN, 2021) and at regional scale (MORIONDO; MASELLI; BINDI, 2007; BECKER-RESHEF et al., 2010; PANEK; GOZDOWSKI, 2020). Mkhabela et al. (2011) analyzed 10-day composite NDVI data and crop yield data for barley, canola, field peas, and spring wheat from 2000 to 2006. The results showed that MODIS-NDVI data can be effectively used to predict crop yield, with models developed for each crop, accounting for 32% to 90% of the grain yield variability. This enabled accurate yield forecasts to be made one to two months before harvest.

Bolton and Friedl (2013) used MODIS and empirical models to predict maize and soybean yield in the Central United States and found that EVI2, the EVI without the blue band (JIANG et al., 2008), was the best index for predicting maize yield in non-semi-arid areas, while NDWI performed better in semi-arid areas. NDVI and EVI2 performed equally well in predicting soybean yield. The combination of EVI2 and NDWI had significant benefits for remote sensing-based maize and soybean yield models. Recently, there has been a growing interest in combining remote sensing data with machine learning algorithms or process-based crop models to enhance the accuracy of crop yield predictions, and to overcome the limitations of each approach (AZZARI; JAIN; LOBELL, 2017; CROCI et al., 2023). Cheng et al. (2022) investigated the use of four machine learning algorithms, with multispectral and hyperspectral, to predict wheat yield in China and found that the Long Short-Term Memory (LSTM) model gave the best estimate, with a Root Mean Squared Error (RMSE) of 0.201 t ha^{-1}, outperforming the other methods. The results also showed that hyperspectral data outperformed multispectral data in predicting crop yield, with an RMSE of 0.237 t ha^{-1} compared to 0.3017 t ha^{-1}, respectively. However, when the multispectral data resolution of the Sentinel-2 data improved from 30 m to 10 m, the RMSE improved to 0.219

t ha^{-1}. Additionally, the study found that the greenness vegetation index SR outperformed traditional vegetation indices, and the shortwave infrared bands have the potential to replace visible and near-infrared bands for predicting crop yields. In a more recent study, the researchers used machine learning algorithms to estimate wheat yield using meteorological variables, satellite-driven actual evapotranspiration (ETa), and vegetation indices. ETa was found to be particularly important in improving the accuracy of the model predictions. Both Random Forest (RF) and extreme gradient boosting algorithm (XGB) generated relatively accurate results, with a MAE 0.50 t ha^{-1} and 0.39 t ha^{-1}, respectively. However, both algorithms' performances deteriorated in predicting the yield values beyond the range of the training dataset (NAGHDYZADEGAN JAHROMI et al., 2023).

The integration of remote sensing and process-based crop models has advanced significantly in recent years, with some attempts to effectively combine these two methods to improve yield forecasts. Zhang et al. (2021) developed a process-based and remote sensing driven crop yield model (PRYM-Maize) for estimating regional maize yield in the Northeast China Plain. The study found that a development stage-based grain-filling algorithm coupled with remote sensing phenology and leaf area index, presented a correlation of 0.61 and a RMSE of 1.33 t ha^{-1}, in regional maize yield estimate. In another approach, Azzari, Jain and Lobell (2017) aimed to test the performance of a new satellite-based crop yield mapping method called Scalable satellite-based Crop Yield Mapper (SCYM), which combines remote sensing imagery, crop models and weather data to generate yield estimates at a 30m resolution without ground calibration. The study compared SCYM with a simpler empirical approach (PEAKVI) using data from three regions with varying crops, field sizes, and landscape heterogeneity, with maize in the US corn belt, maize in Southern Zambia, and wheat in northern India. The results showed that SCYM outperformed PEAKVI in tracking temporal yield variations in the US due to its explicit consideration of weather. However, both methods failed to track temporal yield changes in India, possibly due to various reasons. In Zambia, PEAKVI applied to MODIS tracked yield variations better than any other yield estimate due to frequent cloud cover in this region. Overall, the study demonstrated successful approaches to yield estimation in each region and the importance of distinguishing between accuracy for spatial and temporal variation.

Remote sensing has proven to be a valuable tool for farmers, researchers, and policy-makers, providing critical information for decision-making related to crop management and food security. Remote sensing images provide timely and cost-efficient information about the Earth's surface and can be used to determine and monitor the features of the surface, making it a useful tool for agrometeorological, canopy, and soil investigations (WEISS; JACOB; DUVEILLER,

2020). Besides that, the detection of green vegetation has been significantly improved by the use of simple VIs combining visible and NIR bands. However, the characteristics of different environments vary, and thus, the choice of a specific VI should be made cautiously by comprehensively considering and analyzing the advantages and limitations of existing VI's. Additionally, applications of remote sensing may be restricted to regions where intercropping is not prevalent, as it is challenging to evaluate yields in mixed-cropping systems. To advance this field, future research should focus on enhancing and validating algorithms for yield estimation, specifically for non-cereal and rainfed crops, as well as comparing and integrating remote sensing with experimental and simulation-based studies of yield gaps (LOBELL, 2013). The recent advancements in data generation, including enhanced temporal and spatial resolutions, greater computational power from cloud computing and machine learning algorithms, and integration with other technologies like process-based crop models, have the potential to enhance the scalability and robustness of yield forecast frameworks.

3.3. Process-based crop models for yield forecasting

3.3.1. Definition of process-based crop models

Crop modeling has emerged as one of the most reliable approaches for estimating crop productivity. While empirical models are suitable for practical applications, mechanistic or process-based crop models (PBM) are more complex and were designed to understanding the underlying processes (HE et al., 2017). PBM's are designed to simulate the biophysical processes that occur between the plant and the environment by assimilating information from soil, plant, and meteorological variables, which influence the productive response of the agricultural crop (THORNLEY; JOHNSON, 2000). The main objective of a PBM is to understand how these different components interact with each other and influence plant growth. For this purpose, crop models employ a complex network of simple algorithms, with each algorithm describing a specific interaction between the components of the model. By understanding these interactions, PBM's are capable of simulating crop processes such as the growth and development, biomass and nutrient allocation pathways, dynamics of nitrogen and water in the soil and plant, as well as the impact of management practices. This is achieved using algorithms and equations dependent on variables of the soil, plant and atmosphere data. Consequently, PBM's differ from statistical models in that they do not rely on regression of end-of-season yield with within-season plant or environmental observations. Rather, they take into account the complex interactions between weather, soil, crop,

and management to simulate biomass and yield (BASSO; LIU, 2019). As a result, crop models can be employed as a reliable tool for simulating different management practices and evaluate production systems, forecast and monitor crops, and assess the impact of climate change (ASSENG et al., 2014).

Despite PBM's have been widely used to support practical management decisions, such as input of resources (e.g. water, fertilizer, pesticides), planning of processes (e.g. planting, harvesting) and simulation of the effects of different management scenarios on crop growth and yield, one of its key uses is to gather scientific research on crop modeling (PASLEY et al., 2023). PBM's can integrate diverse knowledge on crop growth and development, including crop physiology, soil science, meteorology, and agronomy. Process-based crop models can also, be used to test different hypotheses regarding the effects of environmental factors and management practices on crop growth and yield in a quantitative manner (KASAMPALIS et al., 2018). The models can be calibrated and validated using field experiments, and then be used to simulate the effects of different scenarios, such as different weather conditions or irrigation regimes, on crop performance. The comprehension gained through plant modeling can inevitably lead researchers to identify knowledge gaps and uncertainties in our understanding of crop growth and development (CHAPAGAIN et al., 2022). By comparing the model simulations with experimental data, researchers can identify areas where the model performance is poor and guide future research efforts to improve the model. Finally, PBM's can be used to extrapolate the effects of environmental factors beyond the range of conditions covered by field experiments. For example, crop models can be used to simulate the effects of future climate scenarios, such as increased temperature, changes in precipitation patterns, and elevated atmospheric CO_2 on crop growth and yield (ASSENG et al., 2014).

Crop simulation modeling started in the 1960s, where the initial associations between biomass growth and solar radiation were established (DE WIT, 1965; MONTEITH, 1965). In the 1970s, the first crop models were published (LOOMIS; RABBINGE; NG, 1979). Further progress in crop modeling occurred in the 1980s, including the creation of the wheat models named ARCWHEAT1 (PORTER, 1984) and CERES-Wheat (RITCHIE; OTTER, 1985). As described by Asseng et al. (2014), later in the 1990s, crop models for various crops were merged into crop modeling platforms such as the Decision Support System for Agrotechnology Transfer (DSSAT) (JONES et al., 2003), Agricultural Production Systems Simulator (APSIM) (KEATING et al., 2003), Environmental Policy Integrated Climate model (EPIC) (KINIRY et al., 1995) CropSyst (STÖCKLE; DONATELLI; NELSON, 2003), and Simulateur multidisciplinaire pour les Cultures Standard (STICS) (BRISSON et al., 2003). Initially, process-based crop models were designed to

model processes at plant scale, to comprehensively simulate plant functions, however, the increasing demand to simulate at field, regional and global scales, has led to the development of less complex and scalable models, which often incorporate a mixture of canopy processes and broader scale crop-climate relationships (VAN ITTERSUM; DONATELLI, 2003).

The main components of a PBM are the key processes that drives the plant growth and development, they are generally synthesized in modules that interact with each other, self-regulate, and are limited in space and time, with well-defined variables as inputs and outputs. Although, these models are designed to represent the dynamics of the soil-plant-atmosphere system and interact with climate and crop management, they may not necessarily simulate all the processes of this system. In addition, the processes it simulates always will present different sources of model uncertainties such as in the model structure, model parameters and model inputs including climate, soil, and crop management practices (CHAPAGAIN et al., 2022).

The aim of process-based crop models is to simulate real-world processes, which inevitably involves some simplification due to limitations in input variables and assumptions made to generalize model functions. While crop models use mathematical equations and functions to represent important physiological and physical processes, they are still simplified representations of the actual system processes. This simplification can lead to both known and unknown errors and biases, and the ability of process-based crop models to predict real-world outcomes is not well understood (ROBERTS et al., 2017). Sinclair and Seligman (2000) identified that a significant challenge in using these models is the need for calibration and validation through experimental field data. The input data for weather, soil, crop or genetic parameters and management practices are one of the main uncertainties on the simulation of processes including photosynthesis, respiration, biomass partitioning, nutrient and water uptake and yield formation. Also, another limitation to the PBM approach is the accuracy of sampling variables, i.e. collecting precise phenotypic information under the variety of alternative management practices and seasonal conditions required to estimate model parameters requires specialized training and facilities. Furthermore, it may be difficult or impractical to generate extensive training datasets for subtle variations in management practices to empirically optimize parameters of PBM's (ERSOZ; MARTIN; STAPLETON, 2020). However, in addition to the heterogeneity of data and the range of factors influencing the model outputs, some of these inputs can be challenging to measure and may not be available at the desired location or for a sufficient time period (e.g. soil and weather data) (BELLOCCHI et al., 2010). Therefore, estimating these inputs may require statistical or geostatistical methods, which can introduce additional uncertainties.

To generate yield forecast in the middle of growing season, PBM's require weather forecasting data to harvesting date, as input to run its processes and algorithms. As stated by Basso and Liu (2019), to predict crop yield using crop simulation models, various methods have been employed by researchers. These methods include using historical weather scenarios, averaged historical weather, weather data generated by weather generators, climate model output and satellite-derived weather data to provide the necessary weather inputs for running crop simulation models and obtaining end-of-season crop yield. In the historical weather approaches, the weather input was constructed by combining real-time weather data up to the forecasting date and historical weather scenarios (e.g. good, normal and bad climatic years) from the forecasting date until maturity (WANG et al., 2009). On the averaged historical weather, researchers construct a mean historical weather input for crop simulation models by combining real-time weather data until the forecasting date and averaged historical weather data from the forecasting date to harvesting (DUMONT et al., 2014). The weather generators consist in the use of weather generators such as LARS, SIMMETEO or WGEN (SENTELHAS et al., 2001; SOLTANI; HOOGENBOOM, 2007), which provide weather data for the remaining period after the forecasting date. The assembled weather input comprises real-time weather data until the forecasting date and the generated weather data from the weather generators (BANNAYAN; CROUT; HOOGENBOOM, 2003). The climate forecast models approach involves combining real-time weather data up to the forecasting date with weather output obtained from climate forecast models (SINGH et al., 2017; TOGLIATTI et al., 2017). In the satellite-derived weather data method, the required weather input was obtained from satellites, such as METEOSAT (THORNTON et al., 1997).

Process-based crop models are designed to be used in large homogeneous production areas. Thus, they do not consider the spatial variability of soil attributes, management, and environmental conditions, leading to uncertainties when simulating for large heterogeneous regions (HANSEN; JONES, 2000). To overcome this limitation, the simulation of plant growth in environments with great meteorological, management and soil variability requires high-resolution spatial measurements of input variables, which may pose cost and processing limitations. In this context, remote sensing is a widely employed technique to generate soil and weather data as input to crop models. Therefore, vegetation indices can be used to define spatially distinct zones in which the crop responds spectrally differently, and remote sensing products can be combined with crop growth models to improve their performance (BOUMAN, 1995; FANG et al., 2008; JIANG et al., 2014; HUANG et al., 2015). Another current difficulty with plant models is their limitation in representing the effect of factors such as pest, disease and weed control problems and management practices (O'LEARY et al., 2016; DONATELLI et al., 2017), and these variables can be measured

indirectly by remote sensing (FRANKE; MENZ, 2007; OLIVEIRA et al., 2018; BHATTARAI; SCHMID; MCCORNACK, 2019). In this sense, remote sensing allows for simple acquisition of high-resolution spatial data to simulate variations observed in the field that impact crop yield, thereby eliminating the need for measuring soil parameters, meteorological, biotical stresses and other plant state variables in a large number of sampling points in the field.

3.3.2. Assimilation of remote sensing data into process-based crop models

Remote sensing data has been integrated into PBM's through forcing method, direct use of remote sensing data, calibration of parameters or updating model simulations (INES et al., 2013; JIN et al., 2018; BASSO; LIU, 2019). In the first case the plant state variables simulated by the model are replaced by those obtained by remote sensing (e.g., biomass, leaf area index). A primary technique for integrating remote sensing data and PBM's involves aligning the simulated leaf area index (LAI) with the LAI estimated by remote sensing (BASSO; CAMMARANO; CARFAGNA, 2013). LAI is a significant agricultural parameter, as it is the primary site for gas exchange between plants and the atmosphere. Additionally, LAI is used to model crop evapotranspiration, biomass accumulation, and influences final yield (BRÉDA, 2008). The estimated LAI obtained from remote sensing data was commonly used as the true state variable and replaced the state variables from the process-based crop model (BOUMAN, 1995; THORP; HUNSAKER; FRENCH, 2010; YAO et al., 2015). In some studies, other state plant variables have been used to replace simulated values by remote sensed variables such as aboveground biomass, crop transpiration or yield to enhance the simulation results (JIN et al., 2018). Morel et al. (2012) used a forcing method for coupling remote sensing data to the MOSICAS model. The authors used the estimated interception efficiency index and FAPAR as input to estimate the yield of sugarcane. The study compared the use of satellite data with the raw simulation and two forcing methods, a partial and a complete forcing. The complete forcing approach showed the most significant improvement in yield estimation, with an estimation of the yield 8.3% superior to the observed yield. The partial forcing approach had minimal differences from the raw simulation, with a mean overestimation of respectively 34.7 and 35.4% from the observed yield.

The most used technique is the direct use of remote sensing satellite products as inputs to PBM's, such as rainfall, air temperature, solar radiation and evapotranspiration. Manatsa et al. (2011) used satellite-derived rainfall estimates in a crop water balance model to calculate the Water Requirement Satisfaction Index (WRSI) which was then regressed with historical yield data. The results showed that high skill yield forecasts can be made, even in areas with sparse or no ground

rainfall measurements. The study suggested that early estimations of maize yield are feasible using WRSI and can be useful for national and small-scale commercial farming sectors.

The calibration method involves adjusting the initial parameters of crop models to achieve an optimal agreement between the remote sensing data and the simulated state variables of the model (JIN et al., 2018). The calibration methods can be manually or automatically run using a realistic scope of different parameter values to minimize the error of simulated and observed data, when run automatically some optimization algorithms may be used such as Least Squares Method (LSM), Maximum Likelihood Solution (MLS), Particle Swarm Optimization Algorithm (PSO), among others (JIN et al., 2018). Ren et al. (2010) evaluated the use of the Shuffled Complex Evolution method (SCE-UA) algorithm to integrate remotely sensed LAI data with the EPIC crop growth model to improve the accuracy of crop growth monitoring and yield estimation. The study focused on summer maize in Huanghuaihai Plain, and the results showed that the integration led to a relative error of 4.37% and RMSE of 0.44t/ha in estimating summer maize yield. The simulated sowing date, plant density, and net nitrogen fertilization application rate also had acceptable levels of accuracy, which could meet the need of crop monitoring at a regional scale. Overall, the study concluded that integrating remotely sensed LAI with EPIC model based on SCE-UA was feasible for simulating crop growth and yield. In another research, the calibration method was used to improve the accuracy of wheat yield predictions, the calibration consisted of optimizing the set of input parameters that included the sowing date, the soil wilting point and field capacity, in order to seek the optimal input parameters which minimized the difference between remote sensing LAI and CERES-Wheat model LAI. The method was tested in Southern Italy, and the results showed that the proposed method can effectively reduce the estimation errors of yield maps, with RMSE ranging from 360 kg ha^{-1} to 420 kg ha^{-1} (DENTE et al., 2008).

The updating method includes continuously updating crop model simulation data. In summary, it is used an algorithm (e.g. Kalman Filter, 4DVar, etc.) to assimilate remote sensing data, such as LAI, into crop models by updating the system state in a sequential manner and reinitializing the model simulations based on the optimized estimates. This method can improve the accuracy of crop yield estimation by incorporating additional information from remote sensing data (JIN et al., 2018). Zhao, Chen and Shen (2013) applied this technique in a study where they evaluated the performance of the process-based crop model PyWOFOST in simulating maize growth and yield in Northeastern China, using MODIS LAI as a coupling point. The authors used the Ensemble Kalman Filter (EnKF) to optimize the estimate of the system state by separately weighting the observational and modeling errors with the observed results. The updated system state led to the reinitialization that continued the forward integration until new observations came in. Results

showed that simulated maize yield significantly improved with assimilation compared to without assimilation, with errors varying from 12.71%-10.48% compared to 14.04%, respectively. Simulated LAI with assimilation agreed better with field observations than without assimilation. The study demonstrated that it is applicable to simulate crop growth using a PBM assimilated with remote sensing data based on Ensemble Kalman Filter, and it is significant to estimate the uncertainties in crop yield estimation.

The different methods are valuable tools to assimilate remote sensing data to process-based crop models. As described by Jin et al. (2018), the calibration method requires a lot of optimization iterations and more computing time, while the updating methods significantly reduces computation time as only the crop model is run, however, the minimization algorithms errors are brought into the crop model, requires the most expensive calculation and measurement uncertainty and the date of selected remote sensing images and phenological shifts can affect the efficiency of data assimilation. The use of forcing methods are easy to operate, it just replaces the values simulated by crop models by remote sensing data, and in some cases, they provide high precision state variables. The main drawback of this method is that they assume the remote sensing data is better than the crop model data, however, remote sensing data also include errors, and these errors are propagated into crop models (INES et al., 2013). The same can be considered in the direct use of remote sensing data as inputs of process-based crop models. Soil parameters and weather variables derived from satellite are estimates and has the potential to introduce uncertainties in PBM's.

3.3.3. Sugarcane process-based crop models

For the sugarcane crop, the main simulation models in use worldwide are the DSSAT-CANEGRO (INMAN-BAMBER, 1991), APSIM-Sugar (KEATING et al., 1999), SWAP-WOFOST-Sugarcane (SUPIT; HOOIJER; VAN DIEPEN, 1994; VAN DAM et al., 1997), STICS-Sugarcane (VALADE et al., 2014) and MOSICAS (MARTINÉ; SIBAND; BONHOMME, 1999; MARTINÉ, 2003). In these models, the accumulation of biomass is governed by two primary mechanisms, the first is the sugarcane canopy's growth and interception of solar radiation, and the second is the transformation of intercepted solar radiation into carbohydrates through photosynthesis. In conditions where there is no stress, the expansion of the canopy is driven by temperature, and the conversion of radiation each day is defined by the maximum possible radiation use efficiency (RUE) and air temperature, additionally, the APSIM-Sugar model accounts for carbohydrate availability from photosynthesis (JONES et al., 2021).

The Decision Support System for Agrotechnology Transfer, also known as DSSAT, is a widely used simulation platform globally. Among the models available for sugarcane in DSSAT, the CANEGRO model (INMAN-BAMBER, 1991) is based on the CERES-Maize model (JONES and KINIRY, 1986) and is designed to simulate the physiological processes of sugarcane and the production system used by the South African industry (INMAN-BAMBER, 1991). The MOSICAS sugarcane simulation model was developed by CIRAD in 1995, and it can calculate daily crop growth variables and environmental output variables such as total biomass, root, leaf, stem, and tiller biomass, stem height, and root depth. The model uses input parameters such as meteorological, soil, plant, and variety data, as well as crop and management data. It also incorporates a soil water balance calculation routine, which enables the simulation of irrigation by adding water to precipitation (MARTINÉ, 2003).

Another PBM is the STICS model, an agronomical model that focuses on site-specific applications, providing detailed information on soil and crop processes associated with specific crop varieties and management practices. It can describe various crop species through specific parameterizations using common equations, making it a generic crop model (BRISSON et al., 1998). It offers outputs on aboveground biomass, biomass nitrogen content, water and nitrogen content in soil, yield, and root density (BRISSON et al., 2003). The model has been validated for various cropping situations, including sugarcane (VALADE et al., 2014). The Sugarcane model SWAP-WOFOST-Sugarcane is derived from the SWAP hydrological model that can simulate water flow, heat, solute transport and the crop growth component SWAP (WOrld-FOod-Study). The combination of both models results in a versatile process-based crop model that incorporates various driving and controlling processes, such as light interception, CO_2 assimilation, phenological development, respiration, dry matter formation, and assimilate partitioning. The model describes phenological development in terms of development stage, which includes vegetative and reproductive stages for many annual crops. However, when modeling sugarcane growth, only the vegetative stage is considered in SWAP-WOFOST-Sugarcane (SUPIT; HOOIJER; VAN DIEPEN, 1994; VAN DAM et al., 1997; HU et al., 2019).

3.3.4. APSIM-Sugar model overview

The APSIM platform is described in more details in the next session due to its direct utilization in the current thesis. The APSIM simulation platform is a comprehensive tool that incorporates some of the most widely used crop models in the world. The platform consists of modules that rely on algorithms and equations that take into account plant variables, soil, and

meteorological data. It enables users to simulate crop growth and development, assess management practices, and gain insights into nitrogen and water dynamics in the soil (KEATING et al., 2003). The APSIM-Sugar model enables the simulation of sugarcane's growth, development, stem yield, and Nitrogen accumulation on a per-unit-area and time basis. It requires daily meteorological data such as global solar radiation, maximum and minimum air temperature, and rainfall, as well as soil parameters and cultivar definition for its application. Fertilization, management, and irrigation information can also be included in the system for simulations. The key features of the sugarcane model in the APSIM platform are that the crop dry weight accumulation is simulated through radiation-use efficiency (RUE), which can be reduced by temperature, water, and Nitrogen limitations. Canopy expansion is determined by temperature and the same limitations mentioned above. Biomass partitioning is determined by phenological stage, and the model also simulates Nitrogen uptake and return to soil. Additionally, it simulates water content in addition to dry weight, and it can simulate differences between plant and ratoon crops based on physiological variations (KEATING et al., 1999).

The APSIM model utilizes a generic plant model with genetic adaptations and may have varying genetic coefficients for different genotypes within a species. The phenology of APSIM-Sugar is divided into phases, with each phase's duration defined by thermal time (°C d) based on lower (9°C), optimum temperature (32°C) and maximum temperature (45°C) basal temperatures. Leaf area is described by emergence, expansion, and senescence functions and solar radiation interception is determined by the leaf area index and radiation extinction coefficient of 0.38. Aboveground biomass potential is simulated based on global solar radiation use efficiency (1.8 g MJ^{-1} for plant crop and 1.65 g MJ^{-1} for ratoon crop) but limited by water availability (KEATING et al., 2003). Originally, the edaphoclimatic conditions, management and varieties described therein are different from those used in Brazil. Therefore, for the proper functioning of the model, it is necessary to calibrate it to local conditions (JAME; CUTFORTH, 1996). Model calibration consists of changing the genetic coefficients of the genotype and soil parameters to adapt the simulations to local cultivation conditions, and this could be done through field experiments. The calibration process involves monitoring the growth and development of crops, including phenology, leaf area, number of leaves, phyllochron, productivity, weather and soil conditions, and adjusting model parameters to ensure accurate simulation results (KEATING et al., 2003). Therefore, calibration of process-based crop models is critical for adequate modeling of crops under Brazilian cultivation conditions and genotypes. Some studies calibrated the APSIM-Sugar model for the Brazilian varieties and environmental conditions, such as the sugarcane variety SP80-1842 in São Paulo state by Costa et al. (2014), the varieties RB867515 and RB83594 with experimental data from Southeast

and Northeast regions from Brazil (MARIN et al., 2014) and the soil parameterization for sugarcane simulation in Espírito Santo state (OLIVEIRA et al., 2015). Although in the latter study, the authors simulated with the standard crop physiological parameters from the APSIM module for the Australian sugarcane variety Q117, as there were no Brazilian varieties parameterized in the model and insufficient physiological information on the sugarcane variety used in the experiment. More recently, the APSIM-Sugar model was calibrated for different varieties (RB867515, RB92579, RB931003, RB961003, RB98710 and SP94-3206) to optimize some parameters to account for high sugarcane yields in tropical environments (DIAS et al., 2019) and to simulate future climate scenarios in various main producer regions from Brazil (DIAS et al., 2021).

In general, these models, and the ones specific for sugarcane, have become a valuable tool in agricultural research and are being integrated into agronomic decision support systems to evaluate production systems, simulate different management practices, forecast and monitor crops, and even predict the impacts of future climate change. With their ability to provide insights into crop performance and resource use, process-based crop models can play a crucial role in improving agricultural practices and supporting plant breeding programs (HEINEMANN; STONE; SILVA, 2010; ROSENZWEIG et al., 2013).

3.4. Machine learning as a strategy to integrate remote sensing and crop model variables to improve crop yield forecast

3.4.1. Introduction to machine learning

Machine learning (ML) is the subfield of computer science that enables machines to learn and improve their performance over time without being explicitly programmed (SAMUEL, 1959). This is achieved through a learning process that involves training data as examples, which are described by a set of attributes or features that can be nominal, binary, ordinal or numeric. In this way, ML algorithms aim to optimize the performance of a task by leveraging examples or past experience and thus, the more data used, the better the ML algorithm works (SHARMA et al., 2021). The primary goal of ML is generalizability, that is, the ability to provide correct predictions when new data is presented based on learned rules from previous exposure to similar data. Data is described by a group of characteristics, or features, that form a feature vector used as input in the learning or training phase. During this phase, the machine learns from experience to perform the task, after that the developed model can then be used to classify, cluster, or predict the target variable based on unseen data from features. In summary, a computer is supplied with a dataset

and associated outputs, and the ML algorithm learns to describes the relationship between the two (CHOI et al., 2020). The performance of a machine learning model is measured by a performance metric that is enhanced over time with more training data. To calculate the performance of machine learning models and algorithms, statistical and mathematical models are commonly used. Once the learning process is completed, the trained model can be used to predict, classify or cluster new data based on the experience obtained during the training phase (BENOS et al., 2021).

In order to apply a machine learning method some crucial are involved, such as data collection, data preprocessing, feature selection and feature engineering, model selection, model training and evaluation, which involves hyperparameter tuning, and finally, deployment and monitoring of the machine learning model (JAMES et al., 2013). The first step is data collection, which is a crucial component of any machine learning project. However, researchers often encounter challenges such as a lack of data, unavailability of data in the required format, poor data quality, and data containing irrelevant features (MESHRAM et al., 2021). To prepare raw data for machine learning, pre-processing is required, which involves data cleaning to remove inconsistent or missing items and noise, data integration when there are many data sources, and data transformation, including normalization and discretization (BENOS et al., 2021). The next step is feature selection, which aims to identify the most informative subset of features for the learning model to be trained on. In some datasets, it may be necessary to reduce the number of features by using a dimensionality reduction algorithm. This is typically done as a preprocessing step before the actual machine learning task, such as classification or regression, and it can be useful in situations where the dataset has a large number of features or variables that are irrelevant or redundant. In these cases, dimensionality reduction can help to reduce the computational complexity of the task and prevent overfitting, where the model becomes highly accurate on the training data and performs poorly on unseen data (LIAKOS et al., 2018). However, it is important to note that dimensionality reduction can also lead to information loss, and the choice of algorithm and parameters should be carefully considered based on the specific dataset. Depending on the dataset, the features have to be converted to another format or some calculations are needed, in order to extract meaningful data to the learning algorithm, this process is the feature engineering (JAMES et al., 2013).

The model selection is one of the main steps in the machine learning process. There is a wide list of algorithms, which can make it challenging to determine the most appropriate one for building a customized model. Researchers can select the best algorithm for the dataset after comparing results of multiple algorithms by their statistical performance. There is no one-size-fits-all method in machine learning that can be applied to all datasets. A specific method may work well

on one dataset, but may not be the best choice for another dataset that is similar but different. Therefore, it is crucial to determine which method is most effective for a particular set of data. Choosing the appropriate approach can be one of the most difficult aspects of practical statistical learning (JAMES et al., 2013). After that, to build an accurate model, large amounts of data are needed for training, and testing and validation are crucial to ensure high model accuracy. However, building a model from scratch to achieve the desired outcome requires extensive training and testing, which can be time-consuming and resource-intensive (MESHRAM et al., 2021). Additionally, overfitting and underfitting are common challenges in building models (JAMES et al., 2013). When a model is finalized, the deployment in platforms to the final user, and implementation as routine to predict on new data is a challenging task. The models need to be continuously monitored and updated to maintain its accuracy and performance over time. It is also important to test the model in a production-like environment before deployment to ensure that it works as expected and can handle real-world data (MESHRAM et al., 2021).

Machine learning has been increasingly applied in various scientific fields such as bioinformatics, biochemistry, medicine, meteorology, economics, robotics, aquaculture, food security, and climatology. For example, in medicine, machine learning techniques are used to develop predictive models for disease diagnosis and prognosis, and to improve patient outcomes. Also, ML can enhanced comprehension of human health and diseases, by enabling comprehensive analysis of vast datasets of multi-modal data in precision medicine (MACEACHERN; FORKERT, 2021). In meteorology, machine learning models are used to predict weather patterns and improve weather forecasting accuracy, as well as to asses weather forecast uncertainties (SCHER; MESSORI, 2018). In the climatology science field, machine learning techniques are used to analyze climate data and improve our understanding of climate change (HUNTINGFORD et al., 2019). In economics, machine learning algorithms are used in areas such as predicting energy prices (e.g. crude oil, natural gas), demand forecasting, trading strategies, risk management and analyzing macro/energy trends (GHODDUSI; CREAMER; RAFIZADEH, 2019). In robotics, machine learning techniques are used to develop intelligent robots that can learn and adapt to new situations (SEMERARO; GRIFFITHS; CANGELOSI, 2023). In food security, machine learning algorithms are used to monitor crop yields and predict food shortages (ZHOU et al., 2022).

3.4.2. Types of machine learning algorithms

Machine learning comprises four main learning methods that serve distinct purposes in solving various tasks. Which can be the supervised learning, unsupervised learning, semi-supervised

learning, and reinforcement learning (JAMES et al., 2013). Unsupervised learning is a type of machine learning that focuses on discovering patterns and relationships in data that is not labeled or pre-classified. This means that the algorithm is left to identify the underlying structure on its own without any explicit guidance. One of the key goals of unsupervised learning is to identify clusters of similar cases within a dataset, which can then be further analyzed or visualized, and for this reason, this task is often referred to as a clustering problem. Popular unsupervised machine learning models include principal component analysis, k-nearest-neighbors, and variational autoencoders, which is an unsupervised deep learning architecture. By using unsupervised learning techniques, researchers can gain insights into the hidden structure of complex datasets, which can be valuable for applications such as anomaly detection, clustering, and data visualization.

Supervised machine learning is a technique used to identify patterns in multidimensional data with the help of labelled data. In this approach, a dataset with known ground truth labels is used to train a model. The model learns from the labelled data to predict labels for a new unseen data. This process is also known as classification or regression, depending on the type of problem being addressed (CHOI et al., 2020). One key advantage of supervised learning is that it allows for precise prediction of outcomes based on the input data. However, one major limitation is that it requires labelled data, which can be time-consuming and expensive to acquire (KUHN; JOHNSON, 2013). In summary, classification models in machine learning are utilized to categorize datasets, whereas regression models are commonly used to forecast continuous outcome scores. Some common supervised statistical and machine learning algorithms are support vector machine, random forests, linear models, and deep neural networks. In several cases, corresponding machine learning models are applicable to both classification and regression problems (MACEACHERN; FORKERT, 2021).

Semi-supervised learning is a type of machine learning that deals with data that has both labeled and unlabeled data points. It is considered a hybrid of supervised and unsupervised learning because it uses labeled data to train a model and then applies that model to unlabeled data to make predictions. This approach is particularly useful when the cost of labeling data is high, or when there is a scarcity of labeled data available (CHOI et al., 2020). This technique has applications in various fields such as natural language processing, computer vision, and bioinformatics. For example, in computer vision, a small number of labeled images can be used to train a model that can then classify the rest of the images in a dataset. Hu, Thomasson and Bagavathiannan (2021) proposed a semi-supervised learning pipeline for site-specific weed detection, and the results showed that, without manual labels, the pipeline achieves precision close to supervised learning in detecting weeds by images. A semi-supervised learning can be used to analyze gene expression data

where the cost of labeling genes is prohibitive. Liu et al. (2020) evaluated a semi-supervised learning with local and global consistency method-based classifier, which was employed to predict the essential genes of 41 prokaryotes. The authors concluded that the proposed method can achieve acceptable prediction performance with limited labeled data. Overall, semi-supervised learning is a powerful tool that can improve the performance of machine learning models when labeled data is scarce or expensive to obtain.

Reinforcement learning (RL) is a type of machine learning that trains an algorithm to perform a specific task, where no single answer is correct, but an overall outcome is desired (CHOI et al., 2020). This approach is similar to how humans learn from experience and trial-and-error, making it a powerful tool for solving complex problems. Reinforcement learning involves an agent, environment, and a reward signal, where the agent learns to take actions that maximize its reward within the environment (SUTTON; BARTO, 2018). The agent receives feedback from the environment in the form of rewards or penalties, allowing it to adjust its behavior to achieve the desired outcome (GAUTRON et al., 2022). Reinforcement learning has been applied to a variety of fields, such as robotics, game playing, finance and even agriculture, and has shown promising results in solving challenging problems. However, as stated by Sutton and Barto (2018), it requires careful consideration of the reward structure and exploration-exploitation trade-off, as well as robust techniques for handling high-dimensional and continuous state and action spaces. The attempts to use RL for crop management purposes are scarce or applications only considered simulated environments (GAUTRON et al., 2022). A demonstration of RL use in agriculture was proposed by Chen et al. (2021), in which the authors used a deep Q-learning (DQN) irrigation decision-making strategy based on short-term weather forecasts, with the main goal to conserve water in agriculture for paddy rice grown in Nanchang, China. The DQN irrigation strategy showed strong generalization ability and was able to make irrigation decisions using weather forecasts, resulting in irrigation water savings of 23 mm, reducing drainage by 21 mm and irrigation timing by 1.0 times on average, without significant yield reduction. Therefore, the proposed RL algorithm learned from past irrigation experiences and the uncertainties in weather forecasts and optimized the irrigation system.

3.4.3. Advantages and disadvantages of machine learning algorithms

Machine learning models offer several benefits, one of which is their ability to work with data without requiring strict assumptions about the data distribution, making it possible to combine multiple data sources with different formats, which can reduce the need for extensive data

preprocessing. This is especially helpful in agriculture, given that the data generated in modern agricultural operations is provided by a variety of different sensors, such as harvest machines, soil sensors, weather stations, satellite and UAV hyperspectral data (LIAKOS et al., 2018). Additionally, many machine learning methods use data regularization, which enables them to handle noisy data and large variances within the dataset, although outliers can still be a problem, there are various processes that can be used to mitigate their effects, such as removing outliers, transforming the data, or using robust estimators that are less sensitive to outliers (CHAKRAVARTY; DEMIRHAN; BASER, 2020). Another advantage is the availability of specialized models and architectures that can be trained on small datasets, even when the number of features significantly is higher than the number of observations (MACEACHERN; FORKERT, 2021). In biological systems, complex machine learning models can also identify non-linear patterns in the training data that may be difficult for human observers or simple linear models to detect (ALMEIDA, 2002).

Machine learning methods also has some limitations. Although it excels at identifying complex, non-linear patterns in data, this often comes at the cost of reduced interpretability compared to simpler linear models. In addition, machine learning models require significant amounts of high-quality, well-labeled training data to perform well, which can be a challenge in the agriculture domain (CHLINGARYAN; SUKKARIEH; WHELAN, 2018). Furthermore, these models may struggle with handling rare events, as they have not encountered them before and may not have the ability to generalize effectively (MAESTRINI et al., 2022). Thus, a problem of overfitting is a common problem in machine learning, where the model performs particularly well on the training data, but has unsatisfactory performance on the test data. This happens when the model is too complex and adapts too well to the training data, resulting in decreased generalizability (JAMES et al., 2013; MACEACHERN; FORKERT, 2021). Finally, machine learning models can be computationally expensive to train and deploy, requiring substantial hardware resources and technical expertise to implement and maintain (MESHRAM et al., 2021).

3.4.4. Applications of machine learning in agriculture

3.4.4.1. Machine learning for estimation of soil properties, water and crop management

Machine learning has become an important tool in agriculture, especially in the field of precision farming. Digital agriculture and its technologies have provided data-intensive approaches to better understand crop variability in the field. This enables farmers and growers to make more

informed decisions in crop management, ultimately increasing yields while minimizing the environmental impact of their operations (LIAKOS et al., 2018). Machine learning algorithms can analyze data collected from various sources such as sensors, drones, and satellites to provide insights into crop growth, soil conditions, and weather patterns. In agriculture, machine learning has a wide range of applications. These include research focused on crop management, water resources, soil management and enhancing livestock production. In crop management, machine learning techniques are being developed to improve detecting crop diseases, identifying and managing weeds, predicting crop yields, recognizing different crops (LIAKOS et al., 2018). The majority of studies have implemented machine learning methods to investigate different aspects related to yield prediction, disease detection, crop recognition, and yield quality of the main grain crops, with the top studied crops being maize, wheat, rice and soybean (BENOS et al., 2021).

Accurately predicting soil properties is crucial for various agricultural activities, including crop selection, land preparation, seed selection, crop yield, and fertilizer selection. Soil properties are directly influenced by the geographical and climatic conditions of the land, and soil temperature and humidity variability are reflected in crop yield. Besides that, soil measurements are generally time-consuming and expensive, so a low cost and reliable solution for the accurate estimation of soil can be achieved with ML techniques. By accurately estimating soil conditions, farmers and growers can improve their soil management practices and ultimately enhance their agricultural productivity (SHARMA et al., 2021). The main soil properties of interest in prediction are related to soil nutrients (HENGL et al., 2017), soil organic matter and carbon (JOHN et al., 2020), soil temperature and soil moisture content, as well as the soil hydraulic characteristics such as soil permanent wilting point, saturation and field capacity (GHORBANI et al., 2017). A more detailed review on prediction of soil parameters by machine learning can be accessed in Diaz-Gonzalez et al. (2022) and Wadoux, Minasny and Mcbratney (2020).

Effective water management is crucial for sustainable crop production, water quality improvement, and pollution reduction. Precision agriculture offers the potential for variable rate irrigation, where water is applied at rates that vary according to field variability. Therefore, monitoring soil water status, crop growth conditions, and weather conditions can help in irrigation programming and efficient water management. In this scenario, machine learning can combine weather data, remote sensing and soil properties to improve water management, particularly in arid areas where there is a greater limitation in water availability (BENOS et al., 2021). Seyedzadeh et al. (2020) employed machine learning algorithms to optimize the uniform emitter discharge rate of drip irrigation systems under varying pressure and temperature conditions. The authors used as operating pressure, water temperature, discharge coefficient, pressure exponent, and nominal

discharge as inputs, with the output being the ratio of emitter discharge to nominal discharge. Four different ML algorithms were explored, and simulation results showed that all the applied ML models presented an average mean absolute error (MAE) of 8.8%, thus, showing an acceptable accuracy for estimating the ratio of measured discharge to nominal discharge. ML techniques can also be a tool to asses groundwater quality, in substitution of the traditional methods that are expensive and laborious. In a recent study in Marocco, the authors used simple physical parameters of 520 samples from the Berrechid aquifer as features in ML models to forecast other water parameters qualities. The results showed that Adaboost and Random Forest algorithms have higher prediction performances than Artificial neural networks and Support Vector Regressor, and the developed models provided a low-cost and real-time forecast of groundwater quality, which can support the management of irrigation water strategies (EL BILALI; TALEB; BROUZIYNE, 2021).

In the crop management category, ML techniques are used to manage crops to achieve quantitative and qualitative targets by combining farming techniques to regulate the biological, chemical, and physical crop environment. The automatic recognition and classification of crops has gained attention in various scientific fields, and advancements were made through the employment of ML algorithms and remote sensing, which leveraged the automatic recognition and classification of crops (FENG et al., 2019; LOZANO-GARZON et al., 2022). Besides that, weed detection and management is a significant problem in agriculture as they are one of the most important threat to crop production (WANG; ZHANG; WEI, 2019). Accurate weed detection is crucial for sustainable agriculture as weeds are challenging to identify and distinguish from crops. ML algorithms combined with sensors can accurately detect and discriminate weeds, with low cost and no environmental side effects. Alam et al. (2020) developed a computer vision-based system for real-time weed/crop detection and variable-rate agrochemical spraying. The system used a Random Forest classifier trained with a custom dataset to detect and classify weeds/crops. Agrochemical spraying was done through a fluid flow control system guided by the vision-based feedback system. The machine learning approach showed effectiveness of the proposed system in real-time in field tests, reducing the time of image process and giving an accurate weed detection of 95%. In another study of weed detection, Tellaeche et al. (2011) presents an automatic computer vision system for identifying *Avena sterilis*, a weed seed that grows in cereal crops. The proposed system used a two-step approach involving image segmentation and decision-making using Support Vector Machines. The approach showed optimum performance in terms of computational complexity and memory requirements, using the ML algorithm.

Crop diseases are a major cause of productivity losses, as they can negatively impact agricultural production systems, leading to decreased yield and quality, as well as food insecurity on a global scale (BENOS et al., 2021). ML algorithms are widely applied to timely identify diseases and to allow efficient management of crops and reduce disease losses. In their study, Rumpf et al. (2010) suggested the use of a support vector machine with a radial basis function kernel as a model for detecting and classifying Cercospora leaf spot, leaf rust, and powdery mildew diseases in sugar beet leaves at an early stage. The accuracy of classification between diseased and non-diseased leaves was 97%, and the accuracy of identifying the three diseases was higher than 86%.

In the crop diseases domain, the vast majority of studies applied deep learning algorithms to process images, such as convolutional neural networks (CNN). An artificial neural network (ANN) is a machine learning algorithm inspired by the way the brain works. It consists of interconnected nodes called neurons that process and transmit information. ANN's can be trained to perform various tasks such as pattern recognition, classification, and prediction. A convolutional neural network (CNN) is a type of ANN that is particularly good at recognizing visual patterns such as images or videos. CNN can be viewed as a special case of ANN that preserves the spatial relationship between pixels in an image. It uses a specialized layer called a convolutional layer that applies a set of filters to the input data. These filters detect different features in the data, such as edges or corners, and pass this information on to the next layer (CHOI et al., 2020). Sethy et al. (2020), a hybrid CNN and SVM model was used to detect rice leaf disease. The model involved using a CNN to extract deep features from 5932 diseased rice leaf images, which were then fed into an SVM classifier. The resnet50 with SVM classification model was found to have the highest F1-score metric of 0.98 compared to other models. In another research, the authors proposed a new approach for identifying cucumber leaf diseases under field conditions using deep convolutional neural networks. The method involved a two-stage segmentation process to extract disease spots from the leaves and generate new training samples using generative adversarial networks. The proposed CNN achieved an average identification accuracy of 96.11% and 90.67%, when implemented on the datasets of lesion and raw field diseased leaf images, respectively. The results showed that the approach has the potential to be used in the agricultural Internet of things to recognize three different diseases anthracnose, downy mildew, and powdery mildew in field conditions.

3.4.4.2. Machine learning in crop yield forecasting

Machine learning algorithms in agriculture are increasingly focusing on crop yield forecasting, as it is a crucial factor to consider before harvesting. Soil parameters, weather variables, crop genotype, crop management and biotic factors such as pests, diseases and weeds are responsible for the crop final yield. When machine learning models are applied to a system in a systematic manner, they can serve as a feedforward control mechanism. Accurate machine learning models can anticipate the above mentioned factors that may impact crop yield, allowing corrective action to be taken before any anomalies negatively affect production (SHARMA et al., 2021).

In a machine learning approach that aimed to accurately estimate wheat yields across the Australian wheat belt, the authors found that ML can identify accurately yield gap hotspots (KAMIR; WALDNER; HOCHMAN, 2020). In their study, they used machine learning regression methods with climate data and satellite image time series, and found that, among 9 ML models, support vector regression (SVR) with radial basis function presented the best performance, achieving a yield estimate with an R^2 of 0.77 and an RMSE value of 0.55 t ha^{-1}. Another conclusion was that climate variables, such as maximum temperatures and accumulated rainfall, significantly improved yield predictions, and observations from 75 fields were required for the best single model to reach an R^2 value of 0.7. Also, the study showed that machine learning regression methods can achieve reliable crop yield monitoring across years at both the pixel and country scale.

A study evaluated the performance of Random Forests (RF), a machine learning method, in predicting crop yield responses to climate and biophysical variables at global and regional scales for wheat, maize, and potato. The RF model outperformed the multiple linear regression (MLR) benchmarks in all performance statistics, with RMSE ranging between 6 and 14% for RF models compared to 14% to 49% for MLR models. The study concludes that RF is highly capable of predicting crop yields. However, the authors suggested that RF algorithms may result in a loss of accuracy when predicting the extreme ends or responses beyond the boundaries of the training data (JEONG et al., 2016). Kouadio et al. (2018) assessed the performance of 18 different Extreme Learning Machine (ELM) based models with various predictor variables for predicting coffee yield accurately. The ELM model constructed with soil organic matter, available potassium, boron, sulphur, zinc, phosphorus, nitrogen, exchangeable calcium, magnesium, and pH as predictor variables generated the most accurate coffee yield estimate. The ELM model presented RMSE of 496.35 kg ha^{-1} and MAE of 326.40 kg ha^{-1}, which outperformed the Multiple Linear Regression (MLR), with RMSE of 1072.09 kg ha^{-1} and MAE value of 797.60 kg ha^{-1}, and Random Forest (RF) models, that showed RMSE of 1087.35 kg ha^{-1} and MAE of 769.57 kg ha^{-1}. The study suggests that the ELM model can be used as an improved class of artificial intelligence models for

coffee yield prediction in smallholder farms, and it can be coupled with biophysical-crop models in decision-support systems for precision agriculture.

In recent years, given the deficiencies of single models in processing various patterns and relationships latent in data, hybrid models have emerged as promising techniques. Hybrid models refer to models that combine the strengths of different modeling approaches to produce more accurate and reliable predictions. This hybrid approach to predictions has emerged in various field domains such as of energy consumption, traffic flow, oil price and demand, gross domestic product, inflation, health sciences, meteorology, with rainfall and drought studies and also in agriculture, with crop production (HAJIRAHIMI; KHASHEI, 2019).

In the context of agriculture and environmental sciences, hybrid models often refer to models that combine process-based crop models with machine learning models. By integrating process-based models with ML models, hybrid models can improve the accuracy of predictions, particularly in cases where the underlying processes are complex and poorly understood, and where the available data is limited (FENG et al., 2020). The reasoning for this approach to present great potential in combining the strengths of different models relies on that the process-based model establish causality between inputs and outputs based on biophysical laws, hence being directly interpretable, allowing them to predict the outcome beyond what has been observed in previous field data (REICHSTEIN et al., 2019). In order to successfully build a model, a representative training database must be established with the aid of biophysical constraints and domain knowledge. In the machine learning algorithms part, the models can identify patterns within the training data and build nonlinear relationships much faster than process-based methods, without requiring a biophysical foundation (DANNER et al., 2021). Also, ML models approaches are highly flexible in adapting to data and are responsive to finding unexpected data patterns (REICHSTEIN et al., 2019).

Feng et al. (2020) developed a hybrid yield forecasting method for wheat crops using a combination of process-based crop models variables, vegetation indices from remote sensing and machine learning regression models. The study was conducted in the southeastern Australian wheat belt, and the results showed that the forecasting accuracy increased significantly from 2 months to 1 month before harvest time, with correlation increasing from 0.62 to 0.85 and RMSE reducing from 1.01 to 0.7 t ha^{-1}, respectively. In this study, the system based on random forest outperformed the system based on multiple linear regression. The authors concluded that the hybrid approach identified that drought events throughout the growing season were the main factor causing yield losses in the wheat belt during the past decade. Similarly, Pagani et al. (2017) incorporated outputs from the DSSAT-CANEGRO sugarcane model and agro-climatic indicators into multiple linear

regressions and found that the combined model improved prediction accuracy by approximately 20% when compared to each individual model. In another hybrid modeling approach, Shahhosseini et al. (2021) evaluated whether combining crop modeling and machine learning could enhance corn yield predictions in the US Corn Belt. The research involved designing various ML models and found that integrating crop model variables can decrease yield prediction RMSE by up to 20%. Soil moisture related APSIM variables were found to be the most influential on ML predictions, as simulated APSIM average drought stress and average water table depth were identified as the most important inputs to ML models. The study concluded that ML models require more hydrological inputs to improve yield predictions. A reasonable explanation for hybrid approaches that involves process-based crop models in performing better than statistical methods is that PBM's are a good tool to capture extreme years because responses to environmental predictors mainly rely on the crop physiology rather than historical information (CHIPANSHI et al., 2015).

Everingham et al. (2016) utilized a machine learning technique with random forest algorithm, along with climate indicators and APSIM-simulated biomass as predictors, to develop a model for forecasting regional sugarcane yield in Tully, northeastern Australia. The authors evaluated the time of forecast, so models were generated on 1 September in the year before harvest, and then on 1 January and 1 March in the year of harvest. The R^2 of the random forest regression model gradually improved from 66.76%, in September of the year before harvest, to 79.21% in March of the same year of harvest. Also, it was possible to predict if the production would be above the median as early as September in the year before harvest with an accuracy of 86.36%. However, the authors did not incorporate remotely sensed vegetation indices as predictors, which could have potentially enhanced the prediction accuracy.

4. METHODOLOGY

4.1. Study site

The study was conducted with public data of sugarcane production in the state of São Paulo, Brazil from the IBGE, which provides the total planted and harvested area and crop productivity by municipality, and not the geographic location of sugarcane fields in the municipality (IBGE, 2022). The study was carried out in 10 municipalities in São Paulo State, Brazil, during 10 consecutive cropping seasons. To select the study locations, all municipalities in São Paulo State were ranked by their sugarcane total production between 2010 and 2020, and the top 10 cities with the highest production were chosen (Figure 1).

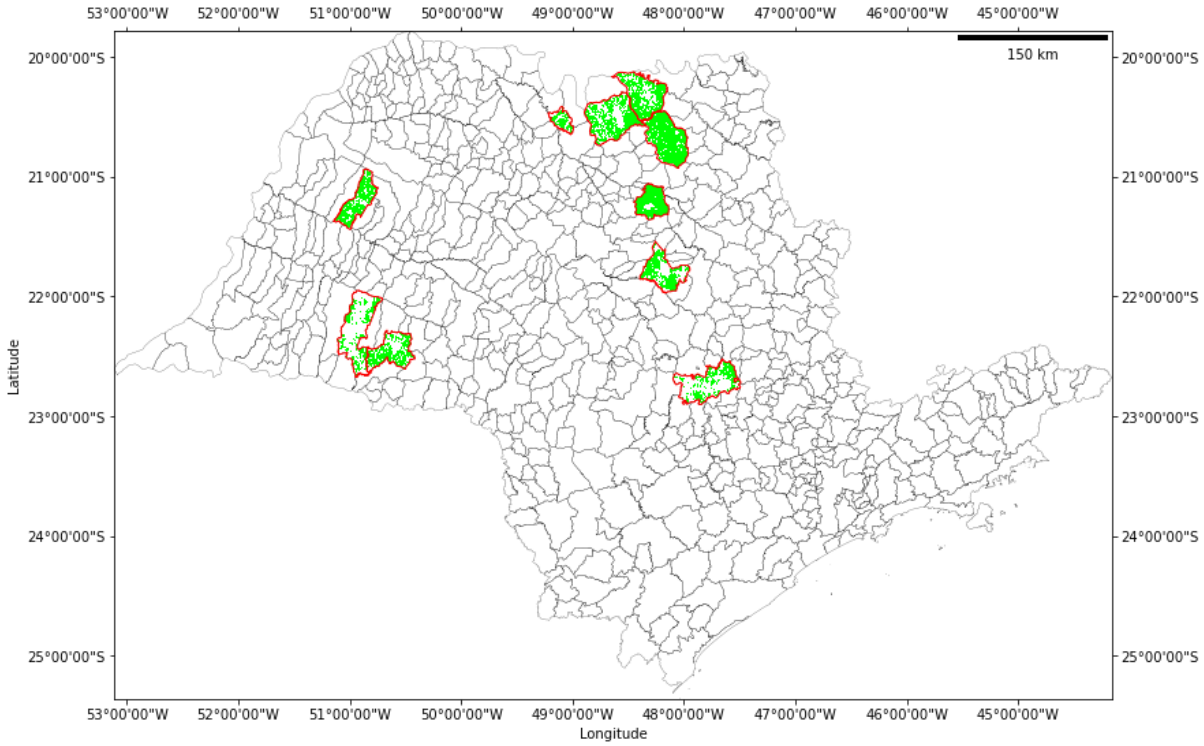

Figure 1. Selected municipalities and the sugarcane areas retrieved from the MapBiomas project, for the period of 2010-2020.

To carry out this study and obtain the multispectral data of the sugarcane crop, the geographic location of the sugarcane producing areas was first defined, using the annual maps of coverage and land use mapping reports from the Mapbiomas project (https://mapbiomas.org/) (SOUZA et al., 2020). The top 10 municipalities, in the state of São Paulo, that had the highest sugarcane production records in IBGE, between the years 2010 and 2020 were included in this step. From this initial selection, maps of land cover and land use from the annual Mapbiomas were

used to extract the geographic location of areas planted with sugarcane within each producing municipality. In this way, all images with multispectral data, extraction of meteorological data by satellite and simulation of sugarcane in the APSIM-Sugar model were carried out specifically on the locations where there were sugarcane crops planted in that period, identified by this map of land cover and land use. The Mapbiomas collection of land cover and land use maps for the time series has an average accuracy of 89%, ranging from 73 to 95% depending on the region and biome of Brazil (SOUZA et al., 2020).

After the planted sugarcane area maps were gathered, the methodology consisted in blend every Mapbiomas maps between 2010 and 2020, resulting in a final sugarcane area map that was used to derive all the subsequent remote sensing data. For each sugarcane field in the top 10 sugarcane-producing municipalities in São Paulo, a single centroid point was extracted, thereby, the latitude and longitude of each centroid point were used to identify the sugarcane fields in each city. A total of 3,838 points were generated, and their location was utilized to extract the meteorological, multispectral and soil data by remote sensing. These data extraction methods will be described in detail in the following sections.

4.2. Multi-source datasets

4.2.1. Weather data

Meteorological variables play a crucial role in crops productivity. However, public surface weather stations don't cover all the municipalities in the state of São Paulo, and there are often significant gaps in the data (SANTOS et al., 2022). To overcome the limitations posed by the lack of a network of public meteorological stations and the spatial variability of some variables, such as rainfall, it was necessary to use meteorological data from public satellites with higher spatial resolution than the state's network of weather stations. This decision was made due to the potential errors that can be introduced by the spatial interpolation of data, which can affect the simulations and interpretation of the results (PORCÙ; MILANI; PETRACCA, 2014).

For this study, the rainfall data were collected from the Google Earth Engine (GEE) platform, a cloud-based platform that allows access to a vast amount of free public satellite imagery and geospatial datasets, including data such as Landsat, Sentinel-2, and MODIS, as well as geophysical, weather, climate, and demographic data (GORELICK et al., 2017). The platform offers tools for processing and analyzing this data by two APIs, a JavaScript API and a Python API. The current study employed the earthengine-api client API module that is implemented in

the Python programming language, and has a structure similar to its JavaScript counterpart (GOMES; QUEIROZ; FERREIRA, 2020). The initial processing to obtain the multi-source datasets, was to use each of 3,838 sugarcane field centroids' latitude and longitude to access the collection of weather and multispectral data and retrieving this information in a daily aggregation.

The Rainfall data were obtained using the Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) GEE collection (FUNK et al., 2015). CHIRPS rainfall data is a collection of 0.05° high resolution, available on daily and monthly scales, covering the globe at latitudes between 50°S to 50°N. CHIRPS builds on approaches used in thermal infrared (TIR) precipitation products such as the National Oceanic and Atmospheric Administration's (NOAA's) precipitation estimate (RFE2) and the African rainfall climatology and weather series products TAMSAT, at the University of Reading (TARCAT). CHIRPS also rely on the rainfall product from the Tropical Rainfall Measuring Mission Multi-Satellite Precipitation Analysis (TRMM) version 7, to calibrate global cold cloud rainfall (CCD) duration estimates. Furthermore, CHIRPS incorporates data from surface weather stations into its algorithm (FUNK et al., 2015). According to Costa et al. (2019), when performing a comparative analysis of rainfall data between surface meteorological stations and CHIRPS, the Southeast region of Brazil is the one with the highest correlation between observed and estimated data, with a coefficient of determination of 0.99 and significant by the t-student test ($p<0.05$).

The other atmospheric variables, maximum and minimum temperature and daily global solar radiation were obtained from the ERA5-Land reanalysis collection, which was also acquired via the GEE platform. To adapt the air temperature and solar radiation products for APSIM simulations, the data were converted to a consistent unit of measurement. The air temperature, which has a temporal resolution of 1 hour, was aggregated to daily values and converted to degrees Celsius to obtain minimum and maximum values. Solar radiation, also with a resolution of 1 hour, was converted to MJ m^2 hour and summed to obtain daily values. ERA5-Land is an extensive collection of atmospheric variables from the globe, with detailed records since 1950, its data have a spatial resolution of 0.1°. The reanalysis combines data from physical models with observations from around the world into a globally complete and consistent dataset, in this way, ERA5-Land is generated by re-running the land component of the ERA5 in a higher resolution of its predecessor, thus ERA5-Land is, theoretically, an improved version of ERA5, and may provide insights over regions where observational data are unavailable (BAKER et al., 2021; MUÑOZ-SABATER et al., 2021).

The quality of solar radiation data from ERA5 was evaluated by Zuluaga et al. (2021), they compared the estimate of global solar radiation from 7 different data sources with observed data

from surface meteorological stations in Brazil, during the period of 1980 to 2016. The authors concluded that the best source of solar radiation data was the ERA5 reanalysis, with an annual bias ranging between ±20 W m-², and with no statistical difference for the data observed by surface meteorological stations. In another study, Zuluaga et al. (2023) evaluated the downward shortwave radiation of various reanalysis data sources and compared to 3 ground weather stations. In the weather stations located in São Paulo state, the authors found a correlation ranging from 0.7 to 0.8, and an RMSE ranging from 55 to 71 w m-². Thus, the dataset proved to be an adequate alternative to estimate the components of the radiation balance in southeastern Brazil. Rainfall, minimum and maximum air temperatures and global solar radiation were used as input data in the APSIM-Sugar crop simulation model to simulate sugarcane growth and development.

4.2.2. Soil data

Soil physical properties, necessary for simulation in APSIM-Sugar, were obtained from the GeoInfo platform by Embrapa Solos, for the layers 0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 depths, with a spatial resolution of 90 m (VASQUES et al., 2021). The soil physical parameters obtained were clay, silt, and sand content, soil bulk density and organic carbon content. Although the conversion factor from soil organic carbon to organic matter can vary considerably from 1.4 to 2.5, in this study it was used the common van Bemmelen factor of 1.724 to estimate soil organic matter (PRIBYL, 2010).

The physical properties were then used to estimate the hydraulic properties of the soil by pedotransfer functions (MEDRADO; LIMA, 2014), with the objective of determining the soil references Field Capacity (θ_{FC}), Wilting Point (θ_{WP}) and Saturation (θ_S) (KIRKHAM, 2005). It was used the pedotransfer functions developed by Medrado and Lima (2014) to estimate soil-water retention in this study. These functions were chosen because they have been shown to outperform the pedotransfer function developed by Tomasella, Hodnett and Rossato (2000), which is widely considered a reference for Brazilian soils, especially for tropical soils. The parameters generated by the pedotransfer functions were then introduced in the Van Genuchten equation (Equation 1) to estimate the soil water retention curve points (VAN GENUCHTEN, 1980), θ_{WP} and θ_{FC}, at the matric potential of -1500 kPa and -10 kPa, respectively. The field capacity was defined as a mean value between a clayey and a sandy soil that represents Brazilian tropical soils, following the literature for Brazilian soils (SILVA et al., 2017).

$$\theta = \theta r + \frac{\theta s - \theta r}{[1 + (\alpha h)^n]^{-m}} \quad (1)$$

where θ is the volume moisture content ($cm^3\ cm^{-3}$), h is soil matrix potential, θr is residual water content ($cm^3\ cm^{-3}$), θs is the saturated water content of the soil ($cm^3\ cm^{-3}$), α (cm^{-1}) and n are the fitting shape parameters of soil water retention curve, and $m = 1-1/n$, $n > 1$.

The Wilting point (θ_{WP}) and field capacity (θ_{FC}) were used as the lower limit and drainage upper limit parameters in APSIM. Following the official documentation to construct soils in APSIM soil module, the first layer's air dry value was set to 50% of the lower limit. As the first layer was thinner (0-5 cm), it was applied the same rule of 50% of the lower limit to better simulate air dry to the second layer as well (5-15 cm). For the 15-30 cm layer, the air dry value used was 80% of the lower limit value. For all other layers, the air dry value were equal to the lower limit value defined previously (DALGLIESH et al., 2016).

In conclusion, using the soil parameters obtained, specific soils were constructed for the 3,838 sugarcane field areas. These different soils constructed with 90 m resolution data allowed for high-resolution simulations, with specific physical and hydraulic parameters adapted in the default soil textures of sand, clay, and loam in the APSIM soil module.

4.2.3. Satellite vegetation indices

The image source to retrieve the vegetation indices was obtained from the Moderate Resolution Imaging Spectroradiometer (MODIS), aboard in two NASA Earth Observing System (EOS) satellites that provides data on various environmental variables, such as land surface temperature, vegetation cover, and ocean color. The MODIS Surface Reflectance products were used, and it specifically estimate the surface spectral reflectance as it would be measured at ground level, corrected for atmospheric gases and aerosols. The MOD09GA version 6.1 was the product utilized, which provides daily gridded data for bands 1-7 in a sinusoidal projection, including reflectance values and, observation and geolocation statistics at a resolution of 500 m and 1 km respectively. The image data comprehending 2010 and 2020 for the sugarcane areas was filtered and processed to calculate vegetation indices from the MODIS surface reflectance bands on the GEE platform. The vegetation indices calculated were NDVI (Normalized Difference Vegetation Index, Equation 2) (ROUSE; HAAS; DEERING, 1973), SAVI (Soil-Adjusted Vegetation Index, Equation 3) (HUETE, 1988), EVI (Enhanced Vegetation Index, Equation 4) (LIU; HUETE,

1995), NDMI (Normalized Difference Moisture Index, Equation 5) (HARDISKY; KLEMAS; SMART, 1983), GNDVI (Green Normalized Difference Vegetation Index, Equation 6) (GITELSON; KAUFMAN; MERZLYAK, 1996) and BAI (Burned Area Index, Equation 7) (CHUVIECO; MARTÍN; PALACIOS, 2002).

$$NDVI = \frac{NIR - R}{NIR + R} \quad ((2)$$

$$SAVI = \frac{NIR - R}{NIR + R + L}\ (1 + L) \quad ((3)$$

$$EVI = \frac{NIR - R}{NIR + C1\ R - C2\ B + L} \quad ((4)$$

$$NDMI = \frac{NIR - SWIR}{NIR + SWIR} \quad ((5)$$

$$GNDVI = \frac{NIR - G}{NIR + G} \quad ((6)$$

$$BAI = \frac{1}{(0.06 - NIR)^2 + \ (0.1 - R)^2} \quad ((7)$$

Where NIR represents the near-infrared spectrum band; R is the red band; B is the blue band; SWIR the short infrared band; L canopy background adjustment factor of 0.5; C1 and C2 are the coefficients of the aerosol resistance term, which uses the blue band to correct for aerosol influences in the red band, values of 6 and 7.5, respectively; G is the green band.

To remove pixels that were affected by complete or partial clouds over the time series in the MODIS data, a cloud mask was applied. This involved using the quality band (state_1km QA) to identify and remove those pixels that were affected by clouds (VERMOTE; VERMEULEN, 1999). Next, the daily vegetation indices were combined with weather variables in a simple data table within the GEE platform for the entire period. This was the final step of data processing in the cloud environment, resulting in weather and calculated vegetation indices in a daily time step. The Python API was then used to retrieve the data table for each sugarcane area from 2010 to 2021, which was then exported to a local computer.

The cloudy pixels can significantly affect vegetation indices and lead to errors in the analysis of crop biomass reflectance (REN et al., 2008), and to address this issue, a cloud mask was

applied. However, the application of the cloud mask resulted in gaps in the time series of vegetation indices on some days. To overcome this limitation, the vegetation indices were temporal interpolated by linear interpolation and were submitted to the Savitzky-Golay (SG) filter method (SAVITZKY; GOLAY, 1964). The SG filter is a method used to reduce contamination in time-series data, caused primarily by cloud contamination and atmospheric variability or lack of data. It is based on a simplified least-squares-fit convolution for smoothing and computing derivatives of a set of consecutive values. The main steps for smoothing a time-series using this filter include linear interpolation of cloudy values, fitting the long-term change trend, determining the weight for each point, generating a new time-series, and fitting it again (CHEN et al., 2004). The Savitzky-Golay filter is described in Equation 8.

$$y^* = \frac{\sum_{i=-m}^{i=m} C_i \, Y_{j+i}}{N} \qquad ((8)$$

Where Y is the original NDVI value, Y* is the resultant NDVI value, *Ci* is the coefficient for the NDVI value of the filter (the smoothing window), N is the number of convoluting integers (equal to the smoothing window size), and *j* is the running index of the original ordinate data table. The smoothing array or filter size is composed of 2m+1 points, where *m* is the half-width of the smoothing window.

Chen et al. (2004) stated that to apply the SG filter for NDVI time-series smoothing, two parameters need to be determined, *m* that is the half-width of the smoothing window, and *d*, which specifies the degree of the smoothing polynomial. For instance, a higher *d* value can reduce the filter bias but may give a noisier result by overfitting the data, and a smaller *d* value can produce smoother results but may introduce bias, thus *d* is typically set in a range from 2 to 4. On the other hand, for the *m* values, a larger value can produce smoother results but flatten sharp peak, conversely, the opposite is also true. Therefore, middle values of *m* in the range of 4 to 7 can be considered as appropriate parameters for generating the long-term change trend curve. In this study, it was used the SG filter with a smoothing window (*m*) of 61 observations and an intermediate smoothing polynomial of order 3. This is because Chen et al. (2004) evaluated a 10-day NDVI timeseries from SPOT VGT product, while in the current study it was used a daily MODIS product. As a result, their 4-7 smoothing window for the 10-day composite is equivalent to a 40-70 day smoothing window in the present study. This method was applied to all vegetation indices derived from the MODIS and it helped to obtain a high-quality time-series that were more useful for the analysis.

Besides that, the MODIS NDVI data was also used to create another feature, a smoother NDVI data with a half-width window of 301 days. The purpose was to create the harvest dates of the sugarcane areas, by detecting the bottom values of the NDVI timeseries for each cropping season. To determine crop growing seasons, vegetation indices can be evaluated during the period. For example, the NDVI timeseries is characterized by a narrow peak followed by a decreasing plateau, which can help us obtain the beginning and end of a growing season, however smoothing techniques are required to remove noisy data (JÖNSSON; EKLUNDH, 2004). In this study, a smoothing technique was employed using a window range that is typical for a sugarcane season. To identify bottom values or inverted peaks, it was inverted the standard peak detection algorithm from the scipy python library's signal module. The general functioning of this module is that the algorithm searches for peaks (local maxima) by comparing neighboring samples and returns those peaks whose properties match specific conditions for their height, prominence, width, threshold, and distance from each other (VIRTANEN et al., 2018). The values utilized in the half-width smoothing window and the polynomial order of the Savitzky-Golay filter were 301 days and 3, respectively. Besides that, it was used a width of 150 days in the find peaks function of scipy Signal module, which is half of a sugarcane cropping season.

A similar approach to detect maize and vineyard phenological stages, planting and harvesting dates by using the Savitzky-Golay filter was evaluated by Duarte et al. (2018), the authors found a median difference of -3 days, with an interquartile range of 10 days, between the observed dates of the end season and the estimated by the algorithm. Other studies have also determined harvesting dates at regional scales, utilizing satellite vegetation indices and smoothing algorithm approaches for rice fields (SAKAMOTO et al., 2005), wheat fields (CHU et al., 2014), sugarcane areas (WANG et al., 2020), forest and agriculture areas (PRIYADARSHI et al., 2018) and general summer and winter crops (PATEL; OZA, 2014). Also, the results of detecting the end of season was in concordance with the months that sugarcane is harvested Figure 2.

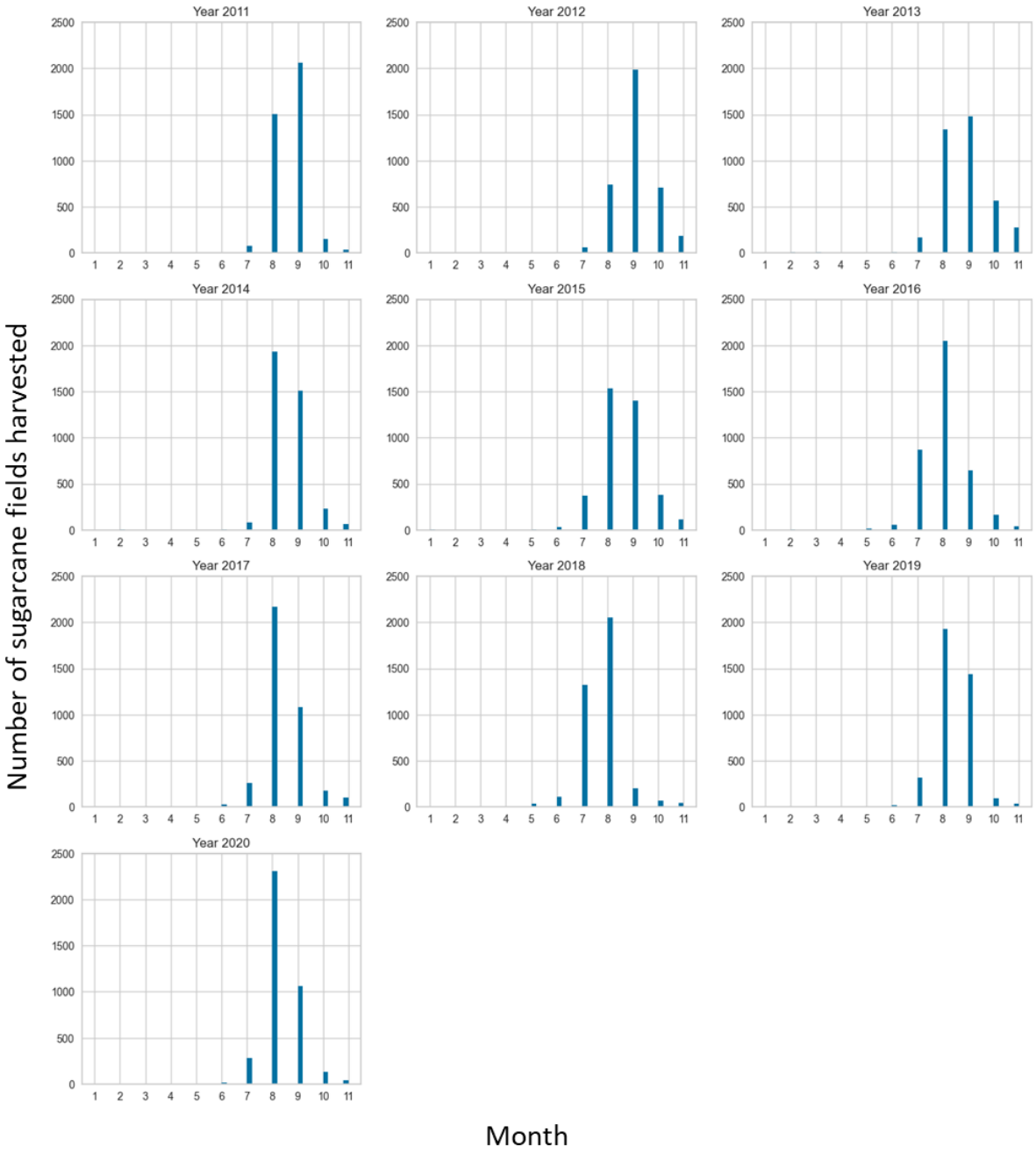

Figure 2. Number of sugarcane fields harvested per month and year detected by the inverted find peaks algorithm. The NDVI bottom values of each sugarcane cropping season defined the harvest dates.

The date of the detected bottom values of the NDVI timeseries for each season were used as the harvest day, and thus used to initialize the planting and harvest days in APSIM-Sugar management module. A typical example extracted from one of the 3,838 points timeseries can be seen in Figure 3. The different vegetation indices were used to identify spatial variability in the sugarcane productive areas, which may be the result of pest infestation, diseases, planting failures, weeds, etc.

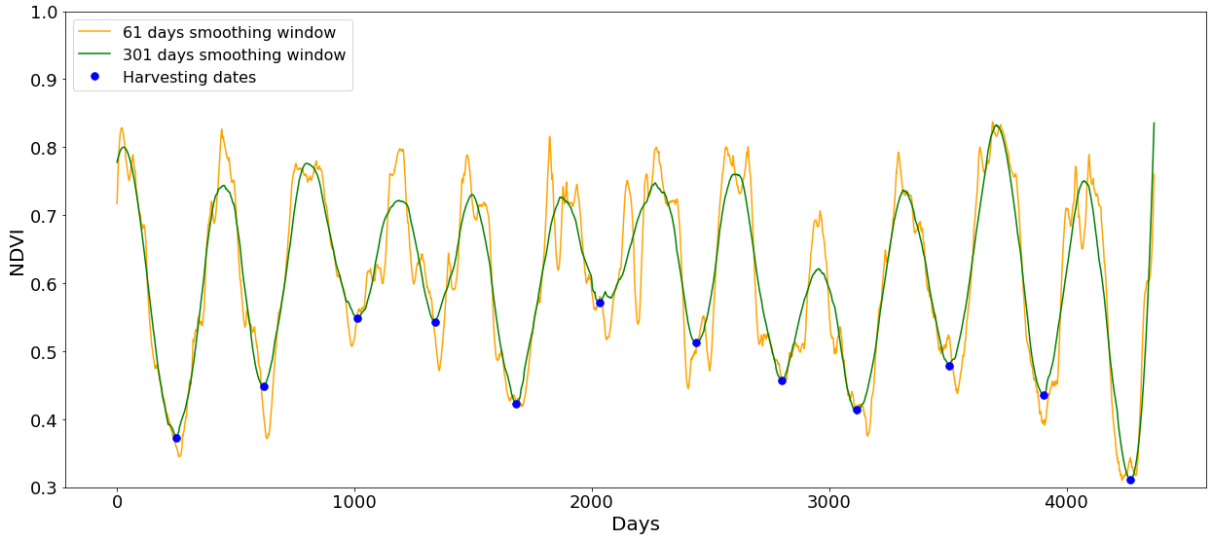

Figure 3. Sugarcane smoothed NDVI values for a single sugarcane field. Bottom values detected by the inverted find peaks algorithm to extract the sugarcane harvesting dates during the cropping seasons. A smoothing period of 301 days was used to estimate harvesting dates by detecting the bottom values of NDVI. A smoothing period of 61 days was used to minimize variability, replace missing values, and remove noisy data to construct a more reliable NDVI profile.

4.3. Modeling methodology

4.3.1. Sugarcane growth and development simulations on the APSIM-Sugar model

In the present study, simulations of sugarcane growth and development were carried out using the APSIM-Sugar model (Agricultural Production System sIMulator) version 7.10. The APSIM-Sugar model includes simulation of canopy development, light interception, nitrogen and soil water dynamics in the biomass accumulation. The model is incorporated as a crop module in the APSIM platform and runs on a daily time-step, influenced by various factors such as genotype, climate, soil water, nitrogen, and crop residues (KEATING et al., 2003). The remote sensing weather variables described in section 4.2.1 and the soil parameters were used to configure the APSIM meteorological files and the soil module. The NDVI bottom values during the growing seasons was used as inputs of the planting and harvesting dates in the management module of APSIM.

Some management practices also interfere with the development of the sugarcane crop, such as nitrogen fertilization, as the APSIM-Sugar model simulates the dynamics of this nutrient in the soil and in the plant. For this reason, nitrogen fertilization containing was included in the sugarcane simulations in APSIM, following the average fertilization recommendations for the state of São Paulo (SCHULTZ; REIS; URQUIAGA, 2015). Thus, the simulations were configured as

sugarcane rainfed systems, with stalk density of plants of 12 and 10, for sugarcane plant crop and ratoon, respectively. Fertilizing module was also applied, with mean annual nitrogen fertilizer applications of 30 kg ha^{-1} and 80 kg ha^{-1}, for plant crop and ratoon respectively.

The APSIM-Sugar model has already been calibrated and tested in different regions of the world, showing satisfactory simulation results for several sugarcane variables (BASNAYAKE et al., 2012; INMAN-BAMBER et al., 2016; GUNARATHNA et al., 2019). Also, the model has been tested and calibrated in various Brazilian soils and environments, including the Southeast and Northeast regions, which are the primary sugarcane producers in Brazil (COSTA et al., 2014; MARIN et al., 2014; OLIVEIRA et al., 2015; DIAS et al., 2019). However, there is no default Brazilian sugarcane variety in the current version 7.10 of APSIM. Therefore, in this study, the plant variety parameters described by previous researches, including Costa et al. (2014), Marin et al. (2014) and Dias et al. (2019), were collected and evaluated against the performance of the default sugarcane APSIM variety Q117. A total of 3,838 simulations were performed to evaluate the mean absolute error (MAE) between simulated and observed sugarcane yield in the top 10 sugarcane producer cities in the São Paulo State, from 2010 to 2020. The default Q117 variety resulted in a MAE of 14.3 t ha^{-1}, and the best performance was achieved with the sugarcane variety parameters derived from Dias et al. (2019), which was derived from the leaf measurements of seven Brazilian varieties in an experiment carried out by Leal (2016), with a MAE of 12.4 t ha^{-1} between simulated and observed yield. The calibration of the APSIM model is necessary to estimate or reduce uncertainties of values, increasing the reliability of the modeling results, since the crop simulation models based on processes are imperfect representations of natural processes. Uncertainty in plant growth prediction is often due to the very structure of the models and their calibration (MURPHY et al., 2004).

The variables described in the previous sections were inputs to APSIM-Sugar model, and were used to generate simulations of sugarcane variables such as biomass, plant height, leaf area index, accumulated crop evapotranspiration, accumulated water deficit, accumulated degree-days for each phenological stage, number of leaves during the cycle, accumulated water consumption per phenological stage, demand for Nitrogen in each phenological stage, average growth rate in each phenological stage, average soil and accumulated temperature. The simulations were performed in a daily time step and the output variables are described in the Appendix A. The objective of this step was to generate plant state variables at different times of the crop cycle, to later integrate these simulated variables with other variables through machine learning algorithms to forecast sugarcane yields in various months before harvest.

The use of variables simulated by process-based models, such as the APSIM-Sugar model, can provide a more complete biophysical description of the interaction between soil, plant and atmosphere. This is mainly because plant state variables represent the physiological response of crops to meteorological and soil variability. Therefore, using these variables directly in machine learning algorithms can be more suitable for estimating crop yield than using raw environmental data (PARK; HWANG; VLEK, 2005). The assumption is that using meteorological and soil variables as inputs to simulate plant growth processes is better for describing final sugarcane yield than using them directly in machine learning algorithms, as the APSIM-Sugar model biophysically describes the interaction processes of sugarcane with the environmental variables.

4.3.2. Machine learning regression models

ML algorithms involve a process of learning from data or variables that describe an outcome. From learning or training the ML algorithm, the resulting model can be used to classify or predict the result based on new data using the experience gained with the training data. Thus, in addition to the plant simulation stage with algorithms based on processes such as APSIM-Sugar, ML algorithms will be used to integrate the available variables and estimate the sugarcane productivity in advance. The algorithms that will be used belong to the supervised learning category, as they describe regression problems since the predicted variable will be sugarcane productivity (ALI et al., 2015; BEHMANN et al., 2015; LIAKOS et al., 2018).

The input data for the machine learning algorithms are comprised of weather, vegetation indices and simulated variables by APSIM-Sugar model, described in the section 4.2 and 4.3.1, but with some preprocesses described in the subsequent sections. Based on the variability in variables units and their distributions, a data preprocess was conducted to prepare it for fitting machine learning models. Scaling the input data between 0 and 1 using min-max scaling (Equation 9) was the first pre-processing task to ensure similar ranges for some machine learning models, and other tasks included cumulative weather, vegetation indices, and APSIM feature construction during growing season, and feature selection.

$$X = \frac{x - x_{min}}{x_{max} - x_{min}} \qquad ((9)$$

where X is the new scaled data, x is the range of original data, while x_{min} and x_{max} are the lowest and highest values of the features.

The scaled data was split into two datasets, with 80% of the data used for model training using repeated k-fold cross-validation with 10 folds, thus, the training data was divided into 10 groups to train the models. The 10-fold cross-validation technique helps evaluate machine learning models with a small sample of data and can reduce overfitting (KOHAVI, 1995). The cross-validation is a wide applied technique in agriculture modeling forecast, which involves splitting a dataset into subsets for analysis and validation. In the preprocessing workflow, it was implemented the leave-one-out cross-validation method, where a single observation is used as the validation value and the rest as training data, and this process is repeated until each observation is used once as a validation point (QIAN et al., 2009). The remaining 20% of the data was completely separated from the training and remained as a test dataset, with the last two years of the dataset (2019 and 2020) removed to verify prediction accuracy. This out-of-sample technique was used to prevent data leakage during the training process, because sugarcane yields can be influenced by previous sugarcane plant and ratoon deficient management practices, variety traits, soil fertility and plant conditions (MILLIGAN; GRAVOIS; MARTIN, 1996; VITTI et al., 2007; RAMBURAN et al., 2013; XU et al., 2021).

The weather, vegetation indices and APSIM variables were temporal aggregated for each sugarcane field and then, spatial aggregated by municipality, to be in the same scale as the municipality observed yield. The aggregation performed were a temporal mean and standard deviation for each sugarcane field, by expanding method. After that, a mean and standard deviation was performed in spatial scale by each municipality, which have led to a total of 158 variables. This great number of variables is due to the soil APSIM variables that have multi layers data for each sugarcane field, and the aggregations of each layer by mean and standard deviation results in many soil variables. However, in predicting yield, not all weather, vegetation indices and APSIM simulated variables are equally important. Some of the variables are redundant or less explanatory, thus, these features can lead to poor performance of machine learning algorithms. To improve accuracy, it is crucial to find important features and discard redundant ones. In this way, it was used a Principal Component Analysis (PCA) algorithm to reduce the number of variables, removing undesirable features and returning a compact data to simplify the prediction process (BRUNI; CARDINALI; VITULANO, 2022). In this study, an iterative process was implemented to determine the best proportion of variance in the PCA, which helped minimize yield prediction errors. The original dataset contained 158 variables, but a transformed dataset with fewer variables (principal components) was generated by applying PCA. Since the number of components is a hyperparameter that does not learn from data, it was determined through an iteration process, varying the hyperparameter proportion of variance from 60 to 98% by increments of 2%. The

proportion of variance that resulted in the lowest MAPE error for sugarcane yield predictions was selected. The PCA technique, developed by Pearson (1901), works by transforming a large set of variables into a smaller set of uncorrelated variables, called principal components. These components represent the underlying patterns in the data, with the first component explaining the largest amount of variability in the data, followed by the second component explaining the second largest amount, and so on. The reduction in the dimensionality space of data by PCA makes it easier to visualize and analyze large datasets, and in feature selection, can be used to identify which features have the most significant contribution to the target variable, allowing for the removal of less informative or redundant features, while retaining most of the important information (JOLLIFE; CADIMA, 2016).

For precise prediction of sugarcane yield, comprehension of the relationship between yield and its influencing factors is crucial. Different machine learning models interpret features and their influence on the target feature in distinct ways. Therefore, there is no single model that can be universally be effective in every scenario, which makes the model selection of great importance (JAMES et al., 2013). In this sense, it was evaluated 16 different ML models, including linear and non-linear models such as tree-based in the Pycaret version 2.3 Python library, which is a wrapper around several machine learning libraries and frameworks such as the scikit-learn, CatBoost, LightGBM and many other machine learning libraries (PEDREGOSA et al., 2011; KE et al., 2017; PROKHORENKOVA et al., 2018; ALI, 2020). The models evaluated were the following regressors including the AdaBoost Regressor, Bayesian Ridge, Decision Tree Regressor, Elastic Net, Extra Trees Regressor, Gradient Boosting Regressor, Huber Regressor, K Neighbors Regressor, Lasso Regression, Least Angle Regression, Light Gradient Boosting Machine, Linear Regression, Orthogonal Matching Pursuit, Passive Aggressive Regressor, Random Forest Regressor and Ridge Regression.

4.3.3. Modelling framework

The APSIM-Sugar module divides sugarcane cultivation into six growth stages, including sowing, sprouting, emergence, begin cane, flowering, and end crop. Although the model structure includes flowering as a phenological stage, it is currently inactive until there is a more reliable physiological basis for prediction. These stages are continuous and dynamic, with sugarcane growth processes transitioning into the next stage once they complete the previous one (KEATING et al., 2003). Since the long cropping season of sugarcane normally takes up from 11 to 14 months, forecasting final yield based on stages would involve too many months in some stages. To address

this issue, it was developed a forecasting approach that focuses on the months within the current harvesting year, specifically from January to November, which covers 11 different dates of forecasting. It was used the available data at the end of each month to forecast the final yield. The approach involved calculating the expanding mean and sum of variables during the cropping season, which was reinitialized at the start of each new crop season. For example, given a sugarcane field that was harvested in July of 2013, the total rainfall between the planting day in July of the last year and the end of January of 2013 was calculated using the expanding mean and sum in the current crop season, and the same process was repeated for the subsequent months until the final harvesting of that sugarcane field. Since the observed yield is at municipality level, it was developed a forecast based on calendar months, in this way, the difference between April forecast and May forecast is that in the first, there is sugarcane fields in different growth stages but a few have been harvested, while in May forecast, more fields have been harvested and there are still some fields not harvested that are now one month ahead in the growth stage. In summary, the variables in each sugarcane field are updated every month until the field is harvested. In the month it is harvested, the data from that field stops being updated. Then, every subsequent forecast will have more harvested fields with completed updated variables. Therefore, early forecasts will not have the completeness of all data, and as the months pass, more data will compound the individual sugarcane fields and the mean municipality forecast tends to be more accurate.

The general flowchart of the processing steps for developing and evaluating an in-season sugarcane yield forecasting model is described in Figure 4. The predictors for the machine learning (ML) models were plant and soil variables simulated by APSIM, as well as climate data and remote sensing information. The forecasting model provided yield forecasts from January to November. The framework involved several steps. Firstly, APSIM simulations were set up using information from remote sensing and literature on weather, soil, and management practices. The simulations were dynamically run based on the known climate data up to the end of each forecast month. Then, at the completion of each month, a forecasting event was triggered, resulting in 11 forecasting data in total. The R statistical software was used to perform 3,838 simulation points, in which a script R ran the APSIM program and collected the output data using the apsimr package (STANFILL, 2015; R CORE TEAM, 2022). Then, a Python script integrated all the APSIM variables, remote sensing and weather data to forecast the end-of-season sugarcane yield using the 16 models previously described (VANROSSUM, 1995; ALI, 2020).

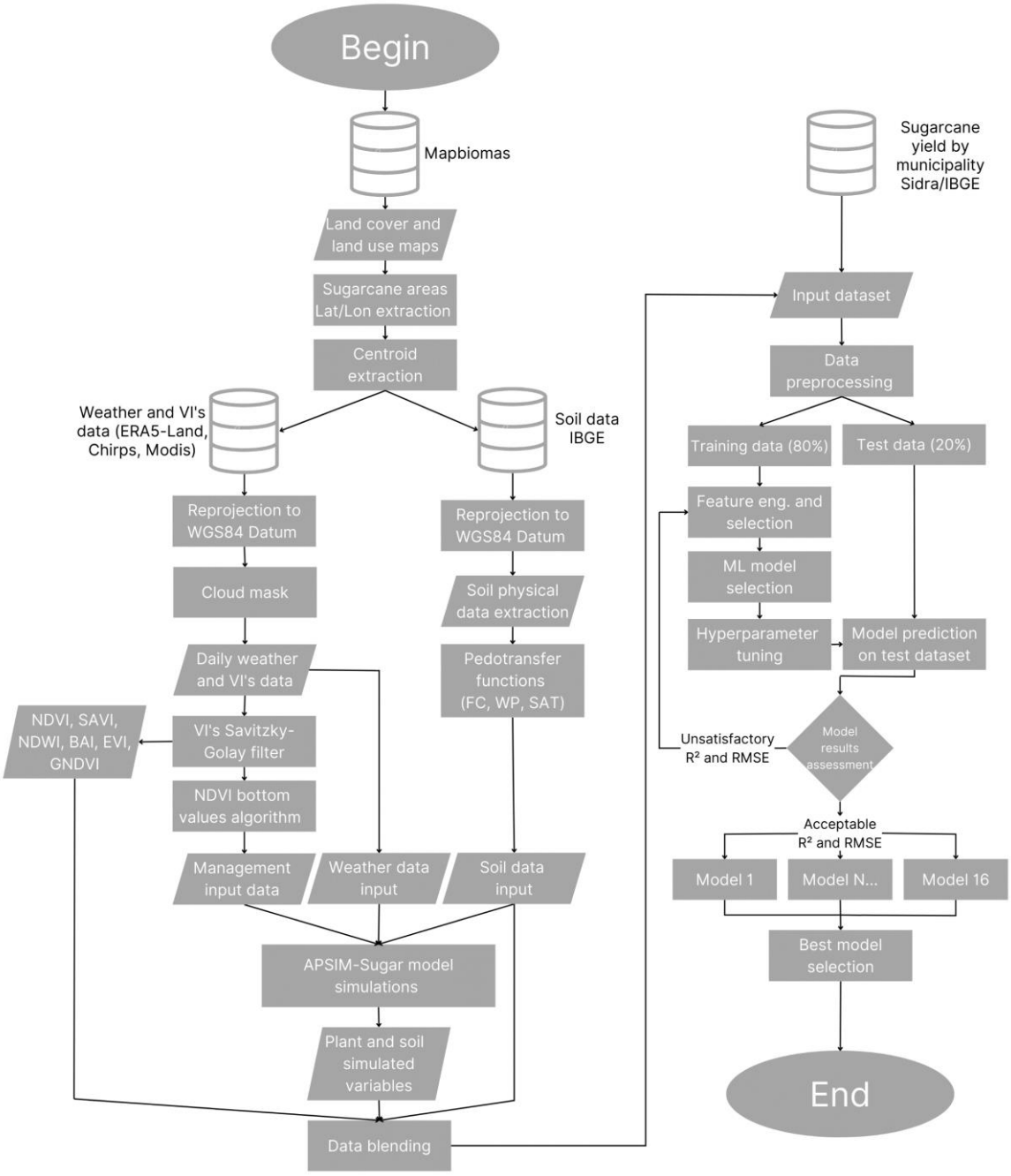

Figure 4. Flowchart overview of the steps involved in data acquisition and processing, simulation of sugarcane, and integration of various variables in machine learning algorithms to estimate sugarcane yield.

4.3.4. Model performance assessment

To monitor and evaluate the performance of the models, performance evaluation metrics were employed. The statistical metrics used were mean absolute error (MAE, Equation 10), root mean squared error (RMSE, Equation 11), root mean square logarithmic error (RMSLE, Equation 12) mean absolute percentage error (MAPE, Equation 13), coefficient of determination (R^2,

Equation 14) and coefficient of correlation (r, Equation 15). Besides that, RMSE was also used in the hyperparameter tuning of the models, in which the best hyperparameters were chosen for each model based on the lowest RMSE result of the cross-validation training step. The best model was selected as the one with the lowest RMSE, MAE, RMSLE and MAPE and the highest values of R^2 and r, in the test dataset, which comprehends data not used in the training process, from 2019 and 2020.

$$MAE = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \quad ((10)$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(\hat{y}_i - y_i)^2}{N}} \quad ((11)$$

$$RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (\log(y_i + 1) - \log(\hat{y}_i + 1))^2} \quad ((12)$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N} \frac{|y_i - \hat{y}_i|}{y_i} \quad ((13)$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad ((14)$$

$$r = \frac{\sum_i (y_i - \bar{y})(\hat{y}_i - \bar{y})}{\sqrt{\sum_i (y_i - \bar{y})^2 \sum_i (\hat{y}_{i-} \bar{y})^2}} \quad ((15)$$

Where $\hat{y}_i$ is the predicted value, y_i is the observed value, $\bar{y}$ is the mean, N is the total sample size. MAE, RMSE, RMSLE and MAPE indicate the average magnitude of the errors in a set of forecasts. R^2 measures the proportion of the variation in the dependent variable that can be explained by the independent variables. Correlation coefficient (r) measures the strength of the linear relationship between observations and forecasts.

5. RESULTS AND DISCUSSION

5.1. Analysis of weather and soil data inputs

Monthly weather conditions during the 2010 and 2021 are shown in Figure 5. Monthly maximum air temperature averaged (Tmax) increased from July to January and has a natural decrease trend from February to June for all cities. Maximum average temperatures never exceeded 40 °C and were mostly between 27-30 °C (Figure 5). Similarly, the monthly minimum air temperature averaged shows the minimum values between June and July and its maximum values in December

The ERA5-Land reanalysis air temperature at 2 m and solar radiation have been reported to agree well with ground truth weather stations observed data worldwide. Vanella et al. (2022) compared the performance of ERA5 single levels and ERA5-Land in representing agrometeorological data from 2008 to 2020 with observational data collected at 66 sites across 7 irrigation districts in Italy. The researchers found that the air temperature estimates offered the most accurate reanalysis predictions, followed by the relative humidity, solar radiation, and wind speed variables, which still provided satisfactory results. Zou et al. (2022) evaluated ERA5-Land hourly air temperature against 1080 weather stations in China, and found that correlation is generally in good agreement with in situ measurements ($r = 0.97$), but there is an overall underestimate (a bias of −0.90 °C). Studies that compares ERA5-Land to ground observations are limited in Brazil and even in South America. Baker et al. (2021) evaluated the ERA5-Land soil moisture, precipitation, surface temperature, evapotranspiration and solar radiation with satellite data and two global climate models: the Brazilian Global Atmospheric Model version 1.2 (BAM-1.2) and the U.K. Hadley Centre Global Environment Model version 3 (HadGEM3). The author found differences between products, with ERA5-Land, HadGEM3, and BAM-1.2 showing opposite interactions to satellites over parts of the Amazon and the Cerrado, and stronger land–atmosphere coupling along the North Atlantic coast. However, due to lacking in ground truth observations, the comparisons between models and satellite radiation data should be interpreted with some caution.

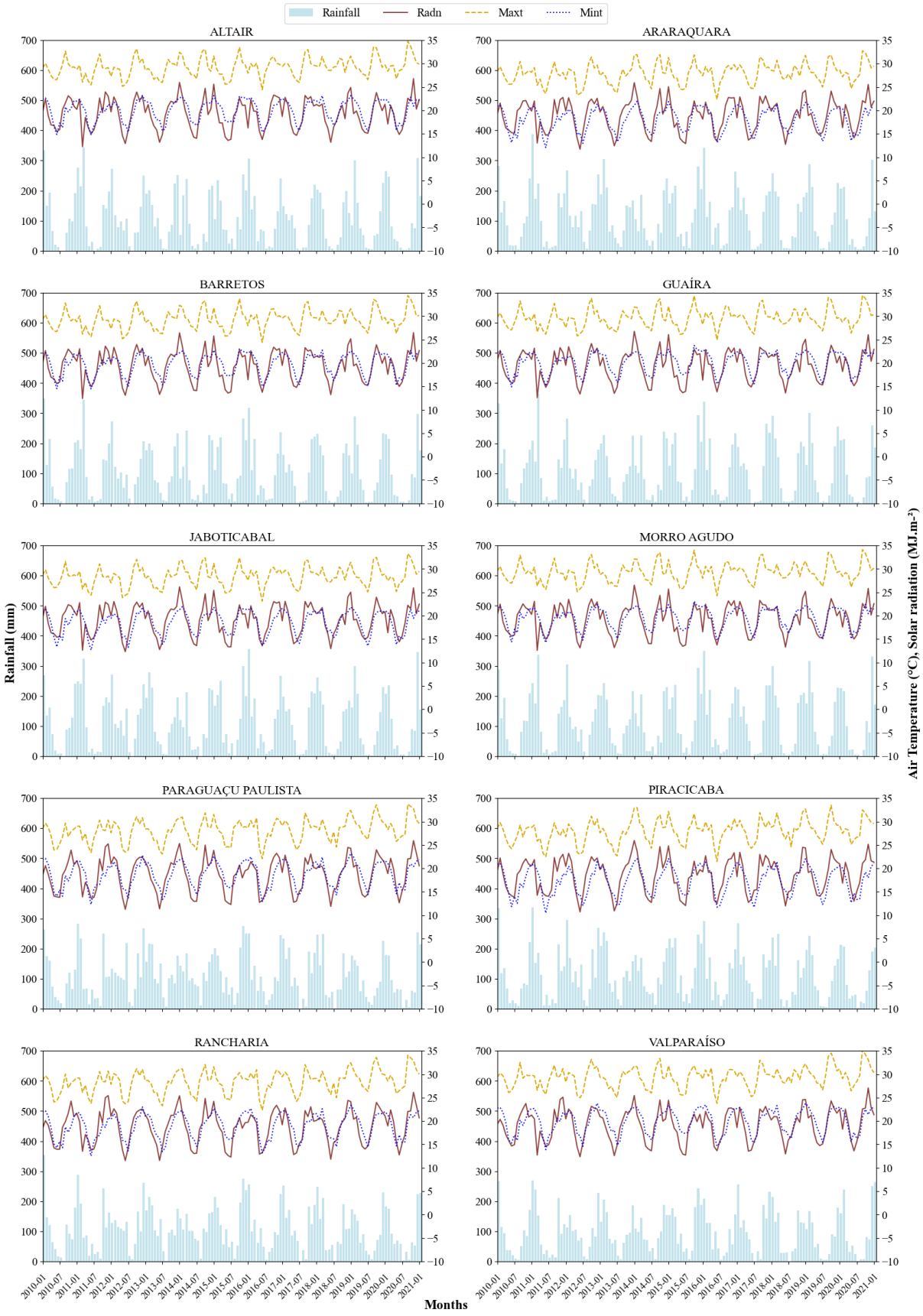

Figure 5. Monthly average of minimum (Mint) and maximum (Maxt) air temperature, global solar radiation (Radn) and accumulated rainfall (Rainfall) during the evaluated years 2010-2021 for the selected 10 municipalities in São Paulo State, Brazil.

The ERA5-Land was also evaluated by Brown et al. (2021), in a study that compared high-resolution global and regional meteorological datasets against flux tower observations at 11 sites across Brazil to assess the quality of four global reanalysis products (ERA5-Land, GLDAS2.0, GLDAS2.1, and MSWEPv2.2) and one regional gridded dataset (Xavier). Results showed that ERA5-Land achieved the best ranking for most variables, while MSWEP performed best for precipitation at the daily scale and Xavier at the monthly scale. Thus, the authors recommended ERA5-Land variables including air temperature, wind speed, pressure, downward shortwave and longwave radiation, and specific humidity for forcing land-surface models in Brazil. Zuluaga et al. (2023) evaluated different high-resolution datasets to represent radiation balance components in the southeast region of Brazil. The datasets evaluated were able to represent the seasonality of the radiation balance components, with ERA5-Land being the most accurate in representing albedo, downwelling longwave radiation, and upwelling longwave radiation, while GLASS was the best for downwelling shortwave radiation estimation. Thus, ERA5-Land and GLASS were recommended for estimating radiation balance components in the region. In Pernambuco state, the ERA5-Land reanalysis mean air temperature was evaluated over 10 years in different cities. The data were compared to 12 weather stations and assessed by R^2 and RMSE, and the results showed that ERA5-Land reanalysis has a high accuracy in estimating average air temperature in almost the entire state of Pernambuco, with the highest R^2 of 0.98 for the city of Ibimirim, while the lowest accuracy was measured in the city of Caruaru, R^2 of 0.57. The RMSE generated by ERA5-Land reanalysis was lower than 0.60°C in most of Pernambuco state (ARAÚJO et al., 2022).

In a study that evaluated the performance of four precipitation databases, including MERGE, CHIRPS, ERA5, and ERA5-Land, in the SEALBA region of Brazil, where there is a lack of weather stations. The analysis is based on seven weather stations in the region from 2001 to 2020, and the results showed that MERGE had the highest correlations and the lowest errors, with correlation of 0.96, followed by CHIRPS with 0.85, ERA5-Land with 0.83, and ERA5 with 0.70. The MERGE rainfall dataset, exhibited better performance with average MAE of 14.3 mm, followed by CHIRPS with 21.3 mm, ERA5-Land with 42.1 mm and ERA5 with 50.1 mm (SILVA et al., 2022). The ERA5-Land performed better than ERA5 due to its finer resolution, similarly the CHIRPS rainfall dataset increases even more the spatial resolution.

The CHIRPS rainfall dataset was also compared to 31 weather stations in the central region of São Paulo state from 1981 to 2020 (SANTOS et al., 2022). They found that CHIRPS presented an overall R^2 value of 0.81 and RMSE values ranging from 36.4 mm to 49.3 mm. Although, the authors noted an abrupt tendency to underestimate precipitation values. The study

highlights the usefulness of CHIRPS for locations that lack rainfall data from public weather gauges.

In the present study, the remote sensing and reanalysis data from ERA5-Land for air temperature and solar radiation, and CHIRPS for rainfall data was used due to the unavailability of public data for these weather variables with consistency and high resolution in multiple locations, which were necessary for the APSIM model simulation. For example, Santos et al. (2022) found that the data collection failure rate in some public rain gauges in São Paulo state reached 13.8%, during the 40 years period of their analyses. Besides that, instrumentation is costly and it only provides measurements that are applicable to small regions and cannot be extrapolated to larger scales, or if it is extrapolated, it certainly introduces further uncertainties by the interpolation method. In this way. to overcome these limitations and data gaps of public data, remote sensing and reanalysis techniques were considered. Reanalysis and remote sensing weather data have been shown to have high correlation with observed data across diverse environments. In addition, these data sources offer high consistency, temporal resolution, and measurements under all sky conditions (ZULUAGA et al., 2023).

The soil parameters showed high variability across different cities, with Valparaíso exhibiting soil sand content mean values above all other cities, while also having the lowest mean values of clay content as expected (Figure 6). The soil's bulk density was also obtained from the GeoInfo Embrapa dataset and it was in agreement with the other soil variables, in which Valparaíso city showed the highest mean value for the entire soil layers. Bulk density is influenced by soil texture, mineral density, organic matter, and packing arrangement. Soils that are loose and rich in organic matter have lower bulk density, while sandy soils have relatively high bulk density due to less total pore space. Finer-textured soils with good structure have higher pore space and lower bulk density compared to sandy soils.

The other estimated parameters by the pedotransfer functions followed the expected results, with Valparaíso city showing the minimum soil parameter values of air dry, drainage upper limit, and saturation. The city of Araraquara had the highest mean value of organic carbon and organic matter content, which were used as inputs for the APSIM soil module. Organic matter content tends to be higher in shallow soil layers, so its average value for the entire profile tends to decrease, which explains the low values presented in Figure 6.

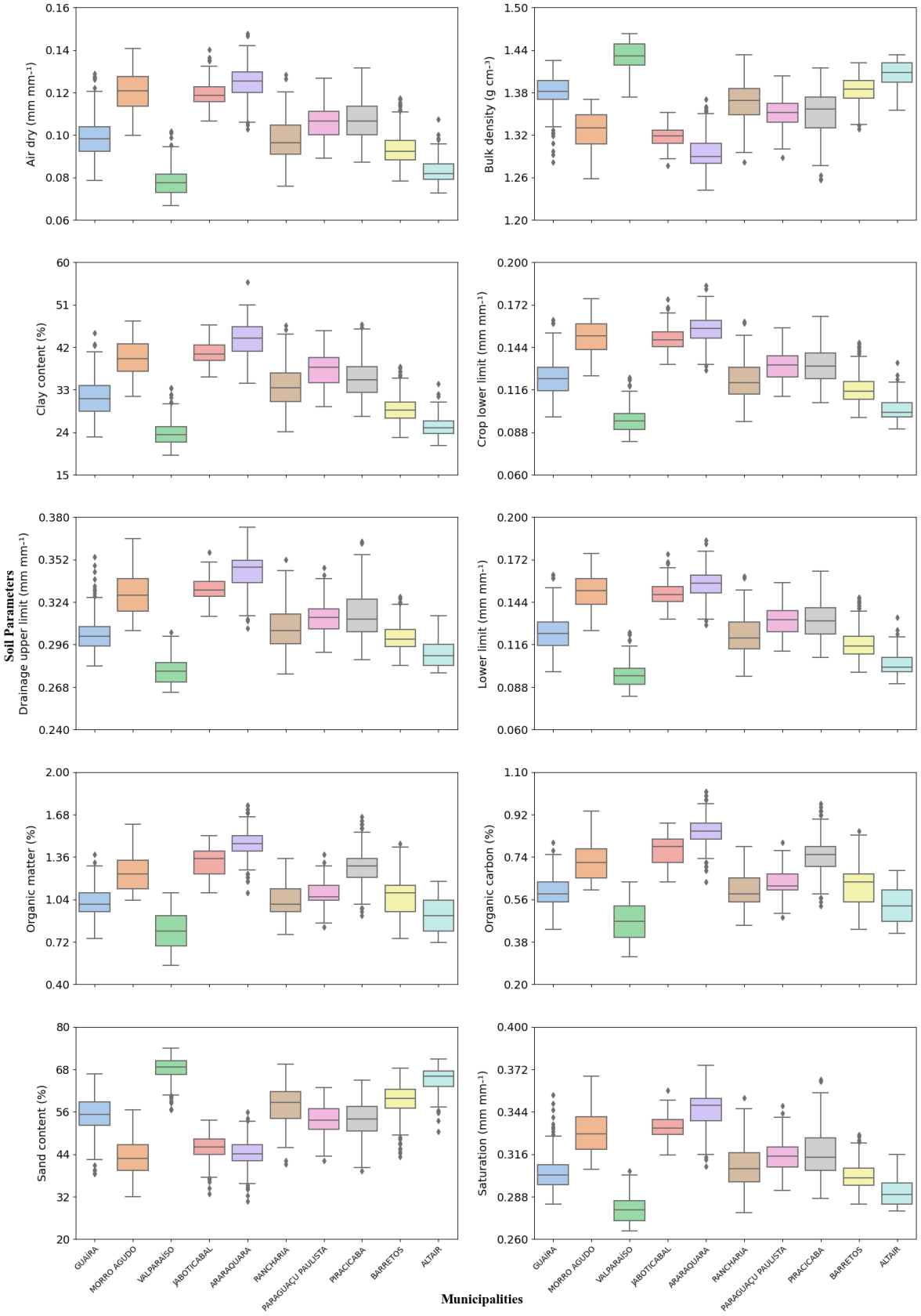

Figure 6. Boxplot of mean soil parameters for all layers used as input in APSIM soil module for each sugarcane point inside municipalities extracted from for the selected 10 municipalities in São Paulo State, Brazil.

Guaíra also showed high variability for the bulk density data, with outliers presenting low values. The opposite occurred with the hydraulic parameter's saturation, drainage upper limit and air dry values, which exhibited high variability but with outliers above the maximum of the boxplot. When soil bulk density decreases, it usually leads to an increase in total porosity and macroporosity. This results in improved soil water infiltration and air movement, which benefits soil hydraulic parameters such as the drainage upper limit, saturation, and air dry (DEC et al., 2008; INDORIA; SHARMA; REDDY, 2020), as can be observed as the presence of outlier values of these soil properties in the case of Guaíra city. Furthermore, these results demonstrate that the pedotransfer functions are producing outputs that are consistent with the physical characteristics observed in the soil. The data from certain cities, such as Guaíra, Piracicaba, and Araraquara, show a high degree of variability in soil properties. This suggests that using a single mean value for these properties in crop simulation models could lead to inaccurate results.

5.2. General results of hybrid modeling

The results of the hybrid model, i.e. machine learning with process-based APSIM variables and remote sensing vegetation index as inputs, showed high variability between machine learning algorithms. Figure 7 shows that K-Neighbors Regressor was the best model in statistical metrics, with the lowest MAE of 3.26 t ha^{-1} and RMSE of 4.48 t ha^{-1}. Also, the RMSLE error was under 0.063 and the mean absolute error relative to the mean (MAPE) showed an error as low as 4.54%. The independent variables explained more than 67% of the sugarcane yield variance and presented a correlation of 0.83 in the test dataset. In contrast, the worst performant model was the extra trees regressor algorithm, which presented the highest MAE, RMSE and RMSLE metrics of 6.13 t ha^{-1}, 8.09 t ha^{-1} and 0.111, respectively. The MAE error converted as relative to the mean sugarcane yield represent an MAPE of 8.5%, which is almost twice the error of the best model.

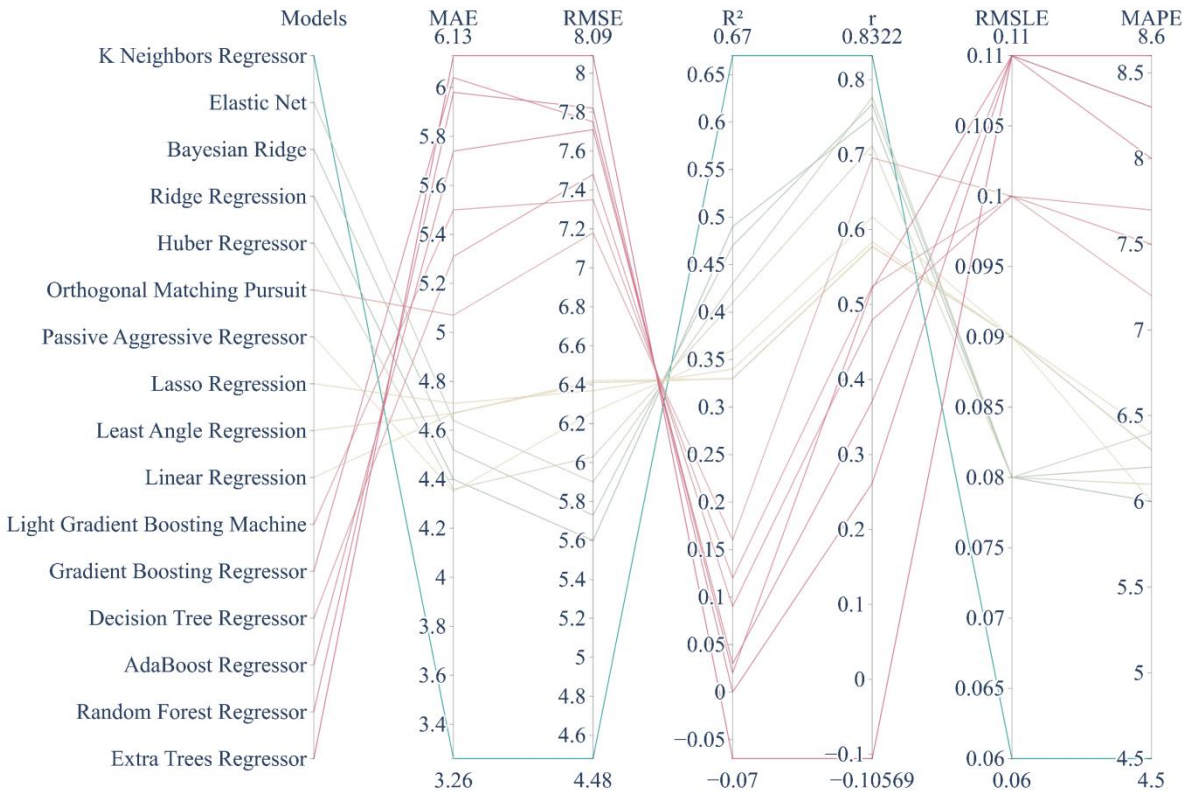

Figure 7. Machine learning model intercomparison in the test dataset for years 2019 and 2020. Predictions comprehend the data until July, 1 month before the majority of areas are harvested. Statistical metrics Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2) Correlation Coefficient (r), Root Mean Squared Logarithmic Error (RMSLE) and Mean Absolute Percentage Error (MAPE).

The agreement of the metrics r and R^2 were negative for the worst model, this indicates that the feature variables were negatively related to the yield, for correlation, while the negative R^2 value of the extra trees regressor indicated that this models' predictions are worse than taking the mean value as predictions. In a study by Hu et al. (2019), on sugarcane yield prediction in China, the SWAP-WOFOST process-based model achieved a standalone RMSE of 10.16 t ha-1 and an R^2 of 0.05. When remote sensing volumetric water content (SWC) and leaf area index (LAI) were individually introduced to the model, there was a slight improvement in yield prediction. However, the most significant performance increase was observed when both variables were included, resulting in a mean RMSE of 6.67 t ha^{-1} and an R^2 of 0.58. The authors evaluated the integration of remote sensing data in different water stress scenario, and concluded that less improvement was made when there was no water limitation factor, with the model that had no assimilation data leading to an RMSE of 7.97 t ha^{-1} and the model with both SWC and LAI data resulting in an RMSE of 6.63 t ha^{-1}. In conditions of severe water stress, both SWC and LAI data are crucial for accurately estimating sugarcane yield. However, as the level of water stress decreases, LAI becomes a more significant factor compared to SWC. When there is adequate water supply in the soil, both LAI and SWC have limited value in yield estimation, but LAI data still plays a crucial role. Morel

et al. (2014) compared three methods of yield estimation based on remote sensing using a dataset of in-farm fields. The three methods were an empirical relationship method based on remote sensing data, including an empirical relationship method with a growing season-integrated NDVI, the Kumar-Monteith efficiency model, and the last method, a forced-coupling method with a sugarcane MOSICAS model and satellite derived fraction of absorbed photosynthetically active radiation. The study found that the linear empirical model produced the best results with an RMSE of 10.4 t ha^{-1}. However, in the same study, the researchers found that the simulated sugarcane yield by MOSICAS coupled with remote sensing data performed better than the standalone model, with a RMSE improvement from 15.3 t ha^{-1} to 12.6 t ha^{-1}, and an R^2 from 0.21 to 0.47.

In Brazil, the main approaches of sugarcane estimation are the process-based crop models for site specific or experimental data or by remote sensing for in-field or large regional scales, but with no integration of the methodologies. Marin et al. (2015) utilized field data from sugarcane plant crop, from seven locations throughout Brazil, with two experiments having two water-limitation treatments and the remaining five experiments grown under rainfed treatments, and five distinct soils types. The authors evaluated the APSIM-Sugar and DSSAT-CANEGRO for sugarcane yield prediction and found a RMSE of 20.09 t ha^{-1} and 18.27 t ha^{-1} and a correlation (r) of 0.94 and 0.95 for the models, respectively. In another research, Dias et al. (2019) implemented a new feature in APSIM-sugar to account for high sugarcane yields in tropical environments, with experimental data from Guadalupe in Piauí state, and the results showed a RMSE of 18.7 t ha^{-1} and a R^2 of 0.69 of the APSIM simulated stalk fresh yield compared to observed stalk fresh yield, in contrast with the default APSIM-Sugar model without the new feature, which had a RMSE of 26.22 t ha^{-1} and a R^2 of 0.86. In a broader regional scale, another study was conducted to predict sugarcane crop yield using the FAO model, a process-based model, with simulations for a period between 1974 and 2003 for 178 locations in São Paulo state, using 10-day period weather data. The model performed with accuracy, as proved by the statistical coefficients R^2 value of 0.58 and RMSE was 5.0 t ha^{-1} (MONTEIRO; SENTELHAS, 2014).

In general, remote sensing combined with machine learning methods tend to provide more accurate predictions at regional scales compared to process-based models, as they require less input data estimates. In a study that aimed to predict sugarcane yield in São Paulo State for 60 municipalities, during 2003 and 2012, using metrics derived from NDVI time series from the MODIS sensor and an ensemble model of artificial neural networks (ANNs), presented an R^2 of 0.61 and a RMSE of 5.7 t ha^{-1} (FERNANDES; EBECKEN; ESQUERDO, 2017). In another experiment, the authors evaluated three machine learning algorithms applied to data from multiple sugar mills in São Paulo, Brazil, and the models were compared to an independent data set. The

ML models had a RMSE ranging from 19.7 t ha^{-1} to 20.03 t ha^{-1}, MAE from 14.8 to 15.3 t ha^{-1} and correlation coefficient ranging from 0.64 to 0.66, when compared to observed yields (HAMMER; SENTELHAS; MARIANO, 2020). They concluded that the ML models precision and accuracy, in the field block level of mills, are still lower than desirable for operational applications in a possible crop forecasting system. The authors also found that, in all ML models evaluated, the most important variable in sugarcane yield prediction was the number of cuts. This highlights that regional public yield estimates, which were used in the present research, lack of important input data for the models, and that obtaining this data at the field scale is crucial for improving forecasts at regional scales. Future research should focus on estimating this variable to improve yield predictions.

Evaluating the results of the proposed method and based on the structure of the algorithms, it is possible to infer that the k-neighbors regressor performed the best because it is a non-parametric algorithm that relies on similarity measures to make predictions. This means that it is not limited by the assumptions of a parametric model and can capture the relationships between the features and the target variable (HAWORTH; CHENG, 2012). On the other hand, the linear algorithms performed intermediate because they are parametric models that make assumptions about the underlying relationship between the features and the target variable. If these assumptions are violated, the performance of the model can suffer (SCHMIDT; FINAN, 2018). Finally, the tree-based algorithms performed worse because they can suffer from overfitting due to their high variance. This means that they can fit the training data excessively, but not generalize well to unseen data (MIENYE; SUN; WANG, 2019). Besides that, it may also be possible to infer that the relationship between the predictor variables and the target variable is not very complex or nonlinear, which is why linear algorithms performed reasonably well. On the other hand, tree-based algorithms tend to perform better when the relationships between the variables are nonlinear or there are interactions between them (GARDNER; DORLING, 2000). A key important factor, that could have influenced the results, and can describe why the linear models performed reasonably well, was that linear algorithms are more likely to benefit from principal component analysis since they assume a linear relationship between the input variables and the output components. By reducing the number of input variables through PCA, the linear algorithm can become more efficient in identifying the underlying linear relationship in the data. Additionally, training classifiers on reduced dimension data leads to enhanced resilience of ML classifiers, since it reduces the weights of less informative and low-variance features (BHAGOJI et al., 2018). While tree-based algorithms, in contrast, are better suited to handle high-dimensional data and can handle interactions between variables without relying on feature reduction techniques such as PCA.

Additionally, the better performance of k-nearest neighbors regressor algorithm suggests that there may be a strong local structure in the data, where points that are closer together in the predictor variable space tend to have similar target variable values. This means that the k-NN regressor algorithm relies heavily on the structure and distribution of the data points in the immediate vicinity of a given data point to make predictions. If the data points are clustered or distributed unevenly, the algorithm may not perform well and could produce inaccurate predictions. In other words, the k-NN regressor algorithm is highly dependent on the locality of the data (ZHANG et al., 2017).

5.3. Results of partial exclusion of variables

This section investigates the effect of partial inclusion of APSIM variables, remote sensing variables vegetation indices and meteorological variables in the performance of ML models on the test years 2019 and 2020. The scenarios are (1) Full dataset without the removal of variables, i.e. dataset includes remote sensing vegetation indices, APSIM variables and meteorological variables; (2) All dataset except APSIM variables; (3) All dataset except remote sensing vegetation indices; (4) All dataset except meteorological variables; (5) Dataset excluding APSIM and meteorological variables; (6) Dataset excluding APSIM and remote sensing vegetation indices; (6) Dataset without meteorological variables and vegetation indices and (7) Dataset excluding APSIM variables and vegetation indices. The averaged statistical metrics of the 7 distinct scenarios of variables are represented in Table 1. The results suggest that the K Neighbors Regressor remains the top-performing ML model, even with varying scenarios, indicating its robustness when features are scarce for predicting sugarcane yield. On average, the model achieved an R^2 of 0.54, a correlation of 0.78, and a MAE below 4 t ha^{-1}, resulting in a mean MAPE of less than 5.4%. Conversely, the Decision Tree Regressor was the worst model to predict sugarcane yield in a variety of missing variables scenarios, which led to poor correlation of 0.32 and R^2 of only 6%, indicating that the yield prediction does not follow the observed yield trend.

Table 1. Averaged statistical metrics for 7 different removal variable scenarios for all machine learning models, in the test dataset for years 2019 and 2020. Predictions comprehend the data until July, 1 month before the majority of areas harvesting. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Percentage Error (MAPE) and Correlation Coefficient (r).

Model	MAE	RMSE	R^2	RMSLE	MAPE	r
	t ha^{-1}	t ha^{-1}			%	
K Neighbors Regressor	3.88	5.27	0.54	0.074	5.4	0.78
Ridge Regression	4.55	5.83	0.44	0.081	6.2	0.74
Bayesian Ridge	4.61	5.94	0.43	0.083	6.3	0.74
Lasso Regression	4.52	5.98	0.42	0.082	6.1	0.66
Least Angle Regression	4.51	5.99	0.41	0.082	6.0	0.65
Linear Regression	4.51	5.99	0.41	0.082	6.0	0.65
Passive Aggressive Regressor	4.38	6.12	0.39	0.085	6.0	0.66
Elastic Net	4.84	6.22	0.37	0.087	6.7	0.75
AdaBoost Regressor	4.79	6.45	0.31	0.089	6.7	0.67
Orthogonal Matching Pursuit	4.79	6.57	0.29	0.092	6.8	0.72
Huber Regressor	4.87	6.57	0.29	0.092	6.8	0.71
Light Gradient Boosting Machine	5.10	6.66	0.27	0.092	7.0	0.64
Gradient Boosting Regressor	5.14	6.70	0.26	0.091	7.0	0.65
Extra Trees Regressor	5.20	6.80	0.23	0.094	7.2	0.58
Random Forest Regressor	5.37	6.90	0.22	0.095	7.4	0.57
Decision Tree Regressor	5.71	7.55	0.06	0.103	7.9	0.32

Upon further analysis of the K Neighbors Regressor model and its variable removal scenarios, it is clear that the APSIM variables hold the most significance in sugarcane yield predictions. This is evident from the fact that scenarios involving the removal of APSIM variables yielded the worst statistical performance, as seen at the bottom of the Table 2.

Table 2. K neighbors Regressor statistical metrics for 7 different removal variable scenarios, in the test dataset for years 2019 and 2020. Predictions comprehend the data until July, 1 month before the majority of areas harvesting. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Percentage Error (MAPE) and Correlation Coefficient (r).

Model	Variable Removal Scenarios	MAE	RMSE	R^2	RMSLE	MAPE	r
		t ha^{-1}	t ha^{-1}			%	
K Neighbors Regressor	No removal (full dataset) (1)	3.26	4.48	0.67	0.063	4.5	0.83
	VEG_IND (3)	3.51	4.66	0.65	0.065	4.9	0.84
	MET (4)	3.77	5.17	0.56	0.073	5.3	0.81
	MET + VEG_IND (6)	3.79	4.98	0.60	0.070	5.3	0.82
	APSIM (2)	3.94	5.40	0.52	0.075	5.4	0.76
	APSIM + VEG_IND (7)	4.28	6.05	0.40	0.083	5.8	0.65
	APSIM + MET (5)	4.62	6.13	0.39	0.086	6.4	0.76

The variable removal scenarios are (1) Full dataset without variables removal, i.e. includes remote sensing vegetation indices, APSIM variables and meteorological variables; (2) Full dataset except APSIM variables; (3) Full dataset except remote sensing vegetation indices; (4) Full dataset except meteorological variables; (5) Dataset excluding APSIM and meteorological variables; (6) Dataset excluding APSIM and remote sensing vegetation indices; (6) Dataset without meteorological variables and vegetation indices and (7) Dataset excluding APSIM variables and vegetation indices

The results suggest that adding APSIM simulation crop model variables as input features to ML models can decrease sugarcane yield prediction RMSE from 7.7% to as high as 26.9%. Shahhosseini et al. (2021) also found that the APSIM variables demonstrated great correlation and reduced maize yield predictions error (RMSE) from 7% to 20%. The inclusion of APSIM variables as input features in the ML models led to enhanced performance in almost all 16 models developed (see appendix B), thus, this highlights the importance of the APSIM variables to reduce the error in predictions. Everingham et al. (2016), evaluated three sugarcane yield forecast dates using climate indices, weather data and biomass accumulation from APSIM with random forest algorithm. The authors found that the earliest forecast date relied most heavily on the APSIM biomass index and El Niño–Southern Oscillation (ENSO) indices, in the year before harvest, which led to a RMSE of 8 t ha^{-1} and R^2 of 0.67. Conversely, the Southern Oscillation Index (SOI) in October was the most important variable for the forecast dates that were performed later, on the year of harvesting. Pagani et al. (2017) found similar results when studying multiple linear regressions to forecast sugarcane yield. The authors reported that the model with only agro-climatic indicators explained 38% of inter-annual yield variability during growth phase of January to April, and explained 73% during the second half of the harvesting period of September and October. When sugarcane DSSAT-CANEGRO model outputs were added as inputs to the regressor model, the variability explained increased to 63% in the early forecast date, and to 90% after the second half of the harvesting period. The results of this study, along with the literature on hybrid modeling, highlight the effectiveness of this approach in forecasting yield.

When excluding only the APSIM variables, results demonstrated the MAE and RMSE increases more than the exclusion of meteorological or vegetation indices data. This highlights that the simulation of APSIM, that uses the weather variables to simulate crop growth, are better than the standalone use of weather variables due to its simulation of the underlying biophysical processes. The removal of other variables such as vegetation indices and weather variables did not impact as much as the APSIM variables in model performance, which has led to another interesting finding, that the scenario where remote sensing vegetation indices were not included had the least impact on the yield predictions.

For the scenarios with vegetation indices, comparing the scenario 6 to 4, the addition of remote sensing data led to an opposite effect, an increase in RMSE from 4.98 t ha^{-1} to 5.17 t ha^{-1} (-3.8%). Comparing the scenarios 7 and 2, the incorporation of vegetation indices has led to a RMSE decrease from 6.05 to 5.40 (10.7%). While the scenarios that compares only the effect of vegetation indices in the full dataset, scenario 3 compared with 1, the RMSE has a small decrease of 4.66 to

4.48 (3.8%). Thus, depending on the data availability, remote sensing data are expected to have ambiguous effects, increasing or decreasing the error of RMSE in the range of -3.8% to 10.7%.

Similar results were also found by Feng et al. (2020), when studying the hybrid modeling approach of combining APSIM simulated variables, NDVI and meteorological variables with machine learning, to forecast wheat yield. The authors concluded that NDVI was relatively unimportant, only being selected in the top feature importance at late wheat growth stages, and even then, NDVI was not highly ranked as an important feature for wheat yield forecasts. A possible explanation for these results can be the spatial resolution of the vegetation indices. The authors, and the present study, utilized the MODIS satellite, which has a resolution nearly 500 m. In this sense, the sugarcane areas can have vegetation indices pixels without sugarcane or with a mixture of crops, or even sugarcane but at distinct growth stages which could influence the NDVI values and hence reduce its importance in sugarcane yield forecast.

Despite that, the partial exclusion of variables highlights some interesting results, generally speaking, the importance of remote sensing data in yield forecasting studies varies depending on the methodology employed. It was observed that when studies rely solely on remote sensing data, these variables are typically the most critical for yield forecasting. However, in studies that use a combination of remote sensing data and process-based crop models, the variables generated by the simulation models tend to be more important for accurate yield forecasting.

5.4. Results of early forecast

To further evaluate the forecast results of K-Neighbors regressor, scatter plots of ground truth yield against the predicted yield for the years of 2019 and 2020 are shown in Figures 8, 9 and 10. As can be seen the figures depicts the scatter plots for the sugarcane yield forecasting before the harvest season started, months from January to March, the beginning of the harvesting season, from April (Figure 8) to May (Figure 9), and during the mid-harvest season, from June to August (Figure 9), and from the end harvest season from September to November (Figure 10). The sugarcane yield scatter plots indicate that the K neighbors regressor can successfully forecast yield months prior to harvesting. As can be seen that the model effectively captured the temporal changes in the observed sugarcane yield, and the precision of the predictions typically increased as sugarcane growth and time progressed.

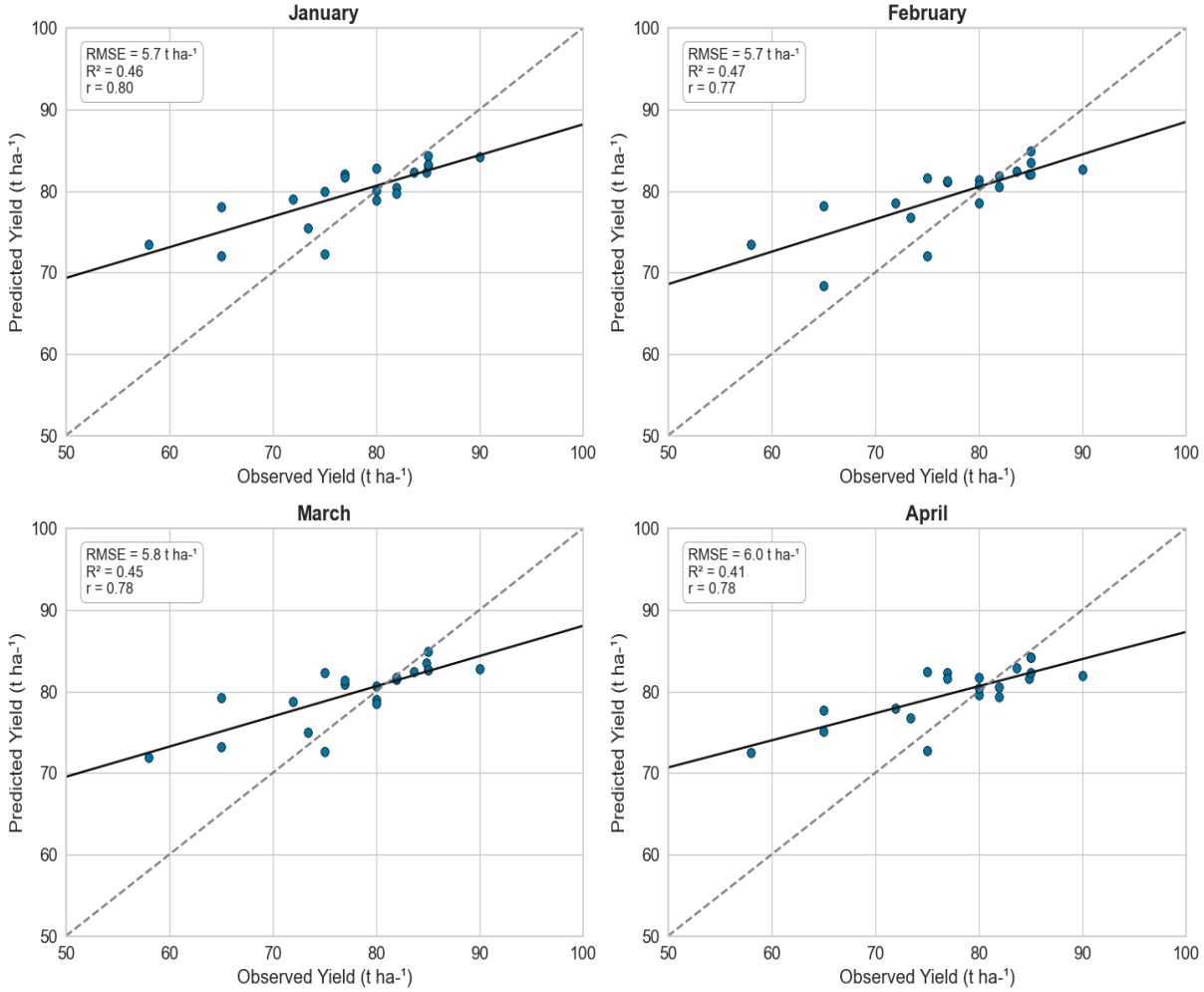

Figure 8. Comparison of observed and model forecasted sugarcane yields at 4 forecasting dates (January, February, March and April) from the K Neighbors Regressor model. The forecasts include input data up to the last day of each month and the data comprehend the test dataset for years 2019 and 2020. Statistical metrics shown are Root Mean Squared Error (RMSE), Coefficient of determination (R^2) and Correlation Coefficient (r). Dashed lines represent the identity lines and solid lines represent the data trendline.

At the first forecasting month events, before the harvesting season begins, January, February and March, showed minor differences in sugarcane yield forecasts, with a RMSE nearly 5.7 t ha^{-1}, a R^2 value under 0.5 and a correlation ranging between 0.77 and 0.8. At the start of growing season in April (Figure 8), the yield forecast slightly underperformed the previous months, as the input variables of remote sensing, APSIM variables and meteorological data were updated. These findings corroborate that sugarcane yield can be forecasted with a degree of accuracy and error as low as 6 t ha^{-1} in the beginning of year, months ahead the starting of harvesting season. Moreover, the accuracy achieved using a 3-month look-ahead to the beginning of harvest season, can have important implications for crop management decisions. Accurate yield predictions at different stages of crop growth can enable decision-makers to modify crop management practices and ensure that yield is optimized throughout the growth cycle.

In the May month forecast event, the predictions accuracy increased significantly, reducing the error to under 5 t ha^{-1}, and achieving a R^2 of 0.63, and the best r value of 0.84. The months shown in Figure 9, May, June and July, are the best sugarcane yield forecasting dates. Besides that, it is in July, 1 month before the majority of areas were identified to have been harvested, was the best month to forecast yield. The performance of the K Neighbors Regressor with the full dataset showed an error of less than 4.5 t ha^{-1}, the observed yield was captured by the predictors by a R^2 of 0.67, and the model presented a good correlation with the observed data, with a correlation of 0.83. As compared, in the section 5.2, the statistical metrics of the hybrid forecasting approach outperforms the yield predictions for sugarcane in the literature, that used one or another forecasting method individually.

Everingham et al. (2016) found that the performance of their sugarcane yield forecasting approach explained sugarcane yield in 67%, on the first early forecast, and increased to 72% and 79%. Also, it decreased the RMSE from 8.0 to 6.3 t ha^{-1}, as the forecast date became later in the season. This study's data were collected from a sugarcane mill located in the Tully region of Australia, which receives an annual average rainfall of 4000 mm and so it is a completely distinct sugarcane growing environment from the present site study and the other cited literature. In another study in Australia, Han, Bishop and Filippi (2022) analyzed harvest data from two sugarcane properties in Queensland, a region with average annual rainfall of 900 mm, from 2007 to 2018. The authors used Random Forest models to forecast yields at two management points, an early-season forecast, in December of previous year of harvest, and a late-season forecast, in June at the harvesting year. Using leave-one-season-out cross-validation, the models achieved a RMSE of 32.1 t ha^{-1} for early-season forecasts and 30.9 t ha^{-1} of RMSE for late-season forecasts.

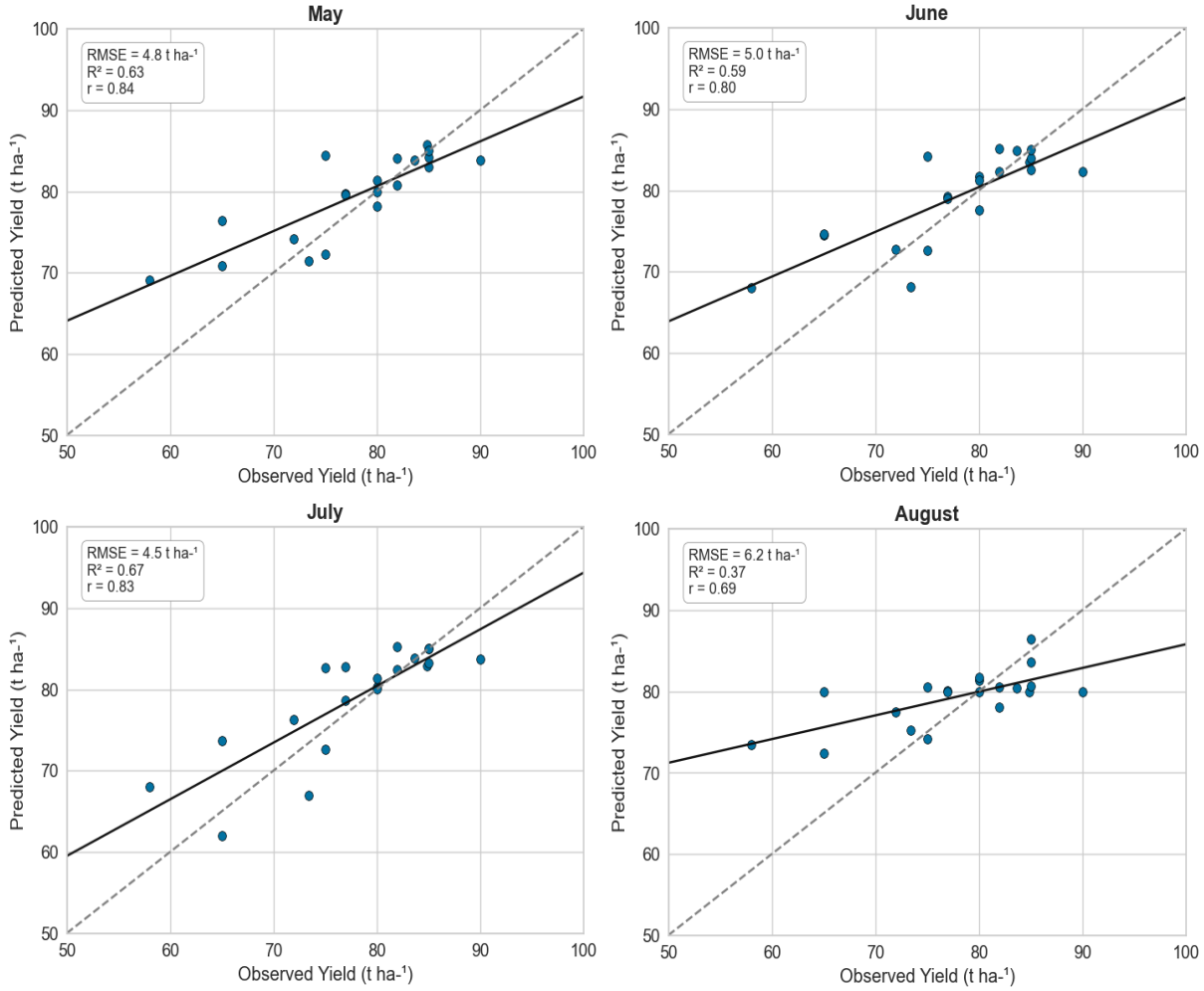

Figure 9. Comparison of observed and model forecasted sugarcane yields at 4 forecasting dates (May, June, July and August) from the K Neighbors Regressor model. The forecasts include input data up to the last day of each month and the data comprehend the test dataset for years 2019 and 2020. Statistical metrics shown are Root Mean Squared Error (RMSE), Coefficient of determination (R^2) and Correlation Coefficient (r). Dashed lines represent the identity lines and solid lines represent the data trendline.

Besides the differences between the environments related to rainfall, the latter study did not use any process-based crop model to generate variables for input in the Random Forest models. In another study, the researchers evaluated machine learning algorithms and various predictor variables derived from satellite imagery at a field level block resolution. The authors achieved a R^2 of 0.39 and RMSE of 19.1 t ha^{-1} for sugarcane yield in January, which improved to 0.51 and 16.0 t ha^{-1} by May (SHENDRYK; DAVY; THORBURN, 2021). This study was conducted in a broader region that comprehends the Wet Tropics of Australia in northeast Queensland. The area has a tropical climate, and the study location ranges from Cairns with an average annual rainfall of 1988 mm to Tully with more than 4000 mm. The authors also concluded that their modeling approach based on combined predictor variables systematically overestimated high values and underestimated low values of sugarcane yield, which was the opposite relationship observed in the present study results. In overall, the results showed that the low observed values were

overestimated by the model and the high observed yields were underestimated, as can be seen in Figures 8, 9 and 10. However, in the July month forecast, the results tended to reduce this systematic bias (Figure 9).

In Brazil, the same pattern found in the present study was identified by Fernandes, Ebecken and Esquerdo (2017), as their model was overestimating low values and underestimating high values of sugarcane yield. The authors also concluded that the yield forecasts improve as the growing season progresses, they reported the 3-months before harvesting forecast with a RMSE of 7.2 and R^2 of 0.38. While at the end of harvest season, the results improved, leading to a R^2 of 0.61 and a RMSE of 5.7 t ha^{-1}. The authors used NDVI time series from the MODIS sensor and an ensemble model of artificial neural networks (ANNs) with sugarcane yield data from 60 municipalities in São Paulo State between 2003 and 2012. The author did not use a process-based crop model and their results are similar to the present study findings without the APSIM variables. Pagani et al. (2017) developed a forecasting system for sugarcane yield using multiple linear regressions that relate agro-climatic indicators and outputs of the sugarcane model DSSAT-CANEGRO to historical yield records at municipalities level in São Paulo. The model was created using the official stalk yields from the period of 2000 to 2013, and they developed forecasting systems in different stages of the sugarcane cycle. The agro-climatic indicators and DSSAT-CANEGRO outputs combined explained the variability led to a R^2 of 0.63 for the boom growth phase and 0.9 after mid harvesting, with the best performances achieved while approaching the end of the harvesting window (i.e. at the beginning of October), with a R^2 of 0.93. The general trend is the same reported in the present study, the agreement increases between model and observed yield, and the prediction errors reduces, as the growing season progresses. However, the authors reported in their study the coefficient of determination (R^2) of the cross validated data, which represents how well a model generalizes to a subset of training data in various different iterations and testing it on the remaining data, but still on training data. Thus, the R^2 value of the test data is typically considered the more important measure, as it reflects the model's ability to make accurate predictions on a complete unseen new data, which was considered in this study and on the other previous cited studies.

The months of August (Figure 9) and September (Figure 10) had the highest sugarcane areas harvested, as seen in the remote sensing time-series where the NDVI decreased significantly (Figure 2). This led to an increase in RMSE, to values above 6 t ha^{-1}, and an R^2 ranging from 0.35 to 0.37. In these months, the predictions underperformed the previous 3 months, and also, these months generated sugarcane yield forecasts worse than the pre-harvesting months from January to March. A reason for this could be the data noise of the remote sensing data, as the data is updating

for each end of the month to compose the input data to the ML model, the changes in remote sensing data and weather data can add data noise which can lead to model uncertainties.

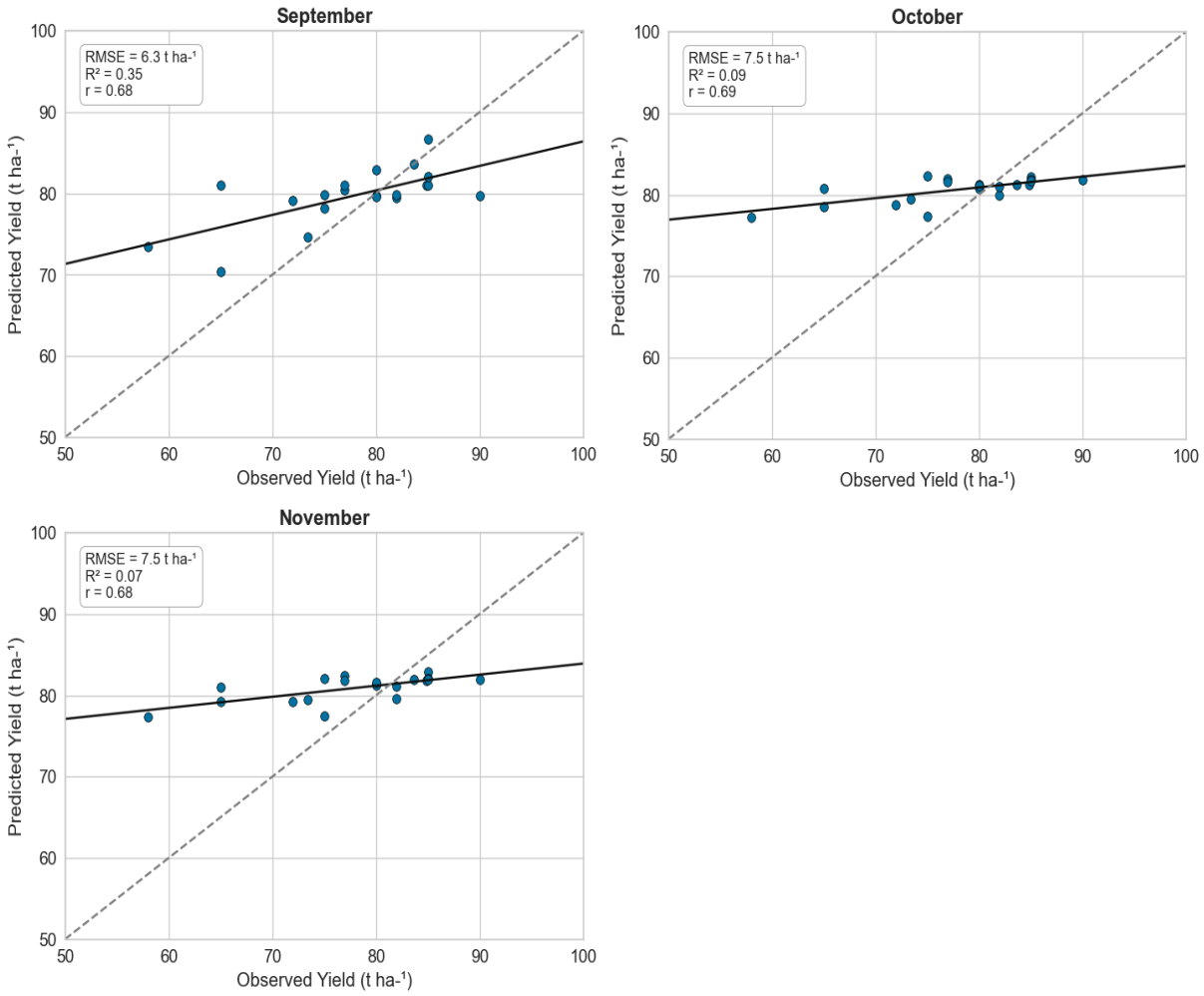

Figure 10. Comparison of observed and model forecasted sugarcane yields at 3 forecasting dates (September, October and November) from the K Neighbors Regressor model. The forecasts include input data up to the last day of each month and the data comprehend the test dataset for years 2019 and 2020. Statistical metrics shown are Root Mean Squared Error (RMSE), Coefficient of determination (R^2) and Correlation Coefficient (r). Dashed lines represent the identity lines and solid lines represent the data trendline.

The same conclusions can be applied to the end-harvesting season of October and November, which presented the worst statistical metrics for sugarcane yield predictions, with a poor R^2 under 0.10, and the sugarcane pixels that are harvested late in the season introduced more uncertainties to the mean yield estimates (Figure 10). Again, the resolution of the MODIS, utilized in this study, can be hypothesized to explain the uncertainties introduced in yield forecasting in the end of harvesting season, as in São Paulo state, a lot of fields are partially harvested and the total area harvest can be at the end of the season, which could have dropped the vegetation indices and compromised the ML model performance. However, the yield forecasting before the harvesting

date, which was one of the objectives of this study, presented greater accuracy than the standalone use of each individual forecast approaches, as denoted by the statistical metrics. In this way, the hybrid approach aligns with the current trend of adopting new technologies in agriculture, as suggested by Keating and Thorburn (2018). This study represents the next step in crop modeling by integrating advanced statistical and mechanistic models in crop-environment research. This integration takes advantage of the increasing availability of data related to farming, climate, and remote sensing, which presents an opportunity to enhance regional yield forecasting and leverage the use of large amounts of agriculture data available. Additionally, it highlights the potential value of incorporating input features from other sources to improve yield forecasts and to apply this methodology for diverse agricultural crops and environments.

However, some aspects need attention when different technologies are blended, as they introduce new concepts, limitations and advantages of each technologies must be well known and balanced. In the study of hybrid modeling for maize proposed by Shahhosseini et al. (2021), they used the APSIM simulations as inputs to ML model, and this simulations carried the full weather of each test year. They recognized that in real world applications, the weather would be unknown. In the current study, the proposed method does not allow unknown future data to be present in the training process due to the APSIM model variables, the weather variables and the remote sensed vegetation indices being limited to the available data at the corresponding end of month (i.e. data in January just have data generated until last day of January), which in theory avoids data leakage to the model forecast and test dataset. Data leakage is when unknown future data are present in training dataset, and this can result in a model highly accurate in the training dataset but it performs poorly in the test dataset, which was not the case in the current study. In this sense, to achieve the best results, it's crucial to take into account the limitations of the methodologies utilized. For instance, the various machine learning algorithms will provide different learning process of data and capture the variability of input variables differently, the impact of different remote sensing resolutions and how to address those limitations, as well as to delineate input variables that account for weather and soil impacts on crop growth and yield. As these blending or hybrid approaches involve multidisciplinary fields of study.

5.5. Principal component analysis and variables contribution

The principal component analysis (PCA) was applied in the 158 input variables to reduce its dimensionality and avoid autocorrelation between variables. However, when applying PCA to a dataset, the most significant challenge is selecting the ideal number of principal components. This

process can be defined as a common hyperparameter tuning of the PCA algorithm, where the optimal hyperparameter value is chosen. After implementing an iterative process to determine the percentage of variance to keep in the PCA, a proportion of 96% was discovered to be the best hyperparameter value that resulted in the lowest MAPE error for sugarcane yield predictions Table 3.

Table 3. K neighbors Regressor statistical metrics for different principal components percentage of variance, in the test dataset for years 2019 and 2020. Predictions comprehend the data until end of July, 1 month before the majority of areas harvesting. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Percentage Error (MAPE) and Correlation Coefficient (r).

PCA percentage of variance threshold	MAE	RMSE	R^2	RMSLE	MAPE	r
	t ha^{-1}	t ha^{-1}			%	
96%	3.26	4.48	0.67	0.06	4.56	0.83
92%	3.26	4.66	0.65	0.07	4.57	0.82
90%	3.35	4.72	0.64	0.07	4.64	0.82
78%	3.47	4.84	0.62	0.07	4.82	0.80
80%	3.47	4.84	0.62	0.07	4.82	0.80
82%	3.72	5.07	0.58	0.07	5.13	0.78
74%	3.83	5.16	0.57	0.07	5.38	0.77
76%	3.83	5.16	0.57	0.07	5.38	0.77
86%	3.55	5.17	0.56	0.07	5.04	0.83
84%	3.63	5.21	0.56	0.07	5.14	0.82
88%	3.69	5.29	0.54	0.07	5.15	0.78
94%	3.50	5.30	0.54	0.07	4.94	0.79
98%	3.92	5.54	0.50	0.08	5.52	0.80
64%	4.71	6.70	0.27	0.09	6.73	0.81
62%	4.71	6.70	0.27	0.09	6.73	0.81
66%	4.71	6.70	0.27	0.09	6.73	0.81
60%	4.45	6.77	0.25	0.09	6.30	0.59

The PCA percentage of variance of 96%, that performed best in the K neighbors regressor reduced the dataset to 19 components. Which means that 19 components were used as inputs to the regressors. Figure 11 shows the explained variance percentage of individual components and the percentage of total variance captured by all principal components. The first three principal components were responsible for almost 55% of total variance captured in the dataset. The last five components represented less than 5% of total variance (Figure 11), however in the iteration process of finding the best PCA threshold for the number of components (Table 3), including these 5 principal components reduced the regression errors of sugarcane yield forecast of the K Neighbors Regressor.

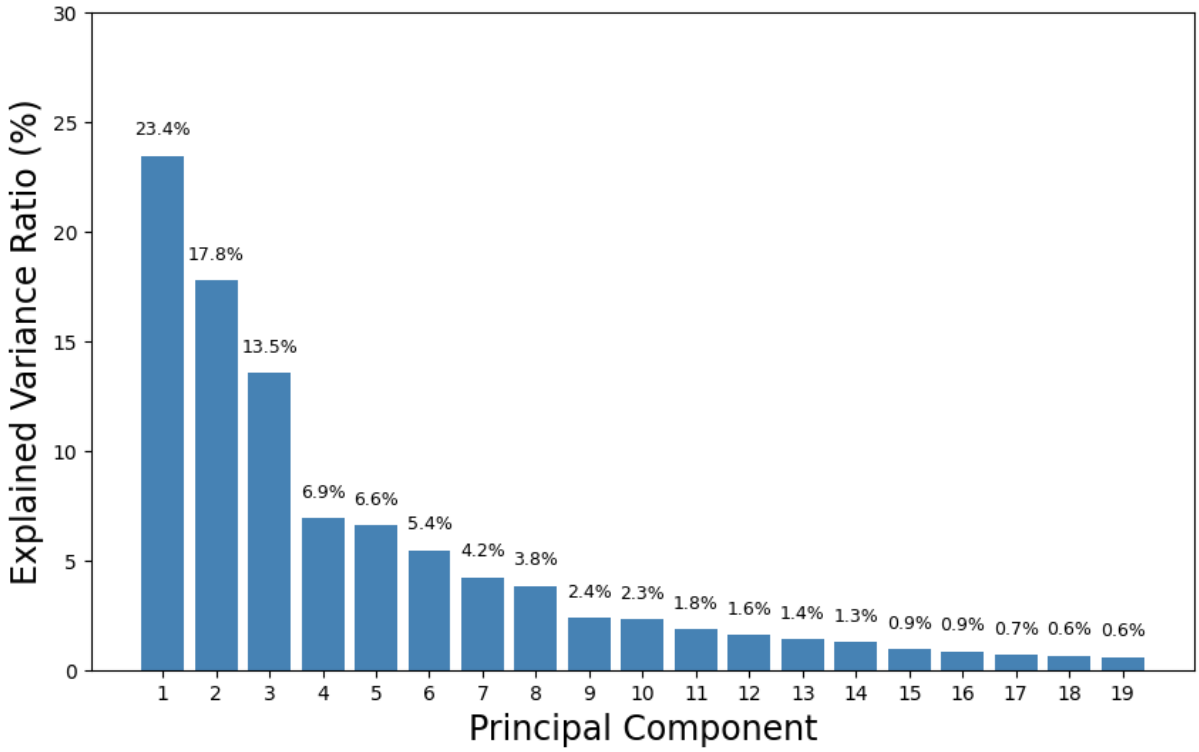

Figure 11. The percentage of total variance captured by the principal components. The input variables contained 158 features and the PCA yielded 19 principal components.

The use of all variable (full dataset) in principal components analysis resulted in 19 principal components. These components represent linear combinations of the original variables that capture the maximum amount of variation in the data, in this way, the main drawback of this technique is the interpretability of principal components, since that they are derived from the original variables and do not have a direct and intuitive meaning. Each principal component is a combination of the original variables, making it difficult to interpret them in terms of the original data. Nevertheless, to overcome the interpretability limitation of the principal components, it was extracted the scores of the features on the first three principal components and identified the top 10 features with the highest absolute score for each component (Figure 12).

The variables that were more important in the first component (PC1), which represents 23.4% (Figure 11), were mean latitude and maximum air temperature variability by municipality that are more related to weather variables and the APSIM simulated variables related to soil temperature and soil water for different soil layers (Figure 12). In the second principal component (PC2), the variable that had the highest contribution was the mean longitude of the municipality, which had a positive score. The remaining nine variables had negative scores, and among them, the variables related to the standard deviation of aggregated variables in the municipality had a dominant influence (std suffix). These variables included simulated APSIM plant state variables such as rootgreenwt_std, senescedwt_std, lai_std, and height_std. Additionally, the sum of degree

days made a significant contribution in PC2, appearing in both the mean variable and standard deviation variables.

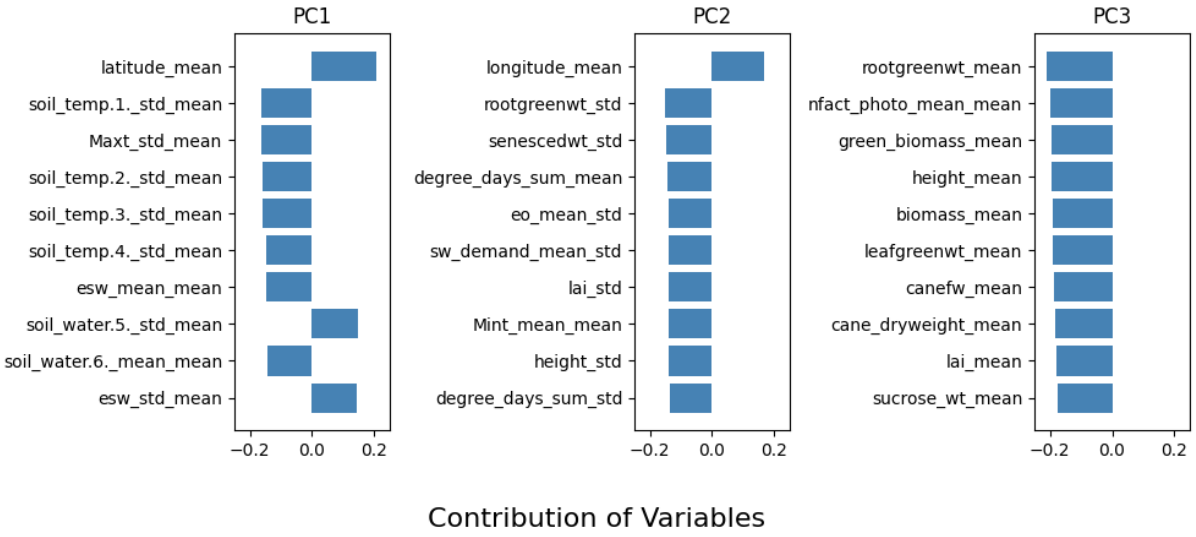

Figure 12. Variables contribution to the sugarcane forecast in first three principal components responsible for almost 55% total variance captured by the principal components.

The top 10 variables in PC3 were all related to the APSIM simulated variables representing the plant state variables, such as root and aboveground biomass that includes fresh and dry stalks, green leaves weight, and sucrose. As can be noted, the appearance of this variables in the same principal component, indicates that they had a high correlation with each other. The std suffix in variables represent the variability (standard deviation) of the variable in the municipality, therefore, higher variability can negatively affect the performance of the regression models in learning the patterns of the data. A more comprehensive description of variables can be seen in the Appendix section (Appendix A). In general, the decomposition of principal components results highlights the importance of geolocation and APSIM variables, as they are one of the most influential variables in the first three principal components.

6. CONCLUSION AND REMARKS

In this study, it was successfully developed a hybrid model forecasting system for sugarcane yield by combining remote sensing weather data, process-based crop model outputs, and vegetation indices information. Also, it was tested the input data into 16 different machine learning regression models. The regional forecasted sugarcane yields showed low bias and good agreement with observed yield, providing one to two-month lead times, and a high degree of accuracy before the harvesting season. This approach can be implemented to monitor and forecast season harvest yield in the main sugarcane producer region and potentially in other similar dryland cropping systems around the world, due to its incorporation of process-based modeling approach. As the advances continues in internet of things (IoT), big data, cloud computing, remote sensing technology, and precision agriculture, it is expected that hybrid approaches can improve crop yield forecasts. This methodology can become increasingly important in providing information for regional planning of sugarcane supply and mitigating the detrimental effects of climate change on global food supply.

Moreover, in this study it was investigated the impact of different source variables on forecasted yield, leading to a better understanding of sugarcane yield variability. Overall, the APSIM simulated variables showed the most positive impact on sugarcane yield forecast hybrid approach. While the remote sensing vegetation indices data presented the least contribution in yield forecasts. However, it may be worthwhile to explore finer spatial resolution in the future to determine whether it further benefits the ML algorithms and minimizes uncertainties.

This approach focused on sugarcane in high-producing cities, but its hybrid modeling framework that combines remote sensing, process-based modeling, and machine learning makes it applicable to sugarcane in any region, indicating its readiness for scalability. While gathering more data could enhance the methodology's accuracy, this remains a potential extension for future work, which could also encompass additional crops and regions. Overall, this study aims to demonstrate the potential of hybrid models for yield prediction and highlights the strength of the approach and its results.

One major limitation of this study was the restricted extent of the study area, which could be addressed by incorporating data from more years, cities, and other producing regions to add more variability and stress to the hybrid model approach. However, gathering more weather data is challenging due to the time required for processing and quality checking for use in crop models. Furthermore, when introducing more regions and years, cloud computing and data storage become crucial, depending on the temporal and spatial scales of process-based simulations. In this study, it

was used data from remote sensing and public databases, and the high spatial simulations could be optimized by reducing and identifying similar data points to simulate only highly different points and extrapolating the simulations with points that had minor differences between input variables, thus reducing processing time in all workflow steps. Future investigations could include applying this study in the field level of sugarcane mills in Brazil, which could improve the accuracy of the yield forecasts in management and operational applications of mills and producers. This hybrid approach developed is designed for all levels of spatial scales, and introducing more data variability could further test the hybrid model approach.

Another suggestion could be the introduction of weather forecasts in the APSIM simulation model, which could improve the sugarcane yield forecasts before harvesting season. APSIM variables could produce simulations of crop to various weather projections, and the forecasts could incorporate weather uncertainties to forecast sugarcane yields as probabilities to fulfill a specific scenario, such as favorable, normal or unfavorable climatic year or it could include a yield forecasting interval.

REFERENCES

AKBARIAN, S. et al. Sugarcane yields prediction at the row level using a novel cross-validation approach to multi-year multispectral images. **Computers and Electronics in Agriculture**, v. 198, p. 107024, 1 jul. 2022.

ALAM, M. et al. Real-Time Machine-Learning Based Crop/Weed Detection and Classification for Variable-Rate Spraying in Precision Agriculture. **2020 7th International Conference on Electrical and Electronics Engineering, ICEEE 2020**, p. 273–280, 1 abr. 2020.

ALI, A. M. et al. Crop Yield Prediction Using Multi Sensors Remote Sensing (Review Article). **The Egyptian Journal of Remote Sensing and Space Science**, v. 25, n. 3, p. 711–716, 1 dez. 2022.

ALI, I. et al. Review of Machine Learning Approaches for Biomass and Soil Moisture Retrievals from Remote Sensing Data. **Remote Sensing**, v. 7, n. 12, p. 16398–16421, 2015.

ALI, M. **PyCaret: An open source, low-code machine learning library in Python**. Disponível em: <https://pycaret.org/>. Acesso em: 26 abr. 2023.

ALMEIDA, J. S. Predictive non-linear modeling of complex data by artificial neural networks. **Current Opinion in Biotechnology**, v. 13, n. 1, p. 72–76, 1 fev. 2002.

ARAÚJO, C. S. P. de et al. Evaluation of air temperature estimated by ERA5-Land reanalysis using surface data in Pernambuco, Brazil. **Environmental Monitoring and Assessment**, v. 194, n. 5, p. 1–13, 1 maio 2022. Disponível em: <https://link.springer.com/article/10.1007/s10661-022-10047-2>. Acesso em: 17 abr. 2023.

ASSENG, S. et al. Simulation Modeling: Applications in Cropping Systems. **Encyclopedia of Agriculture and Food Systems**, p. 102–112, 1 jan. 2014.

ATZBERGER, C. Advances in Remote Sensing of Agriculture: Context Description, Existing Operational Monitoring Systems and Major Information Needs. **Remote Sensing 2013, Vol. 5, Pages 949-981**, v. 5, n. 2, p. 949–981, 22 fev. 2013. Disponível em: <https://www.mdpi.com/2072-4292/5/2/949/htm>. Acesso em: 16 abr. 2023.

AZZARI, G.; JAIN, M.; LOBELL, D. B. Towards fine resolution global maps of crop yields: Testing multiple methods and satellites in three countries. **Remote Sensing of Environment**, v. 202, p. 129–141, 1 dez. 2017.

BAKER, J. C. A. et al. An Assessment of Land–Atmosphere Interactions over South America Using Satellites, Reanalysis, and Two Global Climate Models. **Journal of Hydrometeorology**, v. 22, n. 4, p. 905–922, 1 abr. 2021. Disponível em: <https://journals.ametsoc.org/view/journals/hydr/22/4/JHM-D-20-0132.1.xml>. Acesso em: 17 abr. 2023.

BANNARI, A. et al. A review of vegetation indices. **Remote Sensing Reviews**, v. 13, n. 1–2, p. 95–120, 1995.

BANNAYAN, M.; CROUT, N. M. J.; HOOGENBOOM, G. Application of the CERES-Wheat model for within-season prediction of winter wheat yield in the United Kingdom. **Agronomy Journal**, v. 95, n. 1, p. 114–125, 2003.

BASNAYAKE, J. et al. Sugarcane for water-limited environments. Genetic variation in cane yield and sugar content in response to water stress. **Journal of Experimental Botany**, v. 63, n. 16, p. 6023–6033, 1 out. 2012.

BASSO, B. et al. Spatial validation of crop models for precision agriculture. **Agricultural Systems**, v. 68, n. 2, p. 97–112, 1 maio 2001.

BASSO, B.; CAMMARANO, D.; CARFAGNA, E. Review of Crop Yield Forecasting Methods and Early Warning Systems. In: Proceedings of the First Meeting of the Scientific Advisory Committee of the Global Strategy to Improve Agricultural and Rural Statistics, Rome, Italy. **Anais**... Rome, Italy: FAO Headquarters, 2013.

BASSO, B.; LIU, L. Seasonal crop yield forecast: Methods, applications, and accuracies. **Advances in Agronomy**, v. 154, p. 201–255, 1 jan. 2019.

BATOOL, D. et al. A Hybrid Approach to Tea Crop Yield Prediction Using Simulation Models and Machine Learning. **Plants**, v. 11, n. 15, p. 1925, 25 jul. 2022. Disponível em: <https://www.mdpi.com/2223-7747/11/15/1925/htm>. Acesso em: 2 maio. 2023.

BECKER-RESHEF, I. et al. A generalized regression-based model for forecasting winter wheat yields in Kansas and Ukraine using MODIS data. **Remote Sensing of Environment**, v. 114, n. 6, p. 1312–1323, 15 jun. 2010.

BEHERA, S. K.; RATH, A. K.; SETHY, P. K. Maturity status classification of papaya fruits based on machine learning and transfer learning approach. **Information Processing in Agriculture**, v. 8, n. 2, p. 244–250, 1 jun. 2021.

BEHMANN, J. et al. A review of advanced machine learning methods for the detection of biotic stress in precision crop protection. **Precision Agriculture**, v. 16, n. 3, p. 239–260, 2015.

BELLOCCHI, G. et al. Validation of biophysical models: issues and methodologies. A review. **Agronomy for Sustainable Development 2009 30:1**, v. 30, n. 1, p. 109–130, jan. 2010. Disponível em: <https://link.springer.com/article/10.1051/agro/2009001>. Acesso em: 8 abr. 2023.

BENOS, L. et al. Machine Learning in Agriculture: A Comprehensive Updated Review. **Sensors 2021, Vol. 21, Page 3758**, v. 21, n. 11, p. 3758, 28 maio 2021. Disponível em: <https://www.mdpi.com/1424-8220/21/11/3758/htm>. Acesso em: 11 abr. 2023.

BERTALAN, L. et al. UAV-based multispectral and thermal cameras to predict soil water content – A machine learning approach. **Computers and Electronics in Agriculture**, v. 200, p. 107262, 1 set. 2022.

BHAGOJI, A. N. et al. Enhancing robustness of machine learning systems via data transformations. In: 2018 52nd Annual Conference on Information Sciences and Systems, CISS 2018, Princeton, NJ, USA. **Anais**... Princeton, NJ, USA: 21 maio 2018.

BHATTARAI, G. P.; SCHMID, R. B.; MCCORNACK, B. P. Remote Sensing Data to Detect Hessian Fly Infestation in Commercial Wheat Fields. **Scientific Reports**, v. 9, n. 1, p. 1–8, 1 dez. 2019.

BOLTON, D. K.; FRIEDL, M. A. Forecasting crop yield using remotely sensed vegetation indices and crop phenology metrics. **Agricultural and Forest Meteorology**, v. 173, p. 74–84, 15 maio 2013. Disponível em: <https://www.researchgate.net/publication/236851133_Forecasting_crop_yield_using_remotely_sensed_vegetation_indices_and_crop_phenology_metrics>. Acesso em: 6 abr. 2023.

BOUMAN, B. A. Crop modelling and remote sensing for yield production. **Netherlands Journal of Agricultural Science**, v. 43, n. 2, p. 143–161, 1995. Disponível em: <https://research.wur.nl/en/publications/crop-modelling-and-remote-sensing-for-yield-production>. Acesso em: 8 abr. 2023.

BRÉDA, N. J. J. Leaf Area Index. **Encyclopedia of Ecology, Five-Volume Set**, p. 2148–2154, 1 jan. 2008.

BRISSON, N. et al. STICS: a generic model for the simulation of crops and their water and nitrogen balances. I. Theory and parameterization applied to wheat and corn. **Agronomie**, v. 18, n. 5–6, p. 311–346, 1998. Disponível em: <http://dx.doi.org/10.1051/agro:19980501>. Acesso em: 15 abr. 2023.

BRISSON, N. et al. An overview of the crop model stics. **European Journal of Agronomy**, v. 18, n. 3–4, p. 309–332, 1 jan. 2003.

BROWN, J. et al. Evaluation of high-resolution meteorological global data products using flux tower observations across Brazil. **EGU General Assembly**, 2021. Disponível em: <https://ui.adsabs.harvard.edu/abs/2021EGUGA..2315387B/abstract>. Acesso em: 17 abr. 2023.

BROWN, J. N. et al. Seasonal climate forecasts provide more definitive and accurate crop yield predictions. **Agricultural and Forest Meteorology**, v. 260–261, p. 247–254, 15 out. 2018.

BRUNI, V.; CARDINALI, M. L.; VITULANO, D. A Short Review on Minimum Description Length: An Application to Dimension Reduction in PCA. **Entropy**, v. 24, n. 2, p. 269, 13 fev. 2022. Disponível em: <https://www.mdpi.com/1099-4300/24/2/269/htm>. Acesso em: 26 abr. 2023.

CASADEBAIG, P.; DEBAEKE, P.; WALLACH, D. A new approach to crop model calibration: Phenotyping plus post-processing. **Crop Science**, v. 60, n. 2, p. 709–720, 1 mar. 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/csc2.20016>. Acesso em: 2 maio. 2023.

CHAKRAVARTY, S.; DEMIRHAN, H.; BASER, F. Fuzzy regression functions with a noise cluster and the impact of outliers on mainstream machine learning methods in the regression setting. **Applied Soft Computing**, v. 96, p. 106535, 1 nov. 2020.

CHALONER, T. M.; GURR, S. J.; BEBBER, D. P. Plant pathogen infection risk tracks global crop yields under climate change. **Nature Climate Change 2021 11:8**, v. 11, n. 8, p. 710–715, 5 ago. 2021. Disponível em: <https://www.nature.com/articles/s41558-021-01104-8>. Acesso em: 30 abr. 2023.

CHAPAGAIN, R. et al. Decomposing crop model uncertainty: A systematic review. **Field Crops Research**, v. 279, p. 108448, 1 abr. 2022.

CHEN, J. et al. A simple method for reconstructing a high-quality NDVI time-series data set based on the Savitzky–Golay filter. **Remote Sensing of Environment**, v. 91, n. 3–4, p. 332–344, 30 jun. 2004.

CHEN, M. et al. A reinforcement learning approach to irrigation decision-making for rice using weather forecasts. **Agricultural Water Management**, v. 250, p. 106838, 1 maio 2021.

CHENG, E. et al. Wheat yield estimation using remote sensing data based on machine learning approaches. **Frontiers in Plant Science**, v. 13, p. 5310, 23 dez. 2022.

CHIPANSHI, A. et al. Evaluation of the Integrated Canadian Crop Yield Forecaster (ICCYF) model for in-season prediction of crop yield across the Canadian agricultural landscape. **Agricultural and Forest Meteorology**, v. 206, p. 137–150, 15 jun. 2015.

CHLINGARYAN, A.; SUKKARIEH, S.; WHELAN, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. **Computers and Electronics in Agriculture**, v. 151, p. 61–69, 1 ago. 2018.

CHOI, R. Y. et al. Introduction to Machine Learning, Neural Networks, and Deep Learning. **Translational Vision Science & Technology**, v. 9, n. 2, 2020. Disponível em: </pmc/articles/PMC7347027/>. Acesso em: 11 abr. 2023.

CHU, L. et al. Phenology detection of winter wheat in the Yellow River delta using MODIS NDVI time-series data. **2014 The 3rd International Conference on Agro-Geoinformatics, Agro-Geoinformatics 2014**, 25 set. 2014.

CHUVIECO, E.; MARTÍN, M. P.; PALACIOS, A. Assessment of different spectral indices in the red-near-infrared spectral domain for burned land discrimination. **International Journal of Remote Sensing**, v. 23, n. 23, p. 5103–5110, 10 dez. 2002. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01431160210153129>. Acesso em: 5 abr. 2023.

CORRALES, D. C. et al. A surrogate model based on feature selection techniques and regression learners to improve soybean yield prediction in southern France. **Computers and Electronics in Agriculture**, v. 192, p. 106578, 1 jan. 2022.

COSTA, J. C. et al. VALIDAÇÃO DOS DADOS DE PRECIPITAÇÃO ESTIMADOS PELO CHIRPS PARA O BRASIL. **Revista Brasileira de Climatologia**, v. 24, n. 0, 11 jun. 2019. Disponível em: <https://revistas.ufpr.br/revistaabclima/article/view/60237>. Acesso em: 21 fev. 2022.

COSTA, L. G. et al. Simulação do efeito do manejo da palha e do nitrogênio na produtividade da cana-de-açúcar. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 18, n. 5, p. 469–474, 2014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1415-43662014000500001&lng=pt&tlng=pt>.

CROCI, M. et al. Dynamic Maize Yield Predictions Using Machine Learning on Multi-Source Data. **Remote Sensing**, v. 15, n. 1, p. 100, 1 jan. 2023. Disponível em: <https://www.mdpi.com/2072-4292/15/1/100/htm>. Acesso em: 6 abr. 2023.

DALGLIESH, N. et al. Field Protocol to APSoil characterisations: A Protocol for The Development of Apsoil Parameter Values for Use in APSIM. **CSIRO, Australia**, n. 4, p. 1–24, set. 2016.

DANNER, M. et al. Efficient RTM-based training of machine learning regression algorithms to quantify biophysical & biochemical traits of agricultural crops. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 173, p. 278–296, 1 mar. 2021.

DE WIT, C. T. **Photosynthesis of Leaf CanopiesAgricultural Research Report No. 663**WageningenPUDOC, , 1965. .

DEC, D. et al. Effect of Bulk Density on Hydraulic Properties of Homogenized and Structured Soils. **J. Soil Sc. Plant Nutr**, v. 8, n. 1, p. 1–13, 2008.

DENTE, L. et al. Assimilation of leaf area index derived from ASAR and MERIS data into CERES-Wheat model to map wheat yield. **Remote Sensing of Environment**, v. 112, n. 4, p. 1395–1407, 15 abr. 2008.

DIAS, H. B. et al. New APSIM-Sugar features and parameters required to account for high sugarcane yields in tropical environments. **Field Crops Research**, v. 235, p. 38–53, 1 abr. 2019.

DIAS, H. B. et al. Sugarcane yield future scenarios in Brazil as projected by the APSIM-Sugar model. **Industrial Crops and Products**, v. 171, p. 113918, 1 nov. 2021.

DIAZ-GONZALEZ, F. A. et al. Machine learning and remote sensing techniques applied to estimate soil indicators – Review. **Ecological Indicators**, v. 135, p. 108517, 1 fev. 2022.

DINKU, T. et al. Validation of the CHIRPS satellite rainfall estimates over eastern Africa. **Quarterly Journal of the Royal Meteorological Society**, v. 144, p. 292–312, 1 nov. 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/qj.3244>. Acesso em: 4 abr. 2023.

DLAMINI, N. E.; ZHOU, M. Soils and seasons effect on sugarcane ratoon yield. **Field Crops Research**, v. 284, p. 108588, 1 ago. 2022.

DONATELLI, M. et al. Modelling the impacts of pests and diseases on agricultural systems. **Agricultural Systems**, v. 155, p. 213–224, 1 jul. 2017.

DUARTE, L. et al. QPhenoMetrics: An open source software application to assess vegetation phenology metrics. **Computers and Electronics in Agriculture**, v. 148, p. 82–94, 1 maio 2018.

DUMONT, B. et al. Assessing the potential of an algorithm based on mean climatic data to predict wheat yield. **Precision Agriculture**, v. 15, n. 3, p. 255–272, 1 fev. 2014. Disponível em: <https://link.springer.com/article/10.1007/s11119-014-9346-9>. Acesso em: 8 abr. 2023.

EDALAT, M.; NADERI, R.; EGAN, T. P. Corn nitrogen management using NDVI and SPAD sensor-based data under conventional vs. reduced tillage systems. **https://doi.org/10.1080/01904167.2019.1648686**, v. 42, n. 18, p. 2310–2322, 8 nov. 2019. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01904167.2019.1648686>. Acesso em: 4 abr. 2023.

EL BILALI, A.; TALEB, A.; BROUZIYNE, Y. Groundwater quality forecasting using machine learning algorithms for irrigation purposes. **Agricultural Water Management**, v. 245, p. 106625, 28 fev. 2021.

ELAVARASAN, D. et al. Forecasting yield by integrating agrarian factors and machine learning models: A survey. **Computers and Electronics in Agriculture**, v. 155, p. 257–282, 1 dez. 2018.

ERDLE, K.; MISTELE, B.; SCHMIDHALTER, U. Comparison of active and passive spectral sensors in discriminating biomass parameters and nitrogen status in wheat cultivars. **Field Crops Research**, v. 124, n. 1, p. 74–84, 9 out. 2011.

ERSOZ, E. S.; MARTIN, N. F.; STAPLETON, A. E. On to the next chapter for crop breeding: Convergence with data science. **Crop Science**, v. 60, n. 2, p. 639–655, 1 mar. 2020. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/csc2.20054>. Acesso em: 15 abr. 2023.

EVERINGHAM, Y. et al. Accurate prediction of sugarcane yield using a random forest algorithm. **Agronomy for Sustainable Development**, v. 36, n. 2, p. 1–9, 1 jun. 2016. Disponível em: <https://link.springer.com/article/10.1007/s13593-016-0364-z>. Acesso em: 15 abr. 2023.

EYRE, R. et al. Within-Field Yield Prediction in Cereal Crops Using LiDAR-Derived Topographic Attributes with Geographically Weighted Regression Models. **Remote Sensing 2021, Vol. 13, Page 4152**, v. 13, n. 20, p. 4152, 16 out. 2021. Disponível em: <https://www.mdpi.com/2072-4292/13/20/4152/htm>. Acesso em: 6 abr. 2023.

FANG, H. et al. Corn-yield estimation through assimilation of remotely sensed data into the CSM-CERES-Maize model. **International Journal of Remote Sensing**, v. 29, n. 10, p. 3011–3032, 2008. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01431160701408386>. Acesso em: 8 abr. 2023.

FAO. **Food and Agriculture Organization (FAO)**. Disponível em: <https://www.fao.org/faostat/en/#data/OA>. Acesso em: 16 abr. 2023.

FENG, P. et al. Dynamic wheat yield forecasts are improved by a hybrid approach using a biophysical model and machine learning technique. 2020. Disponível em: <https://doi.org/10.1016/j.agrformet.2020.107922>. Acesso em: 7 abr. 2023.

FENG, S. et al. Crop Type Identification and Mapping Using Machine Learning Algorithms and Sentinel-2 Time Series Data. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 12, n. 9, p. 3295–3306, 1 set. 2019.

FERNANDES, J. L.; EBECKEN, N. F. F.; ESQUERDO, J. C. D. M. Sugarcane yield prediction in Brazil using NDVI time series and neural networks ensemble. **International Journal of Remote Sensing**, v. 38, n. 16, p. 4631–4644, 18 ago. 2017.

FRANKE, J.; MENZ, G. Multi-temporal wheat disease detection by multi-spectral remote sensing. **Precision Agriculture**, v. 8, n. 3, p. 161–172, 24 jun. 2007.

FUNK, C. et al. The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes. **Scientific Data 2015 2:1**, v. 2, n. 1, p. 1–21, 8 dez. 2015. Disponível em: <https://www.nature.com/articles/sdata201566>. Acesso em: 12 fev. 2022.

GAO, B. C. NDWI—A normalized difference water index for remote sensing of vegetation liquid water from space. **Remote Sensing of Environment**, v. 58, n. 3, p. 257–266, 1 dez. 1996.

GARDNER, M. W.; DORLING, S. R. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. **Atmospheric Environment**, v. 34, n. 1, p. 21–34, 1 jan. 2000.

GAUTRON, R. et al. Reinforcement learning for crop management support: Review, prospects and challenges. **Computers and Electronics in Agriculture**, v. 200, p. 107182, 1 set. 2022.

GEIPEL, J. et al. Combined Spectral and Spatial Modeling of Corn Yield Based on Aerial Images and Crop Surface Models Acquired with an Unmanned Aircraft System. **Remote Sensing 2014, Vol. 6, Pages 10335-10355**, v. 6, n. 11, p. 10335–10355, 27 out. 2014. Disponível em: <https://www.mdpi.com/2072-4292/6/11/10335/htm>. Acesso em: 16 abr. 2023.

GHODDUSI, H.; CREAMER, G. G.; RAFIZADEH, N. Machine learning in energy economics and finance: A review. **Energy Economics**, v. 81, p. 709–727, 1 jun. 2019.

GHORBANI, M. A. et al. Application of firefly algorithm-based support vector machines for prediction of field capacity and permanent wilting point. **Soil and Tillage Research**, v. 172, p. 32–38, 1 set. 2017.

GITELSON, A. A.; KAUFMAN, Y. J.; MERZLYAK, M. N. Use of a green channel in remote sensing of global vegetation from EOS-MODIS. **Remote Sensing of Environment**, v. 58, n. 3, p. 289–298, 1 dez. 1996.

GOMES, V. C. F.; QUEIROZ, G. R.; FERREIRA, K. R. An Overview of Platforms for Big Earth Observation Data Management and Analysis. **Remote Sensing** , v. 12, n. 8, p. 1253, 16 abr. 2020. Disponível em: <https://www.mdpi.com/2072-4292/12/8/1253/htm>. Acesso em: 23 abr. 2023.

GONZALEZ-SANCHEZ, A.; FRAUSTO-SOLIS, J.; OJEDA-BUSTAMANTE, W. Predictive ability of machine learning methods for massive crop yield prediction. **Spanish Journal of Agricultural Research**, v. 12, n. 2, p. 313–328, 29 abr. 2014.

GORELICK, N. et al. Google Earth Engine: Planetary-scale geospatial analysis for everyone. **Remote Sensing of Environment**, v. 202, p. 18–27, 1 dez. 2017.

GUNARATHNA, M. H. J. P. et al. Sensitivity Analysis of Plant- and Cultivar-Specific Parameters of APSIM-Sugar Model: Variation between Climates and Management Conditions. **Agronomy**, v. 9, n. 5, p. 242, 14 maio 2019.

GUSSO, A. et al. Spectral Model for Soybean Yield Estimate Using MODIS/EVI Data *. **International Journal of Geosciences**, v. 4, p. 1233–1241, 2013. Disponível em: <http://dx.doi.org/10.4236/ijg.2013.49117>. Acesso em: 4 abr. 2023.

HAJIRAHIMI, Z.; KHASHEI, M. Hybrid structures in time series modeling and forecasting: A review. **Engineering Applications of Artificial Intelligence**, v. 86, p. 83–106, 1 nov. 2019.

HAMMER, R. G.; SENTELHAS, P. C.; MARIANO, J. C. Q. Sugarcane Yield Prediction Through Data Mining and Crop Simulation Models. **Sugar Tech**, v. 22, n. 2, p. 216–225, 1 abr. 2020. Disponível em: <https://link.springer.com/article/10.1007/s12355-019-00776-z>. Acesso em: 19 abr. 2023.

HAN, S. Y.; BISHOP, T.; FILIPPI, P. Data-driven, early-season forecasts of block sugarcane yield for precision agriculture. **Field Crops Research**, v. 276, p. 108360, 1 fev. 2022.

HANSEN, J. W.; JONES, J. W. Scaling-up crop models for climate variability applications. **Agricultural Systems**, v. 65, n. 1, p. 43–72, 1 jul. 2000.

HARDISKY, M. A.; KLEMAS, V.; SMART, R. M. The Influence of Soil Salinity, Growth Form, and Leaf Moisture on-the Spectral Radiance of partina alterniflora Canopies. **Photogrammetric Engineering And Remote Sensing**, v. 1, n. 49, p. 77–83, 1983.

HASEGAWA, T. et al. Extreme climate events increase risk of global food insecurity and adaptation needs. **Nature Food 2021 2:8**, v. 2, n. 8, p. 587–595, 9 ago. 2021. Disponível em: <https://www.nature.com/articles/s43016-021-00335-4>. Acesso em: 16 abr. 2023.

HASSAN, M. A. et al. A rapid monitoring of NDVI across the wheat growth cycle for grain yield prediction using a multi-spectral UAV platform. **Plant Science**, v. 282, p. 95–103, 1 maio 2019.

HATFIELD, J. L. et al. Application of spectral remote sensing for agronomic decisions. **Agronomy Journal**, v. 100, n. 3 SUPPL., p. S-117, maio 2008.

HAWORTH, J.; CHENG, T. Non-parametric regression for space–time forecasting under missing data. **Computers, Environment and Urban Systems**, v. 36, n. 6, p. 538–550, 1 nov. 2012.

HE, D. et al. Data requirement for effective calibration of process-based crop models. **Agricultural and Forest Meteorology**, v. 234–235, p. 136–148, 2017.

HEINEMANN, A. B.; STONE, L. F.; SILVA, S. C. da. **Modelos de simulação do crescimento, desenvolvimento e produtividade na pesquisa agronômica**Santo Antônio de GoiásEmbrapa Arroz e Feijão, , 2010. . Disponível em: <https://ainfo.cnptia.embrapa.br/digital/bitstream/item/34251/1/doc-264.pdf>. Acesso em: 2 maio. 2018.

HENGL, T. et al. Soil nutrient maps of Sub-Saharan Africa: assessment of soil nutrient content at 250 m spatial resolution using machine learning. **Nutrient Cycling in Agroecosystems**, v. 109, n. 1, p. 77–102, 1 set. 2017. Disponível em: <https://link.springer.com/article/10.1007/s10705-017-9870-x>. Acesso em: 13 abr. 2023.

HOOKER, J.; DUVEILLER, G.; CESCATTI, A. A global dataset of air temperature derived from satellite remote sensing and weather stations. **Scientific Data 2018 5:1**, v. 5, n. 1, p. 1–11, 6 nov. 2018. Disponível em: <https://www.nature.com/articles/sdata2018246>. Acesso em: 4 abr. 2023.

HOSSEINI, M. et al. Soybean Yield Forecast Using Dual-Polarimetric C-band Synthetic Aperture Radar. **ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, v. V-3–2022, n. 3, p. 405–410, 17 maio 2022. Disponível em: <https://doi.org/10.5194/isprs-annals-V-3-2022-405-2022>. Acesso em: 6 abr. 2023.

HU, C.; THOMASSON, J. A.; BAGAVATHIANNAN, M. V. A powerful image synthesis and semi-supervised learning pipeline for site-specific weed detection. **Computers and Electronics in Agriculture**, v. 190, p. 106423, 1 nov. 2021.

HU, S. et al. Improvement of sugarcane crop simulation by SWAP-WOFOST model via data assimilation. **Field Crops Research**, v. 232, p. 49–61, 15 fev. 2019.

HUANG, J. et al. Jointly Assimilating MODIS LAI and et Products into the SWAP Model for Winter Wheat Yield Estimation. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 8, n. 8, p. 4060–4071, 1 ago. 2015.

HUETE, A. R. A soil-adjusted vegetation index (SAVI). **Remote Sensing of Environment**, v. 25, n. 3, p. 295–309, 1 ago. 1988.

HUETE, A. R. et al. Modis Vegetation Index (MOD 13) Algorithm Theoretical Basis Document . **The University of Arizona - Vegetation Index and Phenology Lab**, v. Version 3.1, abr. 1999. Disponível em: <http://vip.arizona.edu>. Acesso em: 5 abr. 2023.

HUNTINGFORD, C. et al. Machine learning and artificial intelligence to aid climate change research and preparedness. **Environmental Research Letters**, v. 14, n. 12, p. 124007, 22 nov. 2019. Disponível em: <https://iopscience.iop.org/article/10.1088/1748-9326/ab4e55>. Acesso em: 11 abr. 2023.

INDORIA, A. K.; SHARMA, K. L.; REDDY, K. S. Hydraulic properties of soil under warming climate. In: PRASAD, M. N. V.; PIETRZYKOWSKI, M. (Ed.). **Climate Change and Soil Interactions**. Amsterdam: Elsevier, 2020. p. 473–508.

INES, A. V. M. et al. Assimilation of remotely sensed soil moisture and vegetation with a crop simulation model for maize yield prediction. **Remote Sensing of Environment**, v. 138, p. 149–164, 1 nov. 2013.

INMAN-BAMBER, N. G. A growth model for sugar-cane based on a simple carbon balance and the CERES-Maize water balance. **South African Journal of Plant and Soil**, v. 8, n. 2, p. 93–99, 1991. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/02571862.1991.10634587>. Acesso em: 7 abr. 2023.

INMAN-BAMBER, N. G. et al. Sugarcane for water-limited environments: Enhanced capability of the APSIM sugarcane model for assessing traits for transpiration efficiency and root water supply. **Field Crops Research**, v. 196, p. 112–123, 1 set. 2016.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA [IBGE]. **PAM - Produção Agrícola Municipal. Tabela 1612: Área plantada, área colhida, quantidade produzida, rendimento médio e valor da produção das lavouras temporárias**. Disponível em: <https://sidra.ibge.gov.br/tabela/1612>. Acesso em: 23 abr. 2023.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA, I. **Culturas temporárias e permanentes | IBGE**. Disponível em: <https://www.ibge.gov.br/estatisticas/economicas/agricultura-e-pecuaria/9117-producao-agricola-municipal-culturas-temporarias-e-permanentes.html>. Acesso em: 30 abr. 2023.

JACKSON, T. J. et al. Vegetation water content mapping using Landsat data derived normalized difference water index for corn and soybeans. **Remote Sensing of Environment**, v. 92, n. 4, p. 475–482, 30 set. 2004.

JAME, Y. W.; CUTFORTH, H. W. Crop growth models for decision support systems. **Canadian Journal of Plant Science**, v. 76, n. 1, p. 9–19, jan. 1996. Disponível em: <http://www.nrcresearchpress.com/doi/10.4141/cjps96-003>. Acesso em: 2 maio. 2018.

JAMES, G. et al. **An Introduction to Statistical Learning**. 2. ed. New York, NY: Springer US, 2013.

JEONG, J. H. et al. Random Forests for Global and Regional Crop Yield Predictions. **PLOS ONE**, v. 11, n. 6, p. e0156571, 1 jun. 2016. Disponível em: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0156571>. Acesso em: 13 abr. 2023.

JIANG, Z. et al. Development of a two-band enhanced vegetation index without a blue band. **Remote Sensing of Environment**, v. 112, n. 10, p. 3833–3845, 15 out. 2008.

JIANG, Z. et al. Application of crop model data assimilation with a particle filter for estimating regional winter wheat yields. **IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing**, v. 7, n. 11, p. 4422–4431, 1 nov. 2014.

JIN, X. et al. A review of data assimilation of remote sensing and crop models. **European Journal of Agronomy**, v. 92, p. 141–152, 1 jan. 2018.

JOHN, K. et al. Using Machine Learning Algorithms to Estimate Soil Organic Carbon Variability with Environmental Variables and Soil Nutrient Indicators in an Alluvial Soil. **Land 2020, Vol. 9, Page 487**, v. 9, n. 12, p. 487, 2 dez. 2020. Disponível em: <https://www.mdpi.com/2073-445X/9/12/487/htm>. Acesso em: 13 abr. 2023.

JOHNSON, D. M. An assessment of pre- and within-season remotely sensed variables for forecasting corn and soybean yields in the United States. **Remote Sensing of Environment**, v. 141, p. 116–128, 5 fev. 2014.

JOLLIFE, I. T.; CADIMA, J. Principal component analysis: a review and recent developments. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 374, n. 2065, 13 abr. 2016. Disponível em: <https://royalsocietypublishing.org/doi/10.1098/rsta.2015.0202>. Acesso em: 26 abr. 2023.

JONES, J. . et al. The DSSAT cropping system model. **European Journal of Agronomy**, v. 18, n. 3–4, p. 235–265, 2003.

JONES, M. R. et al. Evaluating process-based sugarcane models for simulating genotypic and environmental effects observed in an international dataset. **Field Crops Research**, v. 260, p. 107983, 1 jan. 2021.

JÖNSSON, P.; EKLUNDH, L. TIMESAT—a program for analyzing time-series of satellite sensor data. **Computers & Geosciences**, v. 30, n. 8, p. 833–845, 1 out. 2004.

KAHRAMAN, S.; BACHER, R. A comprehensive review of hyperspectral data fusion with lidar and sar data. **Annual Reviews in Control**, v. 51, p. 236–253, 1 jan. 2021.

KAMIR, E.; WALDNER, F.; HOCHMAN, Z. Estimating wheat yields in Australia using climate records, satellite image time series and machine learning methods. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 160, p. 124–135, 1 fev. 2020.

KARP, S. G. et al. Bioeconomy and biofuels: the case of sugarcane ethanol in Brazil . **Biofuels Bioproducts and Biorefining**, 2021. Disponível em: <https://www.researchgate.net/publication/349477799>. Acesso em: 30 abr. 2023.

KASAMPALIS, D. A. et al. Contribution of Remote Sensing on Crop Models: A Review. **Journal of Imaging 2018, Vol. 4, Page 52**, v. 4, n. 4, p. 52, 23 mar. 2018. Disponível em: <https://www.mdpi.com/2313-433X/4/4/52/htm>. Acesso em: 8 abr. 2023.

KE, G. et al. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. **Advances in Neural Information Processing Systems**, v. 30, 2017. Disponível em: <https://github.com/Microsoft/LightGBM.>. Acesso em: 26 abr. 2023.

KEATING, B. . et al. An overview of APSIM, a model designed for farming systems simulation. **European Journal of Agronomy**, v. 18, n. 3–4, p. 267–288, 1 jan. 2003.

KEATING, B. A. et al. Modelling sugarcane production systems I. Development and performance of the sugarcane module. **Field Crops Research**, v. 61, n. 3, p. 253–271, 1999.

KEATING, B. A.; THORBURN, P. J. Modelling crops and cropping systems—Evolving purpose, practice and prospects. **European Journal of Agronomy**, v. 100, p. 163–176, 1 out. 2018.

KHAN, N. et al. Prediction of Oil Palm Yield Using Machine Learning in the Perspective of Fluctuating Weather and Soil Moisture Conditions: Evaluation of a Generic Workflow. **Plants**, v. 11, n. 13, p. 1697, 27 jun. 2022. Disponível em: <https://www.mdpi.com/2223-7747/11/13/1697/htm>. Acesso em: 2 maio. 2023.

KINIRY, J. R. et al. EPIC model parameters for cereal, oilseed, and forage crops in the northern Great Plains region. **Canadian Journal of Plant Science**, v. 75, n. 3, p. 679–688, 1995. Disponível em: <https://cdnsciencepub.com/doi/10.4141/cjps95-114>. Acesso em: 7 abr. 2023.

KIRKHAM, M. B. Field Capacity, Wilting Point, Available Water, and the Non-Limiting Water Range. In: KIRKHAM, M. B. (Ed.). **Principles of Soil and Plant Water Relations**. London: Academic Press, 2005. p. 101–115.

KOHAVI, R. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection. In: International Joint Conference on Arti_cial Intelligence, San Mateo. **Anais**... San Mateo: Morgan Kaufman, 1995. Disponível em: <http://robotics.stanford.edu/~ronnyk>. Acesso em: 26 abr. 2023.

KOSMOWSKI, F. et al. How accurate are yield estimates from crop cuts? Evidence from smallholder maize farms in Ethiopia. **Food Policy**, v. 102, p. 102122, 1 jul. 2021.

KOUADIO, L. et al. Assessing the Performance of MODIS NDVI and EVI for Seasonal Crop Yield Forecasting at the Ecodistrict Scale. **Remote Sensing 2014, Vol. 6, Pages 10193-10214**, v. 6, n. 10, p. 10193–10214, 23 out. 2014. Disponível em: <https://www.mdpi.com/2072-4292/6/10/10193/htm>. Acesso em: 4 abr. 2023.

KOUADIO, L. et al. Artificial intelligence approach for the prediction of Robusta coffee yield using soil fertility properties. **Computers and Electronics in Agriculture**, v. 155, p. 324–338, 1 dez. 2018.

KUHN, M.; JOHNSON, K. Applied predictive modeling. **Springer**, p. 1–600, 1 jan. 2013.

KUNKEL, V. R.; WELLS, T.; HANCOCK, G. R. Modelling soil organic carbon using vegetation indices across large catchments in eastern Australia. **Science of The Total Environment**, v. 817, p. 152690, 15 abr. 2022.

LABUS, M. P. et al. Wheat yield estimates using multi-temporal NDVI satellite imagery. **International Journal of Remote Sensing**, v. 23, n. 20, p. 4169–4180, 10 out. 2002. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/01431160110107653>. Acesso em: 16 abr. 2023.

LEAL, D. V. P. **Parametrização do modelo CANEGRO (DSSAT) e caracterização biométrica de oito variedades de cana-de-açúcar irrigadas por gotejamento Piracicaba 2016**. 2016. Escola Superior de Agricultura "Luiz de Queiroz" (ESALQ-USP), Piracicaba, 2016.

LI, Y. et al. Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the U.S. **Field Crops Research**, v. 234, p. 55–65, 15 mar. 2019.

LIAKOS, K. et al. Machine Learning in Agriculture: A Review. **Sensors**, v. 18, n. 8, p. 2674, 14 ago. 2018.

LIU, H. Q.; HUETE, A. A feedback based modification of the NDVI to minimize canopy background and atmospheric noise. **IEEE Transactions on Geoscience and Remote Sensing**, v. 33, n. 2, p. 457–465, 28 jun. 1995.

LIU, X. et al. Predicting essential genes of 41 prokaryotes by a semi-supervised method. **Analytical Biochemistry**, v. 609, p. 113919, 15 nov. 2020.

LOBELL, D. B. The use of satellite data for crop yield gap analysis. **Field Crops Research**, v. 143, p. 56–64, 1 mar. 2013.

LOBELL, D. B. et al. A scalable satellite-based crop yield mapper. **Remote Sensing of Environment**, v. 164, p. 324–333, 1 jul. 2015.

LOOMIS, R. S.; RABBINGE, R.; NG, E. Explanatory Models in Crop Physiology. **Annual Review of Plant Physiology and Plant Molecular Biology**, v. 30, n. 1, p. 339–367, 28 nov. 1979. Disponível em: <https://www.annualreviews.org/doi/abs/10.1146/annurev.pp.30.060179.002011>. Acesso em: 7 abr. 2023.

LÓPEZ-LOZANO, R. et al. Towards regional grain yield forecasting with 1 km-resolution EO biophysical products: Strengths and limitations at pan-European level. **Agricultural and Forest Meteorology**, v. 206, p. 12–32, 15 jun. 2015.

LOZANO-GARZON, C. et al. Remote Sensing and Machine Learning Modeling to Support the Identification of Sugarcane Crops. **IEEE Access**, v. 10, p. 17542–17555, 2022.

LU, B. et al. Recent Advances of Hyperspectral Imaging Technology and Applications in Agriculture. **Remote Sensing 2020, Vol. 12, Page 2659**, v. 12, n. 16, p. 2659, 18 ago. 2020. Disponível em: <https://www.mdpi.com/2072-4292/12/16/2659/htm>. Acesso em: 16 abr. 2023.

MA, B. L. et al. Early Prediction of Soybean Yield from Canopy Reflectance Measurements. **Agronomy Journal**, v. 93, n. 6, p. 1227–1234, 1 nov. 2001. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2134/agronj2001.1227>. Acesso em: 4 abr. 2023.

MACEACHERN, S. J.; FORKERT, N. D. Machine learning for precision medicine. **Genome**, p. 416–425, 2021. Disponível em: <www.nrcresearchpress.com/gen>. Acesso em: 11 abr. 2023.

MAESTRINI, B. et al. Mixing process-based and data-driven approaches in yield prediction. **European Journal of Agronomy**, v. 139, p. 126569, 1 set. 2022.

MANATSA, D. et al. Maize yield forecasting for Zimbabwe farming sectors using satellite rainfall estimates. **Natural Hazards**, v. 59, n. 1, p. 447–463, 12 out. 2011. Disponível em: <https://link.springer.com/article/10.1007/s11069-011-9765-0>. Acesso em: 9 abr. 2023.

MARESMA, A. et al. Accuracy of NDVI-derived corn yield predictions is impacted by time of sensing. **Computers and Electronics in Agriculture**, v. 169, p. 105236, 1 fev. 2020.

MARIN, F. R. et al. Simulating Long-Term Effects of Trash Management on Sugarcane Yield for Brazilian Cropping Systems. **Sugar Tech**, v. 16, n. 2, p. 164–173, 2014.

MARIN, F. R. et al. Sugarcane model intercomparison: Structural differences and uncertainties under current and potential future climates. **Environmental Modelling & Software**, v. 72, p. 372–386, 1 out. 2015.

MARTINÉ, J.-F. **Modélisation de la production potentielle de la canne à sucre en zone tropicale, sous conditions thermiques et hydriques contrastées. Applications du modèle**. 2003. s.n., France, 2003.

MARTINÉ, J. F.; SIBAND, P.; BONHOMME, R. Simulation of the maximum yield of sugar cane at different altitudes: effect of temperature on the conversion of radiation into biomass. **Agronomie**, v. 19, n. 1, p. 3–12, 1999. Disponível em: <http://dx.doi.org/10.1051/agro:19990101>. Acesso em: 7 abr. 2023.

MATSUSHITA, B. et al. Sensitivity of the Enhanced Vegetation Index (EVI) and Normalized Difference Vegetation Index (NDVI) to Topographic Effects: A Case Study in High-density Cypress Forest. **Sensors 2007, Vol. 7, Pages 2636-2651**, v. 7, n. 11, p. 2636–2651, 5 nov. 2007. Disponível em: <https://www.mdpi.com/1424-8220/7/11/2636/htm>. Acesso em: 5 abr. 2023.

MEDRADO, E.; LIMA, J. E. F. W. Development of pedotransfer functions for estimating water retention curve for tropical soils of the Brazilian savanna. **Geoderma Regional**, v. 1, n. C, p. 59–66, 1 set. 2014.

MESHRAM, V. et al. Machine learning in agriculture domain: A state-of-art survey. **Artificial Intelligence in the Life Sciences**, v. 1, p. 100010, 1 dez. 2021.

MIENYE, I. D.; SUN, Y.; WANG, Z. Prediction performance of improved decision tree-based algorithms: a review. **Procedia Manufacturing**, v. 35, p. 698–703, 1 jan. 2019.

MILLIGAN, S. B.; GRAVOIS, K. A.; MARTIN, F. A. Inheritance of Sugarcane Ratooning Ability and the Relationship of Younger Crop Traits to Older Crop Traits. **Crop Science**, v. 36, n. 1, p. 45–60, 1 jan. 1996. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2135/cropsci1996.0011183X003600010008x>. Acesso em: 26 abr. 2023.

MKHABELA, M. S. et al. Crop yield forecasting on the Canadian Prairies using MODIS NDVI data. **Agricultural and Forest Meteorology**, v. 151, n. 3, p. 385–393, 15 mar. 2011.

MONTEIRO, J. E. B. de A. et al. Rice yield estimation based on weather conditions and on technological level of production systems in Brazil. **Pesquisa Agropecuária Brasileira**, v. 48, n. 2, p. 123–131, fev. 2013. Disponível em: <http://www.scielo.br/j/pab/a/dtbmVhmn3DBhmNVqwm8JYgK/?lang=en>. Acesso em: 16 abr. 2023.

MONTEIRO, L. A.; SENTELHAS, P. C. Potential and Actual Sugarcane Yields in Southern Brazil as a Function of Climate Conditions and Crop Management. **Sugar Tech**, v. 16, n. 3, p. 264–276, 2014.

MONTEITH, J. L. Light Distribution and Photosynthesis in Field Crops. **Annals of Botany**, v. 29, n. 1, p. 17–37, 1 jan. 1965. Disponível em: <https://academic.oup.com/aob/article/29/1/17/185173>. Acesso em: 7 abr. 2023.

MOREL, J. et al. A comparison of two coupling methods for improving a sugarcane model yield estimation with a NDVI-derived variable. **Proc. SPIE 8531, Remote Sensing for Agriculture, Ecosystems, and Hydrology XIV**, v. 8531, p. 93–102, 19 out. 2012. Disponível em: <https://www.spiedigitallibrary.org/conference-proceedings-of-spie/8531/85310E/A-comparison-of-two-coupling-methods-for-improving-a-sugarcane/10.1117/12.975688.full>. Acesso em: 9 abr. 2023.

MOREL, J. et al. Toward a Satellite-Based System of Sugarcane Yield Estimation and Forecasting in Smallholder Farming Conditions: A Case Study on Reunion Island. **Remote Sensing** , v. 6, n. 7, p. 6620–6635, 18 jul. 2014. Disponível em: <https://www.mdpi.com/2072-4292/6/7/6620/htm>. Acesso em: 19 abr. 2023.

MORIONDO, M.; MASELLI, F.; BINDI, M. A simple model of regional wheat yield based on NDVI data. **European Journal of Agronomy**, v. 26, n. 3, p. 266–274, 1 abr. 2007.

MUÑOZ-SABATER, J. et al. ERA5-Land: A state-of-the-art global reanalysis dataset for land applications. **Earth System Science Data**, v. 13, n. 9, p. 4349–4383, 7 set. 2021.

MURPHY, J. M. et al. Quantification of modelling uncertainties in a large ensemble of climate change simulations. **Nature**, v. 430, n. 7001, p. 768–772, 12 ago. 2004.

NAGHDYZADEGAN JAHROMI, M. et al. Developing machine learning models for wheat yield prediction using ground-based data, satellite-based actual evapotranspiration and vegetation indices. **European Journal of Agronomy**, v. 146, p. 126820, 1 maio 2023. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S1161030123000886>. Acesso em: 6 abr. 2023.

NAGY, A. et al. Wheat Yield Forecasting for the Tisza River Catchment Using Landsat 8 NDVI and SAVI Time Series and Reported Crop Statistics. **Agronomy 2021, Vol. 11, Page 652**, v. 11, n. 4, p. 652, 29 mar. 2021. Disponível em: <https://www.mdpi.com/2073-4395/11/4/652/htm>. Acesso em: 4 abr. 2023.

NAVALGUND, R. R.; JAYARAMAN, V.; ROY, P. S. Remote sensing applications: An overview. v. 93, n. 12, 2007.

O'LEARY, G. J. et al. Modelling soil organic carbon 1. Performance of APSIM crop and pasture modules against long-term experimental data. **Geoderma**, v. 264, p. 227–237, 15 fev. 2016.

OLIVEIRA, A. P. P. de et al. The Response of Sugarcane to Trash Retention and Nitrogen in The Brazilian Coastal Tablelands: A Simulation Study. **Experimental Agriculture**, v. 52, n. 1, p. 69–86, 1 jan. 2015. Disponível em: <https://www.cambridge.org/core/journals/experimental-agriculture/article/response-of-sugarcane-to-trash-retention-and-nitrogen-in-the-brazilian-coastal-tablelands-a-simulation-study/57254454AD2F387C53D10E451457CEC9>. Acesso em: 16 abr. 2023.

OLIVEIRA, H. C. et al. Failure Detection in Row Crops from UAV Images Using Morphological Operators. **IEEE Geoscience and Remote Sensing Letters**, v. 15, n. 7, p. 991–995, 1 jul. 2018.

OLIVER, Y. M.; ROBERTSON, M. J.; WONG, M. T. F. Integrating farmer knowledge, precision agriculture tools, and crop simulation modelling to evaluate management options for poor-performing patches in cropping fields. **European Journal of Agronomy**, v. 32, n. 1, p. 40–50, jan. 2010.

PAGANI, V. et al. Forecasting sugarcane yields using agro-climatic indicators and Canegro model: A case study in the main production region in Brazil. **Agricultural Systems**, v. 154, p. 45–52, 1 jun. 2017.

PANEK, E.; GOZDOWSKI, D. Analysis of relationship between cereal yield and NDVI for selected regions of Central Europe based on MODIS satellite data. **Remote Sensing Applications: Society and Environment**, v. 17, p. 100286, 1 jan. 2020.

PANTAZI, X. E. et al. Wheat yield prediction using machine learning and advanced sensing techniques. **Computers and Electronics in Agriculture**, v. 121, p. 57–65, 1 fev. 2016.

PARK, S. J.; HWANG, C. S.; VLEK, P. L. G. Comparison of adaptive techniques to predict crop yield response under varying soil and land management conditions. **Agricultural Systems**, v. 85, n. 1, p. 59–81, 1 jul. 2005.

PASLEY, H. et al. How to build a crop model. A review. **Agronomy for Sustainable Development**, v. 43, n. 1, p. 1–12, 1 fev. 2023. Disponível em: <https://link.springer.com/article/10.1007/s13593-022-00854-9>. Acesso em: 7 abr. 2023.

PATEL, J. H.; OZA, M. P. Deriving Crop Calendar Using Ndvi Time-Series. **The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences**, n. 8, p. 9–12, 2014. Disponível em: <www.fao.org/>. Acesso em: 25 abr. 2023.

PEARSON, K. F. R. S. LIII. On lines and planes of closest fit to systems of points in space. **The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science** , v. 2, n. 11, p. 559–572, nov. 1901. Disponível em: <https://www.tandfonline.com/doi/abs/10.1080/14786440109462720>. Acesso em: 26 abr. 2023.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python . **Journal of Machine Learning Research**, v. 12, p. 2825–2830, 2011. Disponível em: <http://scikit-learn.sourceforge.net.>. Acesso em: 26 abr. 2023.

PELOSI, A.; CHIRICO, G. B. Regional assessment of daily reference evapotranspiration: Can ground observations be replaced by blending ERA5-Land meteorological reanalysis and CM-SAF satellite-based radiation data? **Agricultural Water Management**, v. 258, p. 107169, 1 dez. 2021.

PETTORELLI, N. et al. Using the satellite-derived NDVI to assess ecological responses to environmental change. **Trends in Ecology & Evolution**, v. 20, n. 9, p. 503–510, 1 set. 2005.

PORCÙ, F.; MILANI, L.; PETRACCA, M. On the uncertainties in validating satellite instantaneous rainfall estimates with raingauge operational network. **Atmospheric Research**, v. 144, p. 73–81, 1 jul. 2014.

PORTER, J. R. A model of canopy development in winter wheat. **The Journal of Agricultural Science**, v. 102, n. 2, p. 383–392, 1984. Disponível em: <https://www.cambridge.org/core/journals/journal-of-agricultural-science/article/abs/model-of-canopy-development-in-winter-wheat/19CC94FABD9E61C521A2CD0BB5A79E0F>. Acesso em: 7 abr. 2023.

PRIBYL, D. W. A critical review of the conventional SOC to SOM conversion factor. **Geoderma**, v. 156, n. 3–4, p. 75–83, 15 maio 2010.

PRIYA, S. R. K. et al. Sugarcane yield forecast using weather based discriminant analysis. **Smart Agricultural Technology**, v. 3, p. 100076, 1 fev. 2023.

PRIYADARSHI, N. et al. Reconstruction of time series MODIS EVI data using de-noising algorithms. **Geocarto International**, v. 33, n. 10, p. 1095–1113, 3 out. 2018.

PROKHORENKOVA, L. et al. CatBoost: unbiased boosting with categorical features. **Advances in Neural Information Processing Systems**, v. 31, 2018. Disponível em: <https://github.com/catboost/catboost>. Acesso em: 26 abr. 2023.

QIAN, B. et al. Statistical spring wheat yield forecasting for the Canadian prairie provinces. **Agricultural and Forest Meteorology**, v. 149, n. 6–7, p. 1022–1031, 15 jun. 2009.

R CORE TEAM. **R: A language and environment for statistical computing. R Foundation for Statistical Computing2021**Vienna, AustriaR Foundation for Statistical Computing, , 2022. . Disponível em: <https://www.r-project.org/>. Acesso em: 27 abr. 2023.

RAMBURAN, S. et al. Genetic, environmental and management contributions to ratoon decline in sugarcane. **Field Crops Research**, v. 146, p. 105–112, 1 maio 2013.

RAMOS, P. J. et al. Automatic fruit count on coffee branches using computer vision. **Computers and Electronics in Agriculture**, v. 137, p. 9–22, 1 maio 2017.

REICHSTEIN, M. et al. Deep learning and process understanding for data-driven Earth system science. **Nature 2019 566:7743**, v. 566, n. 7743, p. 195–204, 13 fev. 2019. Disponível em: <https://www.nature.com/articles/s41586-019-0912-1>. Acesso em: 14 abr. 2023.

REMBOLD, F. et al. Using Low Resolution Satellite Imagery for Yield Prediction and Yield Anomaly Detection. **Remote Sensing 2013, Vol. 5, Pages 1704-1733**, v. 5, n. 4, p. 1704–1733, 8 abr. 2013. Disponível em: <https://www.mdpi.com/2072-4292/5/4/1704/htm>. Acesso em: 16 abr. 2023.

REN, J. et al. Regional yield estimation for winter wheat with MODIS-NDVI data in Shandong, China. **International Journal of Applied Earth Observation and Geoinformation**, v. 10, n. 4, p. 403–413, 1 dez. 2008.

REN, J. et al. Integrating remotely sensed LAI with EPIC model based on global optimization algorithm for regional crop yield assessment. **International Geoscience and Remote Sensing Symposium (IGARSS)**, p. 2147–2150, 2010.

REN, Y. et al. Analysis of Corn Yield Prediction Potential at Various Growth Phases Using a Process-Based Model and Deep Learning. **Plants 2023, Vol. 12, Page 446**, v. 12, n. 3, p. 446, 18 jan. 2023. Disponível em: <https://www.mdpi.com/2223-7747/12/3/446/htm>. Acesso em: 16 abr. 2023.

RITCHIE, J. R.; OTTER, S. Description and performance of CERES-Wheat: a user-oriented wheat yield model. **ARS - United States Department of Agriculture, Agricultural Research Service (USA)**, 1985. Disponível em: <https://agris.fao.org/agris-search/search.do?recordID=US8637126>. Acesso em: 7 abr. 2023.

ROBERTS, M. J. et al. Comparing and combining process-based crop models and statistical models with some implications for climate change. **Environmental Research Letters**, v. 12, n. 9, p. 095010, 1 set. 2017.

ROSENZWEIG, C. et al. The Agricultural Model Intercomparison and Improvement Project (AgMIP): Protocols and pilot studies. **Agricultural and Forest Meteorology**, v. 170, p. 166–182, mar. 2013.

ROUSE, R. W. H.; HAAS, J. A. W.; DEERING, D. W. Monitoring Vegetation Systems in the Great Plains With ERTS. In: Washington. **Anais**... Washington: 1973.

RUMPF, T. et al. Early detection and classification of plant diseases with Support Vector Machines based on hyperspectral reflectance. **Computers and Electronics in Agriculture**, v. 74, n. 1, p. 91–99, 1 out. 2010.

SAHOO, S. et al. Machine learning algorithms for modeling groundwater level changes in agricultural regions of the U.S. **Water Resources Research**, v. 53, n. 5, p. 3878–3895, 2017.

SAKAMOTO, T. et al. A crop phenology detection method using time-series MODIS data. **Remote Sensing of Environment**, v. 96, n. 3–4, p. 366–374, 30 jun. 2005.

SAMUEL, A. L. Some Studies in Machine Learning Using the Game of Checkers. **IBM Journal of Research and Development**, v. 3, n. 3, p. 210–229, jul. 1959. Disponível em: <http://ieeexplore.ieee.org/document/5392560/>. Acesso em: 11 abr. 2023.

SANTOS, B. C. dos et al. Análise espaço-temporal da precipitação na região central do estado de São Paulo utilizando dados CHIRPS. **Revista Brasileira de Geografia Física**, v. 15, n. 5, p. 2582–2600, 2022.

SAVITZKY, A.; GOLAY, M. J. E. Smoothing and Differentiation of Data by Simplified Least Squares Procedures. **Analytical Chemistry**, v. 36, n. 8, p. 1627–1639, 1 jul. 1964. Disponível em: <https://pubs.acs.org/doi/pdf/10.1021/ac60214a047>. Acesso em: 24 abr. 2023.

SCHAUBERGER, B.; JÄGERMEYR, J.; GORNOTT, C. A systematic review of local to regional yield forecasting approaches and frequently used data resources. **European Journal of Agronomy**, v. 120, p. 126153, 1 out. 2020.

SCHEPEN, A.; EVERINGHAM, Y.; WANG, Q. J. An improved workflow for calibration and downscaling of GCM climate forecasts for agricultural applications – A case study on prediction of sugarcane yield in Australia. **Agricultural and Forest Meteorology**, v. 291, p. 107991, 15 set. 2020.

SCHER, S.; MESSORI, G. Predicting weather forecast uncertainty with machine learning. **Quarterly Journal of the Royal Meteorological Society**, v. 144, n. 717, p. 2830–2841, 1 out. 2018. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/qj.3410>. Acesso em: 11 abr. 2023.

SCHMIDT, A. F.; FINAN, C. Linear regression and the normality assumption. **Journal of Clinical Epidemiology**, v. 98, p. 146–151, 1 jun. 2018.

SCHMITT, J. et al. Extreme weather events cause significant crop yield losses at the farm level in German agriculture. **Food Policy**, v. 112, p. 102359, 1 out. 2022.

SCHULTZ, N.; REIS, V. M.; URQUIAGA, S. **Resposta da cana-de-açúcar à adubação nitrogenada: fontes nitrogenadas, formas de aplicação, épocas de aplicação e efeito varietal**. 1. ed. Seropédica, RJ: EMBRAPA Agrobiologia, 2015. v. 1

SCHWALBERT, R. A. et al. Satellite-based soybean yield forecast: Integrating machine learning and weather data for improving crop yield prediction in southern Brazil. **Agricultural and Forest Meteorology**, v. 284, p. 107886, 15 abr. 2020.

SEGARRA, J. et al. Remote Sensing for Precision Agriculture: Sentinel-2 Improved Features and Applications. **Agronomy 2020, Vol. 10, Page 641**, v. 10, n. 5, p. 641, 1 maio 2020. Disponível em: <https://www.mdpi.com/2073-4395/10/5/641/htm>. Acesso em: 16 abr. 2023.

SEMERARO, F.; GRIFFITHS, A.; CANGELOSI, A. Human–robot collaboration and machine learning: A systematic review of recent research. **Robotics and Computer-Integrated Manufacturing**, v. 79, p. 102432, 1 fev. 2023.

SENTELHAS, P. C. et al. Evaluation of the WGEN and SIMMETEO weather generators for the brazilian tropics and subtropics, using crop simulation models . **Revista Brasileira de Agrometeorologia**, v. 9, n. 2, p. 357–376, 2001.

SETHY, P. K. et al. Deep feature based rice leaf disease identification using support vector machine. **Computers and Electronics in Agriculture**, v. 175, p. 105527, 1 ago. 2020.

SETIYONO, T. et al. Rice yield estimation using synthetic aperture radar (SAR) and the ORYZA crop growth model: development and application of the system in South and South-east Asian countries Visual Assessment for Rice and Gardens Soils in the Mekong Delta View project Global Futures for Agriculture and Strategic Foresight View project. **Article in International Journal of Remote Sensing**, 2018. Disponível em: <https://doi.org/10.1080/01431161.2018.1547457>. Acesso em: 5 abr. 2023.

SEYEDZADEH, A. et al. Artificial intelligence approach to estimate discharge of drip tape irrigation based on temperature and pressure. **Agricultural Water Management**, v. 228, p. 105905, 20 fev. 2020.

SHAHHOSSEINI, M. et al. Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. **Scientific Reports 2021 11:1**, v. 11, n. 1, p. 1–15, 15 jan. 2021. Disponível em: <https://www.nature.com/articles/s41598-020-80820-1>. Acesso em: 15 abr. 2023.

SHANAHAN, J. F. et al. Use of Remote-Sensing Imagery to Estimate Corn Grain Yield. **Agronomy Journal**, v. 93, n. 3, p. 583–589, 1 maio 2001. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2134/agronj2001.933583x>. Acesso em: 4 abr. 2023.

SHANMUGAPRIYA, P. et al. Applications of Remote Sensing in Agriculture-A Review. **Int.J.Curr.Microbiol.App.Sci**, v. 8, n. 1, p. 2270–2283, 2019. Disponível em: <https://doi.org/10.20546/ijcmas.2019.801.238>. Acesso em: 4 abr. 2023.

SHARMA, A. et al. Machine Learning Applications for Precision Agriculture: A Comprehensive Review. **IEEE Access**, v. 9, p. 4843–4873, 2021.

SHENDRYK, Y.; DAVY, R.; THORBURN, P. Integrating satellite imagery and environmental data to predict field-level cane and sugar yields in Australia using machine learning. **Field Crops Research**, v. 260, p. 107984, 1 jan. 2021.

SILVA, A. C. et al. SPLINTEX: A physically-based pedotransfer function for modeling soil hydraulic functions. **Soil and Tillage Research**, v. 174, p. 261–272, 1 dez. 2017.

SILVA, E. H. D. L. et al. Performance Assessment of Different Precipitation Databases (Gridded Analyses and Reanalyses) for the New Brazilian Agricultural Frontier: SEALBA. **Water**, v. 14, n. 9, p. 1473, 4 maio 2022. Disponível em: <https://www.mdpi.com/2073-4441/14/9/1473/htm>. Acesso em: 17 abr. 2023.

SINCLAIR, T. R.; SELIGMAN, N. Criteria for publishing papers on crop modeling. **Field Crops Research**, v. 68, n. 3, p. 165–172, 1 nov. 2000.

SINGELS, A. Crop Models. In: MOORE, P. H.; BOTHA, F. C. (Ed.). **Sugarcane: Physiology, Biochemistry, and Functional Biology**. [s.l.] John Wiley & Sons, Ltd, 2013. p. 541–577.

SINGH, P. K. et al. Forecasting of wheat yield in various agro-climatic regions of Bihar by using CERES-wheat model. **Journal of Agrometeorology**, v. 19, n. 4, p. 346–349, 1 dez. 2017.

SIVASANKAR, T. et al. Advances in Radar Remote Sensing of Agricultural Crops: A Review. v. 8, n. 4, 2018.

SKAKUN, S. et al. Assessing within-Field Corn and Soybean Yield Variability from WorldView-3, Planet, Sentinel-2, and Landsat 8 Satellite Imagery. **Remote Sensing 2021, Vol. 13, Page 872**, v. 13, n. 5, p. 872, 26 fev. 2021. Disponível em: <https://www.mdpi.com/2072-4292/13/5/872/htm>. Acesso em: 4 abr. 2023.

SOLTANI, A.; HOOGENBOOM, G. Assessing crop management options with crop simulation models based on generated weather data. **Field Crops Research**, v. 103, n. 3, p. 198–207, 13 set. 2007.

SOUZA, C. M. et al. Reconstructing Three Decades of Land Use and Land Cover Changes in Brazilian Biomes with Landsat Archive and Earth Engine. **Remote Sensing 2020, Vol. 12, Page 2735**, v. 12, n. 17, p. 2735, 25 ago. 2020. Disponível em: <https://www.mdpi.com/2072-4292/12/17/2735/htm>. Acesso em: 12 fev. 2022.

STANFILL, B. **apsimr: Edit, Run and Evaluate APSIM Simulations Easily Using R** . Disponível em: <https://rdrr.io/cran/apsimr/>. Acesso em: 27 abr. 2023.

STÖCKLE, C. O.; DONATELLI, M.; NELSON, R. CropSyst, a cropping systems simulation model. **European Journal of Agronomy**, v. 18, n. 3–4, p. 289–307, 1 jan. 2003.

SUN, L. et al. Monitoring surface soil moisture status based on remotely sensed surface temperature and vegetation index information. **Agricultural and Forest Meteorology**, v. 166–167, p. 175–187, 15 dez. 2012.

SUPIT, I.; HOOIJER, A. A.; VAN DIEPEN, C. A. System description of the WOFOST 6.0 crop simulation model implemented in CGMS. **theory and algorithms**, v. 1, p. 146-, 1994. Disponível em: <https://cir.nii.ac.jp/crid/1573950399770954368>. Acesso em: 7 abr. 2023.

SUTTON, R. S.; BARTO, A. G. **Reinforcement Learning: An Introduction**. 2. ed. Cambridge, MA: The MIT Press, 2018.

TELLAECHE, A. et al. A computer vision approach for weeds identification through Support Vector Machines. **Applied Soft Computing**, v. 11, n. 1, p. 908–915, 1 jan. 2011.

THOMPSON, L. M. Weather and Technology in the Production of Corn in the U. S. Corn Belt1. **Agronomy Journal**, v. 61, n. 3, p. 453–456, 1 maio 1969. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2134/agronj1969.00021962006100030037x>. Acesso em: 2 abr. 2023.

THORNLEY, J. H. M.; JOHNSON, I. R. **Plant and crop modelling : a mathematical approach to plant and crop physiology**. [s.l.] Blackburn Press, 2000.

THORNTON, P. K. et al. Estimating millet production for famine early warning: an application of crop simulation modelling using satellite and ground-based data in Burkina Faso. **Agricultural and Forest Meteorology**, v. 83, n. 1–2, p. 95–112, 1 jan. 1997.

THORP, K. R. et al. Methodology for the use of DSSAT models for precision agriculture decision support. **Computers and Electronics in Agriculture**, v. 64, n. 2, p. 276–285, dez. 2008.

THORP, K. R.; HUNSAKER, D. J.; FRENCH, A. N. Assimilating leaf area index estimates from remote sensing into the simulations of a cropping systems model. **Transactions of the ASABE**, v. 53, n. 1, p. 251–262, 2010.

TOGLIATTI, K. et al. How does inclusion of weather forecasting impact in-season crop model predictions? **Field Crops Research**, v. 214, p. 261–272, 1 dez. 2017.

TOMASELLA, J.; HODNETT, M. G.; ROSSATO, L. Pedotransfer functions for the estimation of soil water retention in Brazilian soils. **Soil Science Society of America Journal**, v. 64, n. 1, p. 327–338, jan. 2000.

VALADE, A. et al. ORCHIDEE-STICS, a process-based model of sugarcane biomass production: calibration of model parameters governing phenology. **GCB Bioenergy**, v. 6, n. 5, p. 606–620, 1 set. 2014. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1111/gcbb.12074>. Acesso em: 7 abr. 2023.

VAN DAM, J. C. et al. Theory of SWAP version 2.0 Simulation of water flow, solute transport and plant growth in the Soil-Water-Atmosphere-Plant environment. **geningen Agricultural University and DLO Winand Staring Centre**, p. 1–168, 1997.

VAN GENUCHTEN, M. T. van. A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils. **Soil Science Society of America Journal**, v. 44, n. 5, p. 892–898, 1 set. 1980. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.2136/sssaj1980.03615995004400050002x>. Acesso em: 24 abr. 2023.

VAN ITTERSUM, M. K.; DONATELLI, M. Modelling cropping systems—highlights of the symposium and preface to the special issues. **European Journal of Agronomy**, v. 18, n. 3–4, p. 187–197, 1 jan. 2003.

VANDENBERGHE, L. P. S. et al. Beyond sugar and ethanol: The future of sugarcane biorefineries in Brazil. **Renewable and Sustainable Energy Reviews**, v. 167, p. 112721, 1 out. 2022.

VANELLA, D. et al. Comparing the use of ERA5 reanalysis dataset and ground-based agrometeorological data under different climates and topography in Italy. **Journal of Hydrology: Regional Studies**, v. 42, p. 101182, 1 ago. 2022.

VANNOPPEN, A.; GOBIN, A. Estimating farm wheat yields from NDVI and meteorological data. **Agronomy**, v. 11, n. 5, 1 maio 2021.

VANROSSUM, G. Python reference manual. **Department of Computer Science [CS]**, n. R 9525, 1 jan. 1995.

VASQUES, G. M. et al. **Soil Clay, Silt and Sand Content Maps for Brazil at 0-5, 5-15, 15-30, 30-60, 60-100 and 100-200 cm Depth Intervals with 90 m Spatial Resolution. Version 2021**. Disponível em: <http://geoinfo.cnps.embrapa.br/maps/3290>. Acesso em: 24 abr. 2023.

VENANCIO, L. P. et al. Forecasting corn yield at the farm level in Brazil based on the FAO-66 approach and soil-adjusted vegetation index (SAVI). **Agricultural Water Management**, v. 225, p. 105779, 20 nov. 2019.

VERMA, A. K. et al. Variety-specific sugarcane yield simulations and climate change impacts on sugarcane yield using DSSAT-CSM-CANEGRO model. **Agricultural Water Management**, v. 275, p. 108034, 1 jan. 2023.

VERMOTE, E. F.; VERMEULEN, A. Atmospheric Correction Algorithm: Spectral Reflectances (MOD09). **MODIS Algorithm Technical Background Document** , v. Version 4, abr. 1999.

VIRTANEN, P. et al. **Scipy/Scipy: Scipy 1.1.0**. Disponível em: <https://ui.adsabs.harvard.edu/abs/2018zndo...1241501V/abstract>. Acesso em: 25 abr. 2023.

VITTI, A. C. et al. Produtividade da cana-de-açúcar relacionada ao nitrogênio residual da adubação e do sistema radicular. **Pesquisa Agropecuária Brasileira**, v. 42, n. 2, p. 249–256, fev. 2007. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-204X2007000200014&lng=pt&tlng=pt>. Acesso em: 2 maio. 2018.

WACHOWIAK, M. P. et al. Visual analytics and remote sensing imagery to support community-based research for precision agriculture in emerging areas. **Computers and Electronics in Agriculture**, v. 143, p. 149–164, dez. 2017.

WADOUX, A. M. J. C.; MINASNY, B.; MCBRATNEY, A. B. Machine learning for digital soil mapping: Applications, challenges and suggested solutions. **Earth-Science Reviews**, v. 210, p. 103359, 1 nov. 2020.

WALDAMICHAEL, F. G. et al. Machine Learning in Cereal Crops Disease Detection: A Review. **Algorithms 2022, Vol. 15, Page 75**, v. 15, n. 3, p. 75, 24 fev. 2022. Disponível em: <https://www.mdpi.com/1999-4893/15/3/75/htm>. Acesso em: 16 abr. 2023.

WANG, A.; ZHANG, W.; WEI, X. A review on weed detection using ground-based machine vision and image processing techniques. **Computers and Electronics in Agriculture**, v. 158, p. 226–240, 1 mar. 2019.

WANG, J. et al. Mapping sugarcane plantation dynamics in Guangxi, China, by time series Sentinel-1, Sentinel-2 and Landsat images. **Remote Sensing of Environment**, v. 247, p. 111951, 15 set. 2020.

WANG, R. et al. Corn Response to Climate Stress Detected with Satellite-Based NDVI Time Series. **Remote Sensing 2016, Vol. 8, Page 269**, v. 8, n. 4, p. 269, 23 mar. 2016. Disponível em: <https://www.mdpi.com/2072-4292/8/4/269/htm>. Acesso em: 4 abr. 2023.

WANG, X. et al. Use of Ceres-wheat model for wheat yield forecast in Beijing. **IFIP Advances in Information and Communication Technology**, v. 293, p. 9–18, 2009. Disponível em: <https://link.springer.com/chapter/10.1007/978-1-4419-0209-2_4>. Acesso em: 8 abr. 2023.

WEISS, M.; JACOB, F.; DUVEILLER, G. Remote sensing for agricultural applications: A meta-review. **Remote Sensing of Environment**, v. 236, p. 111402, 1 jan. 2020.

WÓJTOWICZ, M.; WÓJTOWICZ, A.; PIEKARCZYK, J. Application of remote sensing methods in agriculture. **Communications in Biometry and Crop Science**, v. 11, p. 31–50, 2016.

XU, F. et al. Sugarcane Ratooning Ability: Research Status, Shortcomings, and Prospects. **Biology**, v. 10, n. 10, 1 out. 2021. Disponível em: </pmc/articles/PMC8533141/>. Acesso em: 26 abr. 2023.

YAO, F. et al. Estimation of maize yield by using a process-based model and remote sensing data in the Northeast China Plain. **Physics and Chemistry of the Earth, Parts A/B/C**, v. 87–88, p. 142–152, 1 jan. 2015.

YU, J. et al. Deep learning for image-based weed detection in turfgrass. **European Journal of Agronomy**, v. 104, p. 78–84, 1 mar. 2019.

ZHANG, S. et al. Learning k for kNN classification. **ACM Trans. Intell. Syst. Technol**, v. 8, n. 43, p. 1–19, 2017. Disponível em: <http://dx.doi.org/10.1145/2990508>. Acesso em: 19 abr. 2023.

ZHANG, S. et al. Developing a process-based and remote sensing driven crop yield model for maize (PRYM–Maize) and its validation over the Northeast China Plain. **Journal of Integrative Agriculture**, v. 20, n. 2, p. 408–423, 1 fev. 2021.

ZHAO, Y.; CHEN, S.; SHEN, S. Assimilating remote sensing information with crop model using Ensemble Kalman Filter for improving LAI monitoring and yield estimation. **Ecological Modelling**, v. 270, p. 30–42, 1 dez. 2013.

ZHOU, X. et al. Predicting grain yield in rice using multi-temporal vegetation indices from UAV-based multispectral and digital imagery. **ISPRS Journal of Photogrammetry and Remote Sensing**, v. 130, p. 246–255, 1 ago. 2017.

ZHOU, Y. et al. Machine learning for food security: Principles for transparency and usability. **Applied Economic Perspectives and Policy**, v. 44, n. 2, p. 893–910, 1 jun. 2022. Disponível em: <https://onlinelibrary.wiley.com/doi/full/10.1002/aepp.13214>. Acesso em: 11 abr. 2023.

ZHOU, Z. et al. Assessment for crop water stress with infrared thermal imagery in precision agriculture: A review and future prospects for deep learning applications. **Computers and Electronics in Agriculture**, v. 182, p. 106019, 1 mar. 2021.

ZOU, J. et al. Performance of air temperature from ERA5-Land reanalysis in coastal urban agglomeration of Southeast China. **Science of The Total Environment**, v. 828, p. 154459, 1 jul. 2022.

ZULUAGA, C. F. et al. Climatology and trends of downward shortwave radiation over Brazil. **Atmospheric Research**, v. 250, p. 105347, 1 mar. 2021.

ZULUAGA, C. F. et al. Radiation Balance Estimates Over Southeastern Brazil: Ground Observations, Satellite and Reanalysis. **Revista Brasileira de Meteorologia**, 14 abr. 2023. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-77862023005003201&tlng=en>. Acesso em: 17 abr. 2023.

APPENDICES

Appendix A. Input variables used in the regression models. The variables were aggregated temporal for each sugarcane field and then, by municipality to be in the same scale as the observed yield. Note that soil APSIM variables have multi layers data for each sugarcane field. The aggregation performed were a temporal mean and standard deviation for each sugarcane field, by expanding method. After that, a mean and standard deviation was performed in spatial scale by each municipality, which have led to a total of 158 variables that were then reduced by Principal components technique to be used in the regression models.

Variable	Data source	Description	Units
biomass	APSIM	Crop above-ground biomass (Green + Trash)	g m^{-2}
green_biomass	APSIM	Total green crop above-ground biomass	g m^{-2}
canefw	APSIM	Fresh cane weight	t h^{-1}
lai	APSIM	Leaf area index of green leaves	mm^{-2} mm^{-2}
cane_wt	APSIM	Weight of cane dry matter	g m^{-2}
leafgreenwt	APSIM	Green leaf weight	g m^{-2}
rootgreenwt	APSIM	Green root weight	g m^{-2}
sucrose_wt	APSIM	Sucrose weight	g m^{-2}
senescedwt	APSIM	Senesced biomass weight	g m^{-2}
height	APSIM	Maximum canopy height	mm
sw_demand	APSIM	Daily demand for soil water	mm
esw	APSIM	Extractable Soil water in each soil layer	mm
nfact_photo	APSIM	Nitrogen stress factor for photosynthesis	0-1
ep	APSIM	Evapotranspiration (extraction) for each soil layer	mm
eo	APSIM	Crop potential evapotranspiration	mm
cep	APSIM	Cumulative plant evapotranspiration	mm
swdef_expan	APSIM	Soil water stress factor for cell expansion	0-1
swdef_pheno	APSIM	Soil water stress factor for phenology	0-1
swdef_photo	APSIM	Soil water stress factor for photosynthesis	0-1
swdef_stalk	APSIM	Soil water stress factor for stalks	0-1
st	APSIM	Soil temperature for each layer	°C
sws	APSIM	Soil water content for each layer	cm^{-3} cm^{-3}
Rain	CHIRPS	Rainfall	mm
Radn	ERA5-Land	Global solar radiation	MJ m^{-2} day^{-1}
Maxt	ERA5-Land	Maximum air temperature	°C
Mint	ERA5-Land	Minimum air temperature	°C
Growing degree days	ERA5-Land	Based on mean air temperature (base 16, ceiling 45)	°C day
NDVI	MODIS	Normalized difference vegetation index	-1 to 1
EVI	MODIS	Enhanced vegetation index	-1 to 1
SAVI	MODIS	Soil-adjusted vegetation index	-1 to 1
NDMI	MODIS	Normalized difference moisture index	-1 to 1
GNDVI	MODIS	Green normalized difference vegetation index	-1 to 1
BAI	MODIS	Burned area index	0-1
Latitude	Mapbiomas	Latitude of each point	-
Longitude	Mapbiomas	Longitude of each point	-

Appendix B. Statistical metrics for 7 different removal variable scenarios for all tested machine learning models, in the test dataset for years 2019 and 2020. Predictions comprehend the data until July, 1 month before the majority of areas harvesting. Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Coefficient of determination (R^2), Root Mean Squared Logarithmic Error (RMSLE), Mean Absolute Percentage Error (MAPE) and Correlation Coefficient (r).

Model	Variables Removal	MAE	RMSE	R^2	RMSLE	MAPE	r
		t ha^{-1}	t ha^{-1}			%	
AdaBoost Regressor	APSIM (2)	4.62	6.15	0.38	0.086	6.46	0.73
AdaBoost Regressor	APSIM + MET (5)	3.73	5.27	0.55	0.073	5.06	0.76
AdaBoost Regressor	APSIM + VEG_IND (7)	4.34	5.72	0.47	0.079	5.87	0.69
AdaBoost Regressor	MET (4)	4.77	6.44	0.32	0.090	6.62	0.64
AdaBoost Regressor	MET + VEG_IND (6)	5.03	6.78	0.25	0.095	7.08	0.80
AdaBoost Regressor	No removal (full dataset) (1)	5.74	7.71	0.03	0.106	8.04	0.37
AdaBoost Regressor	VEG_IND (3)	5.31	7.10	0.18	0.099	7.48	0.68
Bayesian Ridge	APSIM (2)	4.69	6.00	0.41	0.084	6.44	0.70
Bayesian Ridge	APSIM + MET (5)	4.69	6.12	0.39	0.086	6.53	0.72
Bayesian Ridge	APSIM + VEG_IND (7)	4.62	6.11	0.39	0.083	6.21	0.65
Bayesian Ridge	MET (4)	4.53	5.76	0.46	0.081	6.25	0.77
Bayesian Ridge	MET + VEG_IND (6)	4.56	5.87	0.44	0.082	6.3	0.80
Bayesian Ridge	No removal (full dataset) (1)	4.52	5.73	0.47	0.080	6.22	0.77
Bayesian Ridge	VEG_IND (3)	4.66	5.96	0.42	0.083	6.43	0.79
Decision Tree Regressor	APSIM (2)	5.34	6.62	0.29	0.092	7.38	0.58
Decision Tree Regressor	APSIM + MET (5)	5.11	6.30	0.35	0.086	6.93	0.63
Decision Tree Regressor	APSIM + VEG_IND (7)	6.97	8.87	-0.28	0.120	9.45	-0.04
Decision Tree Regressor	MET (4)	5.29	7.40	0.11	0.102	7.45	0.48
Decision Tree Regressor	MET + VEG_IND (6)	6.64	8.68	-0.23	0.118	9.11	-0.35
Decision Tree Regressor	No removal (full dataset) (1)	5.31	7.48	0.09	0.103	7.5	0.48
Decision Tree Regressor	VEG_IND (3)	5.33	7.53	0.08	0.104	7.54	0.48
Elastic Net	APSIM (2)	4.66	5.91	0.43	0.082	6.34	0.69
Elastic Net	APSIM + MET (5)	4.96	6.43	0.33	0.090	6.92	0.71
Elastic Net	APSIM + VEG_IND (7)	4.67	6.05	0.40	0.083	6.36	0.69
Elastic Net	MET (4)	4.81	6.17	0.38	0.086	6.69	0.79
Elastic Net	MET + VEG_IND (6)	5.19	6.78	0.25	0.094	7.22	0.80
Elastic Net	No removal (full dataset) (1)	4.64	5.90	0.43	0.083	6.42	0.78
Elastic Net	VEG_IND (3)	4.91	6.33	0.35	0.088	6.81	0.80
Extra Trees Regressor	APSIM (2)	4.47	5.74	0.46	0.080	6.14	0.74
Extra Trees Regressor	APSIM + MET (5)	4.37	5.78	0.46	0.081	6.03	0.76
Extra Trees Regressor	APSIM + VEG_IND (7)	4.45	5.35	0.53	0.073	6	0.75
Extra Trees Regressor	MET (4)	5.44	7.04	0.19	0.097	7.49	0.65
Extra Trees Regressor	MET + VEG_IND (6)	5.85	7.98	-0.04	0.110	8.24	0.64
Extra Trees Regressor	No removal (full dataset) (1)	6.13	8.09	-0.07	0.111	8.55	-0.11
Extra Trees Regressor	VEG_IND (3)	5.68	7.62	0.05	0.105	7.97	0.64
Gradient Boosting Regressor	APSIM (2)	5.77	7.26	0.14	0.100	7.95	0.78
Gradient Boosting Regressor	APSIM + MET (5)	4.08	5.43	0.52	0.071	5.29	0.73
Gradient Boosting Regressor	APSIM + VEG_IND (7)	5.06	6.70	0.27	0.089	6.77	0.59
Gradient Boosting Regressor	MET (4)	4.87	6.79	0.25	0.094	6.77	0.51
Gradient Boosting Regressor	MET + VEG_IND (6)	5.77	7.46	0.09	0.103	7.98	0.64
Gradient Boosting Regressor	No removal (full dataset) (1)	6.04	7.75	0.02	0.106	8.34	0.52

Gradient Boosting Regressor	VEG_IND (3)	4.37	5.49	0.51	0.076	5.94	0.76
Huber Regressor	APSIM (2)	4.49	5.95	0.42	0.083	6.23	0.70
Huber Regressor	APSIM + MET (5)	4.50	6.00	0.41	0.084	6.28	0.71
Huber Regressor	APSIM + VEG_IND (7)	4.67	6.37	0.34	0.088	6.47	0.66
Huber Regressor	MET (4)	4.94	6.65	0.28	0.093	6.98	0.76
Huber Regressor	MET + VEG_IND (6)	5.58	7.52	0.08	0.104	7.89	0.72
Huber Regressor	No removal (full dataset) (1)	4.36	6.03	0.41	0.084	6.11	0.71
Huber Regressor	VEG_IND (3)	5.56	7.49	0.08	0.104	7.86	0.73
K Neighbors Regressor	APSIM (2)	3.94	5.40	0.52	0.075	5.41	0.76
K Neighbors Regressor	APSIM + MET (5)	4.62	6.13	0.39	0.086	6.44	0.76
K Neighbors Regressor	APSIM + VEG_IND (7)	4.28	6.05	0.40	0.083	5.84	0.65
K Neighbors Regressor	MET (4)	3.77	5.17	0.56	0.073	5.3	0.81
K Neighbors Regressor	MET + VEG_IND (6)	3.79	4.98	0.60	0.070	5.31	0.82
K Neighbors Regressor	No removal (full dataset) (1)	3.26	4.48	0.67	0.063	4.54	0.83
K Neighbors Regressor	VEG_IND (3)	3.51	4.66	0.65	0.065	4.86	0.84
Lasso Regression	APSIM (2)	4.86	6.07	0.40	0.083	6.46	0.66
Lasso Regression	APSIM + MET (5)	4.51	5.76	0.46	0.080	6.16	0.72
Lasso Regression	APSIM + VEG_IND (7)	4.48	5.99	0.41	0.081	5.98	0.65
Lasso Regression	MET (4)	4.67	6.30	0.35	0.086	6.25	0.60
Lasso Regression	MET + VEG_IND (6)	4.02	5.43	0.52	0.074	5.38	0.74
Lasso Regression	No removal (full dataset) (1)	4.71	6.37	0.34	0.087	6.37	0.58
Lasso Regression	VEG_IND (3)	4.39	5.93	0.43	0.081	5.88	0.67
Least Angle Regression	APSIM (2)	4.82	6.07	0.40	0.083	6.41	0.65
Least Angle Regression	APSIM + MET (5)	4.47	5.67	0.48	0.078	6.08	0.72
Least Angle Regression	APSIM + VEG_IND (7)	4.50	5.99	0.41	0.081	5.99	0.65
Least Angle Regression	MET (4)	4.69	6.41	0.33	0.087	6.24	0.59
Least Angle Regression	MET + VEG_IND (6)	4.07	5.46	0.51	0.074	5.4	0.73
Least Angle Regression	No removal (full dataset) (1)	4.67	6.41	0.33	0.088	6.29	0.58
Least Angle Regression	VEG_IND (3)	4.31	5.89	0.44	0.080	5.78	0.67
Light Gradient Boosting Machine	APSIM (2)	5.07	6.34	0.35	0.088	6.97	0.69
Light Gradient Boosting Machine	APSIM + MET (5)	4.65	6.17	0.38	0.084	6.37	0.62
Light Gradient Boosting Machine	APSIM + VEG_IND (7)	4.32	6.03	0.41	0.084	6.01	0.79
Light Gradient Boosting Machine	MET (4)	5.19	6.91	0.22	0.096	7.2	0.60
Light Gradient Boosting Machine	MET + VEG_IND (6)	6.10	7.86	-0.01	0.108	8.43	0.58
Light Gradient Boosting Machine	No removal (full dataset) (1)	5.50	7.35	0.12	0.102	7.68	0.52
Light Gradient Boosting Machine	VEG_IND (3)	4.85	5.95	0.42	0.081	6.51	0.65
Linear Regression	APSIM (2)	4.82	6.08	0.40	0.083	6.42	0.65
Linear Regression	APSIM + MET (5)	4.47	5.67	0.48	0.078	6.08	0.72
Linear Regression	APSIM + VEG_IND (7)	4.50	5.99	0.41	0.081	5.99	0.65
Linear Regression	MET (4)	4.69	6.42	0.33	0.087	6.25	0.59
Linear Regression	MET + VEG_IND (6)	4.07	5.46	0.51	0.074	5.4	0.73
Linear Regression	No removal (full dataset) (1)	4.67	6.42	0.33	0.088	6.29	0.58
Linear Regression	VEG_IND (3)	4.31	5.89	0.44	0.080	5.78	0.67
Orthogonal Matching Pursuit	APSIM (2)	4.95	6.62	0.29	0.092	6.96	0.67

Orthogonal Matching Pursuit	APSIM + MET (5)	4.49	6.00	0.41	0.084	6.31	0.73
Orthogonal Matching Pursuit	APSIM + VEG_IND (7)	5.00	6.66	0.28	0.093	7	0.71
Orthogonal Matching Pursuit	MET (4)	4.40	6.09	0.40	0.085	6.23	0.77
Orthogonal Matching Pursuit	MET + VEG_IND (6)	4.46	6.13	0.39	0.086	6.29	0.78
Orthogonal Matching Pursuit	No removal (full dataset) (1)	5.07	7.18	0.16	0.100	7.24	0.70
Orthogonal Matching Pursuit	VEG_IND (3)	5.18	7.28	0.14	0.101	7.38	0.67
Passive Aggressive Regressor	APSIM (2)	4.35	5.68	0.47	0.079	5.9	0.70
Passive Aggressive Regressor	APSIM + MET (5)	4.46	5.92	0.43	0.083	6.2	0.71
Passive Aggressive Regressor	APSIM + VEG_IND (7)	4.43	6.09	0.40	0.084	6.06	0.65
Passive Aggressive Regressor	MET (4)	4.27	6.08	0.40	0.084	5.84	0.65
Passive Aggressive Regressor	MET + VEG_IND (6)	4.36	6.31	0.35	0.087	6.12	0.67
Passive Aggressive Regressor	No removal (full dataset) (1)	4.35	6.26	0.36	0.086	5.98	0.62
Passive Aggressive Regressor	VEG_IND (3)	4.43	6.50	0.31	0.090	6.24	0.65
Random Forest Regressor	APSIM (2)	5.47	6.92	0.22	0.095	7.5	0.61
Random Forest Regressor	APSIM + MET (5)	5.29	6.73	0.26	0.093	7.23	0.60
Random Forest Regressor	APSIM + VEG_IND (7)	5.51	6.62	0.29	0.090	7.44	0.56
Random Forest Regressor	MET (4)	5.46	6.84	0.24	0.094	7.48	0.66
Random Forest Regressor	MET + VEG_IND (6)	5.18	7.24	0.15	0.100	7.27	0.52
Random Forest Regressor	No removal (full dataset) (1)	5.98	7.82	0.00	0.108	8.33	0.26
Random Forest Regressor	VEG_IND (3)	4.70	6.15	0.38	0.086	6.49	0.77
Ridge Regression	APSIM (2)	4.68	5.97	0.42	0.083	6.4	0.69
Ridge Regression	APSIM + MET (5)	4.75	6.19	0.38	0.087	6.62	0.72
Ridge Regression	APSIM + VEG_IND (7)	4.76	6.18	0.38	0.085	6.52	0.70
Ridge Regression	MET (4)	4.40	5.58	0.49	0.078	6.02	0.75
Ridge Regression	MET + VEG_IND (6)	4.39	5.58	0.49	0.078	6.02	0.79
Ridge Regression	No removal (full dataset) (1)	4.40	5.60	0.49	0.078	6.02	0.75
Ridge Regression	VEG_IND (3)	4.49	5.71	0.47	0.079	6.14	0.77

The variable removal scenarios are (1) Full dataset without variables removal, i.e. includes remote sensing vegetation indices, APSIM variables and meteorological variables; (2) Full dataset except APSIM variables; (3) Full dataset except remote sensing vegetation indices; (4) Full dataset except meteorological variables; (5) Dataset excluding APSIM and meteorological variables; (6) Dataset excluding APSIM and remote sensing vegetation indices; (6) Dataset without meteorological variables and vegetation indices and (7) Dataset excluding APSIM variables and vegetation indices